ECIS 2022 Research Papers

ECIS 2022 Proceedings

6-18-2022

GAM(e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints

Patrick Zschech
Friedrich-Alexander-Universität Erlangen-Nürnberg, patrick.zschech@fau.de

Sven Weinzierl
Friedrich-Alexander-Universität Erlangen-Nürnberg, sven.weinzierl@fau.de

Nico Hambauer
Friedrich-Alexander-Universität Erlangen-Nürnberg, nico.hambauer@fau.de

Sandra Zilker
Friedrich-Alexander-Universität Erlangen-Nürnberg, sandra.zilker@fau.de

Mathias Kraus
Friedrich-Alexander-Universität Erlangen-Nürnberg, mathiaskraus@ethz.ch

Recommended Citation

Zschech, Patrick; Weinzierl, Sven; Hambauer, Nico; Zilker, Sandra; and Kraus, Mathias, "GAM(e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints" (2022). *ECIS 2022 Research Papers*. 106.
https://aisel.aisnet.org/ecis2022_rp/106

GAM(E) CHANGER OR NOT? AN EVALUATION OF INTERPRETABLE MACHINE LEARNING MODELS BASED ON ADDITIVE MODEL CONSTRAINTS

Research Paper

Patrick Zschech, FAU Erlangen-Nürnberg, Nürnberg, Germany, patrick.zschech@fau.de
Sven Weinzierl, FAU Erlangen-Nürnberg, Nürnberg, Germany, sven.weinzierl@fau.de
Nico Hambauer, FAU Erlangen-Nürnberg, Nürnberg, Germany, nico.hambauer@fau.de
Sandra Zilker, FAU Erlangen-Nürnberg, Nürnberg, Germany, sandra.zilker@fau.de
Mathias Kraus, FAU Erlangen-Nürnberg, Nürnberg, Germany, mathias.kraus@fau.de

Abstract

The number of information systems (IS) studies dealing with explainable artificial intelligence (XAI) is currently exploding as the field demands more transparency about the internal decision logic of machine learning (ML) models. However, most techniques subsumed under XAI provide post-hoc-analytical explanations, which have to be considered with caution as they only use approximations of the underlying ML model. Therefore, our paper investigates a series of intrinsically interpretable ML models and discusses their suitability for the IS community. More specifically, our focus is on advanced extensions of generalized additive models (GAM) in which predictors are modeled independently in a non-linear way to generate shape functions that can capture arbitrary patterns but remain fully interpretable. In our study, we evaluate the prediction qualities of five GAMs as compared to six traditional ML models and assess their visual outputs for model interpretability. On this basis, we investigate their merits and limitations and derive design implications for further improvements.

Keywords: Predictive Analytics, Interpretable Machine Learning, Generalized Additive Models

1 Introduction

Due to the technological achievements in the field of machine learning (ML), many tasks related to predictive decision-making are increasingly supported by ML models (Janiesch et al., 2021; Kraus et al., 2020). Prominent examples can be found in business process monitoring (Heinrich et al., 2021; Stierle et al., 2021b), hate speech detection (Zinovyeva et al., 2020), medical diagnosis (McKinney et al., 2020), industrial maintenance (Kraus and Feuerriegel, 2019; Zschech et al., 2019), or natural disaster detection (DeVries et al., 2018). However, most advanced ML models such as deep neural networks (DNNs) and gradient boosting machines (GBMs) represent complex mappings between input features and the prediction targets. This results in black-box behavior which does not allow for sufficient model analysis to assess the internal decision logic and consequently leads to a lack of trust (Miller, 2019; Thiebes et al., 2021). For this reason, there is increasing concern about using such models in critical decision-making scenarios, such as healthcare, finance, and criminal justice (N. Agarwal and Das, 2020; Rudin, 2019).

To overcome such challenges, various techniques have been proposed in recent years to make black-box behavior more transparent and comprehensible. A commonly applied solution is to provide additional model explainability by approaches like local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) and Shapley additive explanations (SHAP) (Lundberg et al., 2020). Such techniques offer post-hoc explanations to better understand the predictions made by complex models (Barredo Arrieta et al., 2020). Nevertheless, post-hoc explanations have to be considered with caution as they can only be used to explain predictions of instances which are already known and therefore try to reconstruct the cause of a generated prediction through approximation. Hence, they might not be reliable and can result in misleading conclusions (Babic et al., 2021; John-Mathews, 2021; Rudin, 2019).

Another way of providing transparency is to build intrinsically interpretable ML models (Du et al., 2019). In general, the structure of interpretable models is in some way constrained, such that the resulting models provide a better understanding of how predictions are generated. This may include various constraints such as linearity, additivity, smoothness, and other types of structural simplifications (Rudin, 2019; Sudjianto et al., 2020). Widely known representatives of this class are linear/logistic regression models or decision rules, which can be easily interpreted but are too restrictive to capture more complex effects.

Beyond that, however, there are also more powerful models that are able to better address the trade-off between model accuracy and model transparency. One of the most promising directions in this area is that of *generalized additive models (GAMs)* (Hastie and Tibshirani, 1986; Lou et al., 2012). In GAMs, input variables are mapped independently of each other in a non-linear manner and the mappings are summed up afterward (Barredo Arrieta et al., 2020). Thus, GAMs include additive constraints yet drop the linearity constraint. In this way, it is still possible to extract univariate mappings from the model, which show the relation between single input features and the response variable. These relations are commonly known as *ridge* or *shape functions* (Lou et al., 2012). Since shape functions can take any arbitrary form, GAMs often achieve much better prediction accuracies as compared to simple linear models. However, as they do not contain high-order interactions between features, they can be easily interpreted by model users and developers for decision assessment or model debugging purposes (Lou et al., 2012).

Traditionally, shape functions in GAMs have been learned via splines (Hastie and Tibshirani, 1986), whereas recent proposals consider more advanced techniques, such as bagged and boosted tree ensembles (Caruana et al., 2015; Lou et al., 2012) or specific neural networks (R. Agarwal et al., 2021a). These approaches allow for a higher degree of flexibility and therefore are able to achieve better predictive performance while remaining fully interpretable. Beyond that, more advanced approaches have been proposed most recently with extensions towards generalized additive index models (GAIMs) (e.g., Vaughan et al., 2018; Yang et al., 2021a) and other enhancements to further improve prediction qualities and/or interpretability. Taken together, these approaches show a very promising direction towards intrinsically understandable white-box ML models that provide a technically equivalent, but ethically more acceptable alternative to black-box models, which is of pivotal interest to the information systems (IS) community.

While each GAM extension has already been evaluated in its own setting by the corresponding authors/developers, there is currently no neutral, independent, and, in particular, overarching cross-model comparison. As such, it currently lacks a comprehensive study that (i) examines the advantages and disadvantages of the individual models, (ii) reflects on the suitability for IS research and practice — especially with regards to commonly applied ML models, and (iii) derives implications for further improvements from a design perspective. Thus, we aim to answer the following research questions:

RQ1: *Can interpretable models based on additive model constraints provide competitive prediction results as compared to traditional white-box and black-box ML models?*

RQ2: *How do the outputs of interpretable ML models based on additive model constraints differ from each other objectively?*

To answer these questions, we perform a comprehensive and systematic evaluation study, in which we apply five GAM/GAIM-based models as well as six commonly used ML-models on twelve datasets to assess and compare their predictive performance. Subsequently, we use selected datasets to outline benefits and limitations concerning the interpretability of the intrinsically interpretable models. As a result, we formulate four design principles which may guide future research on the development of interpretable ML models and their integration into ML-based IS. Following this line, the paper is organized as follows: In Section 2, we provide an overview of the conceptual background and related work. Subsequently, we elaborate on the design of our evaluation study in Section 3. In Section 4, we provide a detailed presentation of the results related to the assessment of predictive performance and model interpretability. Finally, we discuss our findings, outline the potential role of GAM-based models for the IS community, and provide an outlook for future work in Section 5.

2 Conceptual Background and Related Work

2.1 Model Explainability and Model Interpretability

There are basically two strands of research that are concerned with the transparency of ML models. The first one refers to *post-hoc model explainability*. It deals with techniques that provide post-hoc explanations for predictions that are made by complex black-box models in the first place (e.g., DNNs). For this purpose, additional/external approaches are used to explain how results are generated by black-box models (Barredo Arrieta et al., 2020). Such approaches are often subsumed under the term explainable artificial intelligence (XAI) and commonly known examples are variable importance (Breiman, 2001), partial dependence plots (Friedman, 2001), individual conditional expectation plots (Goldstein et al., 2015), LIME (Ribeiro et al., 2016), and SHAP (Lundberg et al., 2020).

Post-hoc-analytical techniques have led to insightful analysis into powerful black-box models, such as gradient boosted decision trees or DNNs. However, these techniques can not represent the full, highly complex functioning of black-box models in a simple manner, but rather give snapshots of the functioning based on few instances. By aggregating these snapshots over multiple instances, the user can draw more general conclusions about the model, yet, these conclusions are only based on a limited number of samples. For a new instance, post-hoc analyzed models could still output unforeseeable, and potentially harmful results (Babic et al., 2021; Kaur et al., 2020; Rudin and Radin, 2019).

The second line of research refers to *intrinsic model interpretability* (Du et al., 2019). This field is concerned with ML models that are designed to be inherently interpretable, often due to their simplicity, such as given by generalized linear models, point systems, simple decision trees, decision rules, or naive Bayes classifiers (Lou et al., 2012; Rudin and Radin, 2019; Yang et al., 2021c). To retain transparency, interpretable models are usually restricted in some way by introducing certain model structures and practical constraints, such as linearity, monotonicity, additivity, sparsity, smoothness, and near-orthogonality (Rudin and Radin, 2019; Sudjianto et al., 2020).

While model interpretability and explainability share the same goal of providing users with insights into how models work, they differ substantially in the underlying idea of how to achieve this goal. The research stream of model interpretability designs models in such a way, that the underlying mathematical function (which fully defines the model) is simple enough that users can access and analyze it directly. In contrast, model explainability generally does not constrain models, but rather adds another layer that attempts to simplify the function so that it is digestible by a user. Furthermore, it should be noted that in this paper, the terms explainability and interpretability are not aimed at subjective perceptual processes during cognitive reasoning, but at the functionality of a model to produce objectively comprehensible outputs.

Often, there is the erroneous belief that prediction accuracy must be strongly sacrificed for interpretability/transparency which is why complex black-box models seemed to be favored over interpretable approaches during the past decade of ML research (Rudin and Radin, 2019). However, most recent proposals and studies have demonstrated the opposite as there are seemingly algorithmic approaches that are able to combine both aspects. One of these promising algorithmic approaches is that of GAMs and their subsequent advancements based on additive model constraints (Barredo Arrieta et al., 2020).

2.2 Generalized Additive Models and Recent Advancements

Let $D = (X, y)$ denote a training dataset, where $X = (x_1, ..., x_n) \in \mathbb{R}^{N \times n}$ is the feature matrix comprising N observations and n features. Further, y denotes the corresponding targets. A generalized additive model is then defined as:

$$g(y) = f_1(x_1) + ... + f_n(x_n), \tag{1}$$

where $g(\cdot)$ is called link function and $f_i(\cdot)$ is the shape function for a feature x_i. If the link function is the identity, Equation 1 describes an additive model (e.g., a regression model) and if the link function is the logistic function, Equation 1 describes a generalized additive model (e.g., a classification model) (Hastie and Tibshirani, 1986; Lou et al., 2012). In this paper, we consider both prediction tasks, i.e., regression problems where $y \in \mathbb{R}^N$ as well as binary classification problems where $y \in \{1, 0\}^N$. Given a model F, let $F(X)$ denote the predictions of the model for our data points X. The goal in both tasks is to minimize the expected value of some loss function $L(y, F(X))$. The purpose of GAMs is to infer the shape functions whose aggregate composition approximates the predicted response variable. The overall structure is simply interpretable, as it allows model users and developers to verify the importance of each variable. In other words, it is directly observable how each feature, through its corresponding shape function, affects the predicted output. In Section 4, we will provide several illustrative examples of such shapes when assessing the output of different GAMs. Due to the intrinsic interpretability of non-linear effects, traditional GAMs have been widely used already in various application scenarios, especially in fields related to risk assessment where trust had to be built with the user (Barredo Arrieta et al., 2020), such as finance (Berg, 2007), healthcare (Caruana et al., 2015), energy supply (Pierrot and Goude, 2011), geology (Tomić and Božić, 2014), and environmental studies (Murase et al., 2009).

Originally, shape functions in GAMs were learned via regression *splines* (Hastie and Tibshirani, 1986). These are piecewise polynomial functions that can approximate complex shapes through curve fitting. However, more recent studies have shown that splines are often too smooth for real-world datasets and that higher predictive performance can be achieved using more flexible models. To this end, Lou et al. (2012) proposed to consider bagged and boosted decision trees ensembles for fitting complex shape functions. The authors further enhanced their tree-based approach by considering pairwise interaction terms (Lou et al., 2013), and made it publicly available as an easy-to-use algorithm, known as *explainable boosting machine (EBM)* (Nori et al., 2019).

More recently, further GAM modifications have been proposed based on DNNs. R. Agarwal et al. (2021a) introduced a *neural additive model (NAM)* in which shape functions are learned via individual deep sub-networks with multiple hidden layers and specific neural units. More specifically, they introduced exponential units (ExU) for fitting jagged curves that often appear in real-world datasets. Another DNN-based approach was pursued by Yang et al. (2021c) proposing *GAMI-Net*. To improve predictive performance, GAMI-Net is designed to capture pairwise interactions. This also required further model constraints, such as heredity and marginal clarity constraints, to retain structural interpretability and avoid mutual absorption between main effects and pairwise interactions. Another advanced DNN approach was proposed by Vaughan et al. (2018) presenting *explainable neural network (xNN)*, which was further improved towards *enhanced explainable neural network (ExNN)* (Yang et al., 2021a). Instead of using a simple GAM structure, both models are based on the structure of GAIMs. This structure generally violates the idea of univariate feature mappings due to an additional projection layer that fully connects all input features to the following sub-networks so that each feature can possibly have a partial contribution to all corresponding shape functions.

2.3 Comparative Evaluation Studies

Since all GAM/GAIM-based models introduced above have distinct characteristics with very specific model constraints, it is worthwhile to evaluate and compare their merits and limitations for different prediction tasks and datasets. A few authors and developers have already compared their GAM extensions to some competing models as well as traditional ML baselines to demonstrate their proposed innovations. For example, Lou et al. (2012) performed comprehensive experiments to compare spline-based models with tree-based GAMs and additionally considered logistic/linear regression (LR) and random forest (RF) as lower and upper bound baselines, respectively. When introducing GAMI-Net, Yang et al. (2021c) compared it with EBM, splines, and several other benchmark models, including LR, RF, extreme gradient boosting (XGB), and multi-layer perceptron (MLP), using a large number of datasets (20+). Likewise, R. Agarwal et al. (2021a) benchmarked their NAM against EBM and a number of traditional approaches. However, in their study, the number of evaluated datasets was limited to a smaller amount.

Apart from that, there are only a few studies that have evaluated the properties of different additive models so far. Chang et al. (2021) investigated a series of GAMs including splines and tree-based approaches like EBM. They investigated the models quantitatively and qualitatively using real-world and simulated data. Hohman et al. (2019) integrated GAMs into a visual analytics tool to examine how data scientists interact with shape functions. A slightly different perspective was taken by Kaur et al. (2020). They also examined how data science professionals use and evaluate such interpretable models. However, they additionally considered models with post-hoc explanations provided by SHAP to compare both approaches.

In summary, when considering the focus of related work, it currently lacks a cross-model comparison to evaluate the merits and limitations of different GAMs from a neutral perspective. To close this gap, we contribute new insights with a comparative evaluation study.

3 Research Method

To answer our research questions, we performed a series of computational experiments. In the first part, our focus was on the assessment of the predictive performance (RQ1), and subsequently, we considered the models' interpretability (RQ2). For our cross-model comparison, we examined five different GAM/GAIM approaches, for which publicly accessible implementations are available. This includes (i) **Splines** (Hastie and Tibshirani, 1986), (ii) **EBM** (Nori et al., 2021), (iii) **NAM** (R. Agarwal et al., 2021a), (iv) **GAMI-Net** (Yang et al., 2021c), and (v) **ExNN** (Yang et al., 2021a). The implementations are provided by the respective authors of the proposed models, except for splines, for which we used the Python package pyGAM. Additionally, several benchmark models were included for a broader comparison, including **LR** and **decision tree (DT)** as common representatives of interpretable models, and **RF**, **GBM**, **XGB** and **MLP** as widely known black-box models (Amancio et al., 2014; Roy et al., 2019). For their implementation, we used the corresponding Python packages from the scikit-learn library, except for XGB where we used the XGBoost library. To provide a fair comparison, we integrated all implementations into a shared environment to run the experiments under the same conditions (i.e., workstation with single GPU NVIDIA Quadro RTX A 5000, 8 CPU cores, 24 GB VRAM and 64 GB RAM, Python 3.6.8, Tensorflow 2.3.0). Considering the choice of hyperparameters, we pursued the approach of leaving the models in their default configurations or used the settings recommended by the authors. In addition, we set the number of interactions of the EBM to 10. Further details on the implementation and hyperparameter settings are given in the appendix in Table 6.

To ensure a comprehensive evaluation, we assessed the predictive performance of all models on a wide range of benchmark scenarios. For this purpose, we looked into related evaluation studies (e.g., Roy et al., 2019; Yang et al., 2021c) to identify commonly used repositories that offer publicly available dataset collections, such as Kaggle (https://www.kaggle.com/) and the UCI machine learning repository (http://archive.ics.uci.edu/ml/). We then selected appropriate datasets in such a way that their inherent properties cover a broad and representative range of real-world applications and that the prediction tasks are associated with organizational and/or societal challenges (as opposed to biological, physical or other phenomena). Furthermore, we limited our analysis to medium-sized datasets to keep the computational effort manageable. As a result, we chose twelve datasets with corresponding prediction tasks that are summarized in Table 1. They cover seven (binary) classification (CLS) and five regression (REG) tasks[1]. The number of observations ranges from 205 to 103,904, and the number of predictors varies between 8 and 99 with a mixed combination of numerical and categorical features. Thus, we consider a variety of settings for assessing the models. When loading the datasets, we removed IDs and variables that cause obvious data leakage, cleaned missing values, removed categorical features with more than 25 distinct values to reduce computational complexity as we one-hot encode these, and converted the targets into a common format. Apart from that, we kept the datasets in their default structure without extensive preprocessing.

Type	Dataset	Observations	Features num/cat	Prediction target	Repository
CLS	Water potability	3,276	9/0	Will the water be safe for consumption?	Kaggle (2021d)
	Stroke	5,110	3/7	Will a patient suffer from a stroke?	Kaggle (2021b)
	Telco churn (IBM, 2019)	7,043	3/16	Will a customer leave the company?	Kaggle (2021c)
	FICO credit score	10,459	21/2	Will a client repay within 2 years?	FICO (2021)
	Adult (Kohavi, 1996)	32,561	7/8	Will the income exceed $50.000/year?	UCI (2021a)
	Bank marketing (Moro et al., 2014)	45,211	5/11	Will a client subscribe to a deposit?	UCI (2021c)
	Airline satisfaction	103,904	18/4	Will a passenger be satisfied?	Kaggle (2021a)
REG	Car price (Kibler et al., 1989)	205	13/11	What is the price of a car?	UCI (2021b)
	Student grade (Cortez and Silva, 2008)	649	13/17	What is a student's final grade?	UCI (2021f)
	Crimes (Redmond and Baveja, 2002)	1,994	99/0	How many violent crimes will happen?	UCI (2021e)
	Bike rental (Fanaee-T and Gama, 2014)	17,379	6/6	How many bikes will be rented/hour?	UCI (2021d)
	California housing (Pace and Barry, 1997)	20,640	8/0	What is the value of a house?	S&P (2021)

Table 1. Overview of used benchmark datasets for classification (CLS) and regression (REG) tasks.

To assess the prediction qualities, we used a 5-fold cross validation where we measured the out-of-sample performance on each test fold and subsequently calculated the mean and standard deviation across all values. For classification, we measured accuracy, precision, recall, and F1-score; and for regression, we calculated root mean square error (RMSE) and mean absolute error (MAE) as commonly applied prediction metrics. Additionally, we also measured the training times in seconds. All our experiments with the corresponding implementations and evaluation results can be found in the following GitHub repository: `https://github.com/fau-is/gam_comparison`.

To assess the GAMs' interpretability, we examined and compared the visual outputs of the different models. More specifically, we followed the notion that an interpretable model must be able to provide transparency at the level of the entire model (simulatability) and at the level of individual components (decomposability) to enable an understanding of how a model works (Lipton, 2018). To this end, we looked into the GAMs' entire output as well as specific shape functions and compared both between different models.

4 Results

4.1 Evaluation of Predictive Performance and Training Times

In this section, we present the results of our computational experiments. Table 2 and Table 3 outline the prediction results for the classification and regression tasks, respectively. We report the F1-score and RMSE as our main metrics for comparison. The best overall performance for each dataset is highlighted in bold, whereas the best result among the interpretable models is marked with an underscore. Further results on the remaining metrics can be found in our online repository.

The results demonstrate that the best prediction values for the individual datasets are highly scattered across the variety of ML models. Taken together, the black-box models achieved the best results in 8

[1] The imbalance between the two prediction tasks reflects the fact that the public repositories mentioned above offer far more datasets for classification tasks than for regression tasks.

	Interpretable Models							Black-box Models			
Dataset	Splines	EBM	NAM	GAMI-Net	ExNN	LR	DT	RF	GBM	XGB	MLP
Water	.557±.020	.631±.017	.469±.009	.634±.014	.632±.029	.464±.003	.609±.007	.547±.024	.633±.025	.642±.010	**.646±.017**
Stroke	.928±.001	.927±.001	.928±.001	.928±.001	.927±.004	.928±.001	.917±.001	.928±.001	.928±.003	**.930±.002**	.928±.001
Telco	**.800±.005**	.797±.006	.723±.015	.799±.014	.787±.012	.799±.010	.747±.008	.780±.017	.790±.009	.782±.006	.790±.006
FICO	.725±.012	.725±.009	.619±.065	**.728±.009**	.708±.008	.718±.010	.667±.015	.716±.012	.722±.009	.710±.010	.716±.011
Adult	.854±.001	.866±.001	.727±.024	.856±.002	.851±.005	.846±.002	.846±.004	.826±.004	.867±.001	**.868±.002**	.850±.003
Bank	.893±.004	.895±.002	.833±.010	.893±.003	**.899±.003**	.888±.004	.892±.001	.859±.002	**.899±.003**	**.899±.003**	.898±.004
Airline	.935±.002	.945±.002	.773±.029	.934±.002	.951±.002	.875±.002	.950±.002	.921±.002	.958±.002	**.963±.002**	.958±.002

Table 2. Predictive performance for classification tasks measured by F1-score.

	Interpretable Models							Black-box Models			
Dataset	Splines	EBM	NAM	GAMI-Net	ExNN	LR	DT	RF	GBM	XGB	MLP
Car	.132±.022	**.083±.0310**	1.112±.223	1.001±.100	.129±.024	.090±.033	.148±.061	.093±.028	.098±.036	.097±.038	.106±.035
Student	.732±.129	.730±.139	1.239±.161	1.000±.171	1.390±.226	.731±.140	1.239±.190	**.712±.096**	.757±.121	.800±.092	.781±.147
Crimes	.503±.021	**.311±.026**	1.152±.047	.388±.016	.511±.059	.312±.026	.620±.056	.319±.024	.320±.018	.343±.030	.351±.038
Bike	.182±.006	.060±.001	1.214±.118	.145±.002	.114±.007	.499±.013	.080±.003	.202±.003	.045±.001	**.041±.001**	.079±.007
Housing	.242±.008	.181±.007	1.131±.153	.266±.001	.223±.008	.352±.009	.261±.005	.357±.014	.171±.006	**.163±.007**	.214±.007

Table 3. Predictive performance for regression tasks measured by RMSE.

out of 12 datasets. However, with some overlap, the interpretable GAMs also achieved outperforming results in 5 out of 12 datasets. Additionally, it is notable that the difference between the best-performing models from both groups is marginally small. For the regression tasks, the largest difference in RMSE is 0.019 (bike: EBM 0.06 vs. XGB 0.041), whereas for classification, the largest difference measured by F1-score is only 0.012 (water: GAMI-Net 0.634 vs. MLP 0.646 | airline: ExNN 0.951 vs. XGB 0.963). These results clearly outline that there is no strict trade-off between model accuracy and interpretability.

Furthermore, EBM turns out to be the best performing GAM among all interpretable approaches showing the highest prediction qualities in 6 out of 12 datasets. For the regression tasks, it even outperformed all other white-box models and could achieve the best results on 2 out 5 datasets. For the classification tasks, GAMI-Net showed the best performance among all interpretable models, followed by ExNN and Splines. Thus, there is no single approach that clearly dominates all other interpretable models.

Considering the results of the black-box models, it is notable that XGB offers the strongest prediction qualities for a wide range of dataset properties on both types of prediction tasks. Thus, it shows the best results in 6 out 12 datasets, which justifies why the model is often the first choice among developers in ML competitions which purely aim at achieving high prediction qualities. By contrast, it is intriguing to reveal that LR and DT, as traditional interpretable models, do not necessarily lag behind by large margins. The LR model, for example, achieves the second-highest performance in 3 out of 12 datasets (i.e., stroke, telco, car), affirming the observation by Rudin (2019) that the results of simple models may not significantly differ from those of black-box models when dealing with structured problems. The worst performance across the majority of datasets was achieved by the NAM model due to strong overfitting. We assume that the proposed default hyperparameters are not well-suited across various datasets. Thus, it requires further investigations in subsequent evaluation studies with a particular focus on model tuning.

Apart from the prediction qualities, we also looked into training times to assess the model usability. Table 4 provides an overview of the results measured in average seconds per fold. As expected, the simple models, i.e., LR and DT, achieve the lowest training times with values primarily < 0.1 seconds due to their basic model structures. These models are followed by the more complex black-box models with average training times ranging from around 0.1 up to 12 seconds. Furthermore, the results of the GAMs reveal that all three neural-based approaches require much more training time than EBM and Splines, whereas the latter still show short training times up to 31 and 39 seconds, respectively. Among the neural-based approaches, NAM requires the most time to train, followed by GAMI-Net and ExNN. This order holds for almost all datasets across both tasks.

	Interpretable Models							Black-box Models			
Dataset	**Splines**	**EBM**	**NAM**	**GAMI-Net**	**ExNN**	**LR**	**DT**	**RF**	**GBM**	**XGB**	**MLP**
Water	0.150	1.510	112.940	32.510	11.740	0.020	0.010	0.220	0.660	0.110	0.260
Stroke	1.230	0.690	129.110	46.700	23.360	0.010	0.010	0.110	0.440	0.080	0.430
Telco	2.090	1.510	270.130	57.510	16.720	0.020	0.020	0.150	0.900	0.120	0.660
FICO	2.370	1.660	403.550	95.370	14.520	0.030	0.040	0.270	1.740	0.120	0.950
Adult	38.540	21.640	1,567.940	519.720	44.040	0.100	0.070	0.450	3.480	0.460	3.340
Bank	15.450	9.490	1,802.360	777.130	63.400	0.120	0.090	0.610	3.970	0.470	4.330
Airline	17.750	31.020	3,020.080	627.220	225.090	0.100	0.210	2.020	12.070	0.720	9.470
Car	0.680	1.830	82.760	6.990	8.060	0.001	0.001	0.080	0.050	0.040	0.020
Student	0.410	0.440	70.560	16.010	7.910	0.001	0.001	0.100	0.090	0.050	0.060
Crimes	3.040	1.570	367.670	92.950	9.240	0.001	0.050	1.320	1.910	0.120	0.200
Bike	0.840	3.820	557.280	299.090	42.540	0.001	0.030	0.850	1.270	0.150	1.420
Housing	0.180	6.970	720.800	418.850	46.380	0.001	0.060	1.880	2.830	0.300	1.420

Table 4. Evaluation results of training time measured in average seconds per fold.

4.2 Evaluation of Model Interpretability

The qualitative assessment and discussion of the model interpretability is based on the example of the *adult* dataset (Kohavi, 1996). It contains about 30,000 observations and 15 socio-demographic and job-related features from the 1994 Census database. The prediction task is to determine whether a person makes over \$50,000 per year. After training all models on the entire dataset, we generated feature plots to visualize the resulting shape functions. Figure 1 summarizes exemplary plots for (a) Splines, (b) EBM, and (c) GAMI-Net, whereas Figure 2 shows selected excerpts for (d) NAM and (e) ExNN, respectively.

Focusing on the first group of models (a-c), it can be seen that all models basically provide the same output representation, i.e., it is displayed which impact a single feature or a pair-wise feature interaction has on the target variable (i.e., *income > \$50.000/year*). The impact of numerical features is shown by curves with different shapes, whereas the impact of categorical features (one-hot encoded) is shown by bar charts for the feature values 0 vs. 1. The x-axes represent the feature values and the y-axes represent the impact on the target variable (i.e., positive impact for $y > 0$, negative impact for $y < 0$). For interaction terms, both x-axis and y-axis represent the feature values, whereas the impact on the target is highlighted with a corresponding color scheme, resulting in a two-dimensional heatmap. In addition, EBM and GAMI-Net provide a calculation of global feature importance, which is displayed by bar charts with a decreasing order of relevant features. Due to space restrictions, only the ten most relevant features are shown.

To compare the same output and highlight differences between shape functions among all models, we selected a subset of important numerical features, i.e., *age*, *capital gain* and *hours per week*, and one categorical feature, i.e., *sex = female*, which is only displayed for EBM and GAMI-Net. Furthermore, since EBM and GAMI-Net can detect and capture relevant interaction terms automatically, we added an exemplary interaction plot (i.e., *hours per week* × *marital status = married-civ-spouse*).

From the detailed plots in Figure 1, we can see that GAMs learn different patterns and relations from the training data depending on the underlying model structure. In the case of Splines (a), for example, it is observable that the feature *age* has a positive effect on the target variable from a value of about 30 years and that the effect continues to increase until about 50 years. After this point, the impact slightly decreases again until 70 years, where we observe another turn towards a higher effect until 80 years. The other two models capture similar effects for that feature. However, since EBM (b) relies on an ensemble of decision-tree learners, the shape function is piece-wise constant, resulting in a step curve with sharp jumps at discrete values. On this basis, more fine-granular patterns are captured. Thus, we see more peaks with smaller ranges at certain feature values. The shape plot of GAMI-Net (c), on the other hand, shows a smooth behavior which is similar to that of Splines. Overall, such learned representations can be easily understood for interpretation purposes as they graphically reveal non-linear patterns that cannot

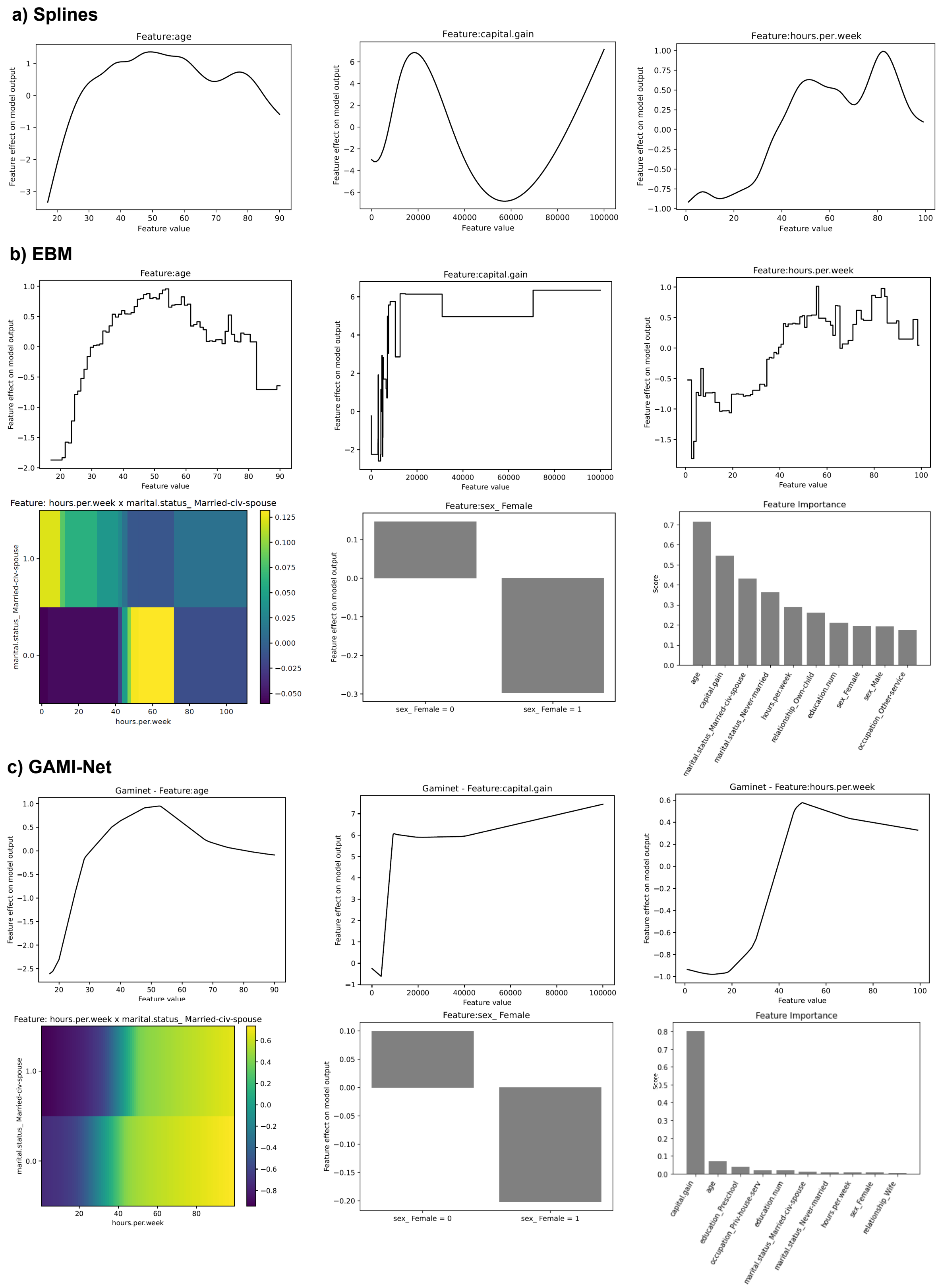

Figure 1. *Visual model outputs for Splines (top), EBM (middle) and GAMI-Net (bottom).*

be represented by simple LR models. Furthermore, in contrast to post-hoc-analytical methods, the plots directly display the actual decision logic of the final models without any approximations.

When comparing the other feature plots, it is also notable that the different models do not always represent the exact same patterns. Some illustrative examples are the deviating shapes for the feature *capital gain*. The difference can be explained by the fact that each model uses distinct model structures (i.e., splines, vs. trees vs. neural networks) to map the feature space to the target space. This can also be seen by the different feature importance plots, where EBM due to missing sparsity constraints considers several features as almost equally important, whereas in GAMI-Net, the prediction is mostly driven by a single feature which possibly absorbs partial effects of the remaining variables.

The other two models, NAM and ExNN, were excluded from our previous considerations because they have shown a fundamentally different behavior. Due to strong overfitting effects of NAM, which we could not resolve without intensive hyperparameter tuning, the model produces extremely jagged shape functions as shown in Figure 2 on the left side. As such, we could not generate comprehensible feature plots, which will be part of subsequent studies when tuning all models for specific data properties. The ExNN, on the other hand, is a special case as it is based on the structure of an additive index model. For the interpretation of the model, this means that not only a single feature is covered by a corresponding shape function, but that an entire set of features can possibly provide partial contributions to this shape. Figure 2 on the right side shows an example of such an output, which is hardly interpretable in a meaningful way if many features are involved.

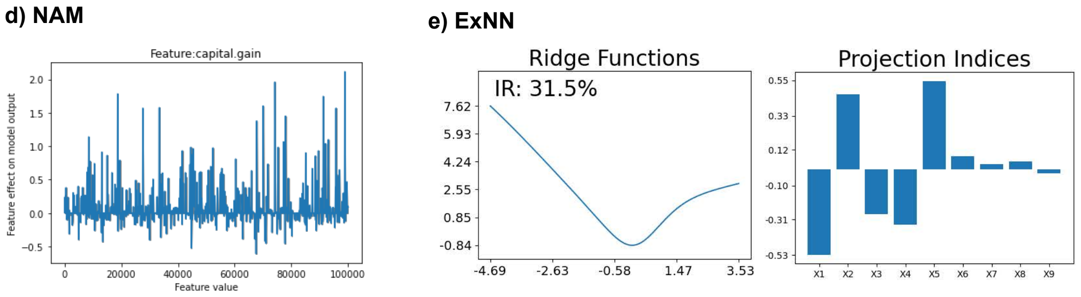

Figure 2. Visual model outputs for NAM (left) and ExNN (right).

4.3 Summary of Cross-Model Comparison

After examining and comparing the GAMs from different perspectives, we derive four design principles by reflecting on the merits and limitations of the individual approaches. Table 5 provides an overview of the compared characteristics upon which the design principles are derived. These can serve as a summary to guide upcoming research in the field of GAM extensions and interpretable ML.

Principle 1: *Provide constraints to derive shape functions that can capture complex patterns but remain comprehensible*. As shown before, simple Splines have the advantage of being simply comprehensible due to their strict smoothness, but this can also impair predictive performance as they cannot capture more complex patterns. NAM and EBM, on the other hand, are able to extract and depict non-smooth effects, such as sharp jumps due to abrupt changes or discrete thresholds given in real-world settings. Nevertheless, this behavior also bears the risk of producing unexpected jumps that are difficult to justify, which may become even worse in the presence of noisy samples and outliers. GAMI-Net offers a compromise by providing smooth shape outputs while still guaranteeing strong prediction qualities based on neural feature subnets and the identification of relevant pair-wise interactions. To take account for different patterns and shapes, a future innovation could be, that a model is able to capture different shapes, depending on whether a user prefers strictly smooth curves or jagged behavior with sharp jumps.

Model	Feature Shapes	Model Compactness	Controllability	Training Effort
Splines	Smooth behavior	Low complexity	Controllable sparsity, controllable smoothing	Simple configuration, medium training times
EBM	Piece-wise constant with sharp jumps	Increased complexity due to missing sparsity	Controllable number of interactions	Simple configuration, fast training times
NAM	Jagged behavior	Low complexity	Controllable regularization for limiting jagged behavior	Extensive HP tuning, slow training times
GAMI-Net	Smooth behavior	Low complexity	Controllable sparsity, controllable smoothness, controllable interactions	Moderate HP tuning, moderate training times
ExNN	Smooth behavior of feature projections	High complexity due to feature projections	Controllable sparsity, controllable smoothness	Moderate HP tuning, moderate training times

Table 5. Comparison of examined GAM characteristics.

Principle 2: *Provide constraints to keep the model's complexity manageable.* The ExNN is an example of a model where the internal decision logic is transparent but difficult to comprehend — especially for real-world problems where many features are involved. Thus, although the additional projection layer may boost predictive performance, it is hard to explain the practical meaning of partial feature combinations that are spread over multiple shape functions. The other models, on the other hand, only depict a single feature at a time — or at most a pair-wise interaction — which remains simply comprehensible. Furthermore, to keep model complexity manageable, some models such as GAMI-Net and Splines provide sparsity constraints to receive a more compact representation that concentrates on fewer features. Since EBM does not provide such constraints, the overall model complexity cannot be controlled.

Principle 3: *Provide constraints to control the model's overall structure and its components.* All five approaches provide some sort of controllability to modify the model's structure, such as the number of interactions (e.g., EBM), smoothness (e.g., ExNN), and/or sparsity (e.g., GAMI-Net). Such control appears to be crucial when model developers and domain experts want to integrate their knowledge into the model structure, depending on the situation. For example, by controlling the sparsity, a developer can counteract the problem of multicollinearity in case of highly correlated but irrelevant features. By contrast, a non-sparse model might help to identify biases within the training data which in case of a strong default regularization could be compiled into other correlated features and remain unrecognized (Chang et al., 2021). Therefore, it is helpful to provide different mechanisms for model modification so that users can create models according to their specific needs and intentions.

Principle 4: *Provide mechanisms to allow for simple usability with minimal training effort.* While EBM and Splines are comparatively fast and can be simply applied without much configuration effort, the neural-based GAMs suffer from the limitation that they are computationally intensive and require careful parameter tuning. Especially for NAM, it appears highly critical to choose a suitable set of hyperparameters (HP) that regulate the overfitting effect. Consequently, such approaches lack user-friendly applicability so that a model can be quickly applied without much training effort. This seems essential for interpretable models since users and developers may go through several iterations of model adjustments after spotting the performance and the visual output of individual shape functions.

5 Discussion and Outlook

ML models are highly beneficial for capturing hidden patterns and crucial nuances in large datasets. It can therefore be expected that the importance and prevalence of ML models will continue to increase in the future to assist or complement human decision-making and prediction tasks. Thus, the central question will not be *whether* ML models should be employed, but instead *which* type of ML models. In several domains, such as healthcare or financial risk assessment, simple regression models are still the most widely applied models as transparency is mandated due to the severe consequences of incorrect model predictions or because regulatory requirements enforce it (e.g., Bertoncelli et al., 2020; Shipe et al., 2019; Valaskova et al., 2018). At the same time, many research communities currently emphasize the development and application of XAI approaches to make complex black-box models more transparent and comprehensible. A similar development can be observed in our IS discipline, with an increasing focus on post-hoc-analytical methods such as LIME or SHAP (e.g., Jussupow et al., 2021; Mehdiyev and Fettke, 2021; Schemmer et al., 2021; Stierle et al., 2021a; Wanner et al., 2020a,b; Wastensteiner et al., 2021; Zhang et al., 2020). However, the concerns being raised about post-hoc-analytical methods should also be taken seriously in our field (Rudin, 2019). The option of choosing a complex black-box model, which subsequently requires a posteriori explanations, should only be considered if there is no better alternative to it. To this end, Rudin (2019) emphasizes the need in critical applications to first prove that high predictive performance can only be provided by black-box models, which in turn cannot be achieved with interpretable alternatives.

The results of our study directly contribute to this central debate. Thus, we were able to show that intrinsically interpretable ML models can indeed achieve competitive prediction qualities, providing further evidence that there is no strict trade-off between model interpretability and model accuracy. In fact, the performance of the examined GAMs was much closer to the prediction results of the black-box models than to those of the traditional white-box models, if not even outperforming them in 5 out of 12 prediction tasks. Consequently, we argue that advanced GAM models such as EBM or GAMI-Net should be firmly established as first-choice models in predictive modeling projects as we see large potential of these models to change the game when developing and applying ML approaches in research and industry alike.

Against this background, we also see an important role of these models in our socio-technical discipline of IS research. Due to their flexibility to capture complex patterns and their simplicity to produce easily understandable outputs, they provide a technically equivalent, but ethically more acceptable alternative to black-box models. In the past, we saw many examples where bias issues and fairness problems have been reported in ML applications. Prominent examples stem from recruiting applications where women have been discriminated or in pretrial detention and release decisions with skewed predictions against African Americans (Janiesch et al., 2021; Mehrabi et al., 2021). While detecting biases in training data is considered a tedious task, with advanced GAMs such distortions can be better spotted by means of the shape functions within the final model in order to avoid racial, sexual and other kind of discrimination. For instance, taking our demonstration case from Section 4.2 as an example, it can be quickly spotted (i) which model incorporates critical features such as gender, age and marital status for the income prediction, (ii) which feature values have a stronger tendency towards a higher income (e.g., sex = male, age between 30 and 60), and (iii) whether the models captured interaction effects of critical features with other variables that need to be considered (e.g., marital status × hours per week).

Likewise, model users can better interact with such intrinsically interpretable ML models, since they directly see how individual features influence a model's outcome. Thus, domain knowledge can be incorporated into the model by removing or adding certain features, adjusting specific model properties (e.g., smoothness), or adding further model constraints (e.g., upper and lower bounds in shape functions) to take better account for desired patterns. Because of all these properties, GAMs can be considered promising for critical applications in practice. For example, in the health sector, full transparency of diagnoses is essential for medical professionals to reconcile the model's results with their own experience and build trust in the prediction. Other examples may refer to safety-critical applications, such as natural disaster detection, crime prediction and financial fraud detection, in which it is crucial to gather a full understanding about a model's predictor variables and their impact on the target.

Nevertheless, to avoid any confusion, we also want to emphasize the point that careful conclusions have to be drawn on the basis of such interpretable models. Using the words of Caruana et al. (2015), *"it is tempting to interpret them causally"*, but even though the evaluated GAMs offer transparent explanations of how predictions are derived, they are still based on correlations. As such, it cannot be said for certain why some of the effects shown in the feature plots are present. This could be due to overfitting, interactions and correlations with other (unmeasured) variables, or due to other underlying phenomena. Nevertheless, even though such ML models cannot guarantee causality by themselves, they provide some suggestive indications for further investigations that can be helpful for the identification of causal pathways.

As with any research, our work is not free of limitations. In the current study, we restricted our analyses to medium-sized datasets to keep the computational costs manageable. This was necessary as we could see that almost all GAM-based models require high computational resources with large training times. Thus, investigating datasets with other properties (e.g., large-scale collections with several million observations and many more features) is an open issue for future work to provide more evidence for our findings.

Likewise, we explicitly focused on prediction tasks with tabular data that usually contain naturally meaningful features for interpretation purposes. In domains with higher-dimensional data such as images, text, and event logs, the results of this study may not directly apply. For this purpose, some upstream feature engineering methods are required that transform raw input data like signals, pixels or text snippets into higher-level features before feeding them into GAMs to produce interpretable shape functions.

Furthermore, we refrained from performing extensive hyperparameter tuning for each model and dataset. To this end, it can be assumed that both the traditional models as well as the GAM-based approaches can achieve even higher prediction performance. However, since we worked mostly with default settings in both groups and avoided explicit optimizations, the comparison can be considered fair. In subsequent work, we will conduct further experiments on this open issue.

A last limitation concerns the evaluation of the model interpretability. As a first step, we tried to pursue an objective evaluation of the learned shape functions of interpretable models to derive a better understanding of the various models and their properties before moving on to a socio-technical investigation in the next step. For this purpose, we compared the outputs of different GAMs qualitatively, and derived statements about their merits and limitations. Thus, the evaluation of the subjective perception by model users, in which all GAMs with their characteristic outputs are evaluated in realistic decision-making scenarios, is currently missing. To this end, it is planned to conduct field experiments with data science experts and decision-makers from different domains and evaluate the usefulness of the different models using real-life prediction problems.

Another avenue for future research is to investigate under which circumstances one interpretable model should be preferred over another. To this end, our initial findings on the merits and limitations of the different GAM models provide a valuable starting point for identifying more specific selection criteria and deriving recommendations as to which model might be more appropriate in a given situation.

Appendix

Models	Python Implementations	Hyperparameter Setting
Splines	pyGAM[.LogisticGAM/.LinearGAM] (Servén, 2021)	Default
EBM	interpret.ExplainableBoosting[Classifier/Regressor] (Nori et al., 2021)	Interactions = 10
NAM	nam (R. Agarwal et al., 2021b)	Default
GAMI-Net	gaminet (Yang et al., 2021b)	Default
ExNN	exnn (Yang, 2021)	Default
LR	sklearn.linear_models[LogisticRegression/Ridge]	L2 regularization (logistic), ridge regression (linear)
DT	sklearn.tree.DecisionTree[Classifier/Regressor]	Max depth = 12
RF	sklearn.ensemble.RandomForest[Classifier/Regressor]	Max depth = 5, number of estimators = 100
GBM	sklearn.ensemble.GradientBoosting[Classifier/Regressor]	Max depth = 5, number of estimators = 100
XGB	xgboost.XGB[Classifier/Regressor]	max depth = 5, number of estimators = 100
MLP	sklearn.neural_network.MLP[Classifier/Regressor]	Number of hidden layers = 1, number of neurons = 40, number of epochs = 100, activation function = ReLU

Table 6. Overview of applied models with implementations and configurations.

References

Agarwal, N. and S. Das (2020). "Interpretable Machine Learning Tools: A Survey." In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1528–1534. DOI: 10.1109/SSCI47803.2020.9308260.

Agarwal, R., L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton (2021a). "Neural additive models: Interpretable machine learning with neural nets." *arXiv preprint arXiv:2004.13912*.

— (2021b). *NAM: Neural Additive Models - Interpretable Machine Learning with Neural Nets*. URL: https://github.com/google-research/google-research/tree/master/neural_additive_models.

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. d. F. Costa (2014). "A Systematic Comparison of Supervised Classifiers." *PLOS ONE* 9 (4), e94137. DOI: 10.1371/journal.pone.0094137.

Babic, B., S. Gerke, T. Evgeniou, and I. G. Cohen (2021). "Beware explanations from AI in health care." *Science* 373 (6552), 284–286. DOI: 10.1126/science.abg1834.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58, 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Berg, D. (2007). "Bankruptcy prediction by generalized additive models." *Applied Stochastic Models in Business and Industry* 23 (2), 129–143. DOI: 10.1002/asmb.658.

Bertoncelli, C. M., P. Altamura, E. R. Vieira, S. S. Iyengar, F. Solla, and D. Bertoncelli (2020). "PredictMed: A logistic regression–based model to predict health conditions in cerebral palsy." *Health Informatics Journal* 26 (3), 2105–2118. DOI: 10.1177/1460458219898568.

Breiman, L. (2001). "Random Forests." *Machine Learning* 45 (1), 5–32. DOI: 10.1023/A:1010933404324.

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730.

Chang, C.-H., S. Tan, B. Lengerich, A. Goldenberg, and R. Caruana (Aug. 2021). "How Interpretable and Trustworthy are GAMs?" en. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Virtual Event Singapore: ACM, pp. 95–105. ISBN: 978-1-4503-8332-5. DOI: 10.1145/3447548.3467453. URL: https://dl.acm.org/doi/10.1145/3447548.3467453 (visited on 08/18/2021).

Cortez, P. and A. Silva (2008). "Using Data Mining to Predict Secondary School Student Performance." In: *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*. Porto, Portugal: EUROSIS, p. 8.

DeVries, P. M. R., F. Viégas, M. Wattenberg, and B. J. Meade (2018). "Deep learning of aftershock patterns following large earthquakes." *Nature* 560 (7720). DOI: 10.1038/s41586-018-0438-y. (Visited on 10/23/2021).

Du, M., N. Liu, and X. Hu (2019). "Techniques for interpretable machine learning." *Communications of the ACM* 63 (1), 68–77. DOI: 10.1145/3359786.

Fanaee-T, H. and J. Gama (2014). "Event labeling combining ensemble detectors and background knowledge." *Progress in Artificial Intelligence* 2 (2), 113–127. DOI: 10.1007/s13748-013-0040-3.

FICO (2021). *Explainable Machine Learning Challenge*. URL: https://community.fico.com/s/explainable-machine-learning-challenge?tabset-158d9=3 (visited on 11/11/2021).

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics* 29 (5). DOI: 10.1214/aos/1013203451.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation." *Journal of Computational and Graphical Statistics* 24 (1), 44–65.

Hastie, T. and R. Tibshirani (1986). “Generalized Additive Models.” *Statistical Science* 1 (3). DOI: 10.1214/ss/1177013604.

Heinrich, K., P. Zschech, C. Janiesch, and M. Bonin (2021). “Process data properties matter: Introducing gated convolutional neural networks (GCNN) and key-value-predict attention networks (KVP) for next event prediction with deep learning.” *Decision Support Systems* 143, 113494. DOI: 10.1016/j.dss.2021.113494.

Hohman, F., A. Head, R. Caruana, R. DeLine, and S. M. Drucker (2019). “Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models.” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–13. DOI: 10.1145/3290605.3300809.

IBM (2019). *Telco customer churn*. https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

Janiesch, C., P. Zschech, and K. Heinrich (2021). “Machine learning and deep learning.” *Electronic Markets* 31 (3), 685–695. DOI: 10.1007/s12525-021-00475-2.

John-Mathews, J.-M. (2021). “A Critical Empirical Study of Black-box Explanations in AI.” In: *Proceedings of the 42nd International Conference on Information Systems (ICIS)*.

Jussupow, E., M. M. Martínez, A. Maedche, and A. Heinzl (2021). “Is This System Biased? – How Users React to Gender Bias in an Explainable AI System.” In: *Proceedings of the 42nd International Conference on Information Systems (ICIS)*.

Kaggle (2021a). *Airline Passenger Satisfaction*. URL: https://kaggle.com/teejmahal20/airline-passenger-satisfaction (visited on 11/11/2021).

— (2021b). *Stroke Prediction Dataset*. URL: https://kaggle.com/fedesoriano/stroke-prediction-dataset (visited on 11/11/2021).

— (2021c). *Telco Customer Churn*. URL: https://kaggle.com/blastchar/telco-customer-churn (visited on 11/11/2021).

— (2021d). *Water Quality*. URL: https://kaggle.com/adityakadiwal/water-potability (visited on 11/11/2021).

Kaur, H., H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan (2020). “Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, pp. 1–14. DOI: 10.1145/3313831.3376219.

Kibler, D., D. W. Aha, and M. K. Albert (1989). “Instance-based prediction of real-valued attributes.” *Computational Intelligence* 5 (2), 51–57. ISSN: 0824-7935, 1467-8640. DOI: 10.1111/j.1467-8640.1989.tb00315.x.

Kohavi, R. (1996). “Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.” In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. Portland, Oregon: AAAI Press, pp. 202–207.

Kraus, M. and S. Feuerriegel (2019). “Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences.” *Decision Support Systems* 125, 113100. DOI: 10.1016/j.dss.2019.113100.

Kraus, M., S. Feuerriegel, and A. Oztekin (2020). “Deep learning in business analytics and operations research: Models, applications and managerial implications.” en. *European Journal of Operational Research* 281 (3), 628–641. DOI: 10.1016/j.ejor.2019.09.018.

Lipton, Z. C. (2018). “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue* 16 (3), 31–57. DOI: 10.1145/3236386.3241340.

Lou, Y., R. Caruana, and J. Gehrke (2012). “Intelligible models for classification and regression.” In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12*. Beijing, China: ACM Press, p. 150. DOI: 10.1145/2339530.2339556.

Lou, Y., R. Caruana, J. Gehrke, and G. Hooker (2013). "Accurate intelligible models with pairwise interactions." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. Chicago Illinois USA: ACM, pp. 623–631. DOI: 10.1145/2487575.2487579.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020). "From local explanations to global understanding with explainable AI for trees." *Nature Machine Intelligence* 2 (1), 56–67. DOI: 10.1038/s42256-019-0138-9.

McKinney, S. M., M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty (2020). "International evaluation of an AI system for breast cancer screening." *Nature* 577 (7788), 89–94. DOI: 10.1038/s41586-019-1799-6.

Mehdiyev, N. and P. Fettke (2021). "Local Post-hoc Explanations for Predictive Process Monitoring in Manufacturing." In: *Proceedings of the 29th European Conference on Information Systems (ECIS)*.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6), 115:1–115:35. DOI: 10.1145/3457607.

Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267, 1–38. DOI: 10.1016/j.artint.2018.07.007.

Moro, S., P. Cortez, and P. Rita (2014). "A data-driven approach to predict the success of bank telemarketing." *Decision Support Systems* 62, 22–31. ISSN: 0167-9236.

Murase, H., H. Nagashima, S. Yonezaki, R. Matsukura, and T. Kitakado (2009). "Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in Sendai Bay, Japan." *ICES Journal of Marine Science* 66 (6), 1417–1424. DOI: 10.1093/icesjms/fsp105.

Nori, H., S. Jenkins, P. Koch, and R. Caruana (2019). "InterpretML: A Unified Framework for Machine Learning Interpretability." *arXiv:1909.09223 [cs, stat]*.

— (2021). *InterpretML - Alpha Release*. URL: https://github.com/interpretml/interpret.

Pace, K. R. and R. Barry (1997). "Sparse spatial autoregressions." *Statistics & Probability Letters* 33 (3), 291–297. DOI: 10.1016/S0167-7152(96)00140-X.

Pierrot, A. and Y. Goude (2011). "Short-Term Electricity Load Forecasting With Generalized Additive Models." In: *Proceedings of the 16th Intelligent System Applications to Power Systems Conference, ISAP*. IEEE, pp. 410–415.

Redmond, M. and A. Baveja (2002). "A data-driven software tool for enabling cooperative information sharing among police departments." *European Journal of Operational Research* 141 (3), 660–678. DOI: 10.1016/S0377-2217(01)00264-8.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "”Why Should I Trust You?”: Explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. San Francisco, California, USA: ACM Press, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

Roy, A., S. Qureshi, K. Pande, D. Nair, K. Gairola, P. Jain, S. Singh, K. Sharma, A. Jagadale, Y.-Y. Lin, S. Sharma, R. Gotety, Y. Zhang, J. Tang, T. Mehta, H. Sindhanuru, N. Okafor, S. Das, C. N. Gopal, S. B. Rudraraju, and A. V. Kakarlapudi (2019). "Performance Comparison of Machine Learning Platforms." *INFORMS Journal on Computing* 31 (2), 207–225. DOI: 10.1287/ijoc.2018.0825.

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1 (5), 206–215.

Rudin, C. and J. Radin (2019). "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition." *Harvard Data Science Review* 1 (2). DOI: 10.1162/99608f92.5a8a3a3d.

S&P (2021). *S&P Letters Data: California Housing*. URL: `https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html` (visited on 11/11/2021).

Schemmer, M., N. Kühl, and G. Satzger (2021). "Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Work Through Explainable Artificial Intelligence." *arXiv:2109.13827 [cs]*.

Servén, D. (2021). *pyGAM*. URL: `https://github.com/dswah/pyGAM`.

Shipe, M. E., S. A. Deppen, F. Farjah, and E. L. Grogan (2019). "Developing prediction models for clinical use using logistic regression: an overview." *Journal of Thoracic Disease* 11 (S4), S574–S584. DOI: `10.21037/jtd.2019.01.25`.

Stierle, M., J. Brunk, S. Weinzierl, S. Zilker, M. Matzner, and J. Becker (2021a). "Bringing Light Into the Darkness - A Systematic Literature Review on Explainable Predictive Business Process Monitoring Techniques." In: *Proceedings of the 29th European Conference on Information Systems (ECIS)*.

Stierle, M., S. Weinzierl, M. Harl, and M. Matzner (May 2021b). "A technique for determining relevance scores of process activities using graph-based neural networks." en. *Decision Support Systems* 144, 113511. ISSN: 01679236. DOI: `10.1016/j.dss.2021.113511`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S016792362100021X` (visited on 07/26/2021).

Sudjianto, A., W. Knauth, R. Singh, Z. Yang, and A. Zhang (2020). "Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification." *arXiv:2011.04041 [cs, stat]*. arXiv: 2011.04041.

Thiebes, S., S. Lins, and A. Sunyaev (2021). "Trustworthy artificial intelligence." *Electronic Markets* 31 (2), 447–464. DOI: `10.1007/s12525-020-00441-4`.

Tomić, N. and S. Božić (2014). "A modified Geosite Assessment Model (M-GAM) and its Application on the Lazar Canyon area (Serbia)." *International Journal of Environmental Research* 8 (4). DOI: `10.22059/ijer.2014.798`.

UCI (2021a). *UCI Machine Learning Repository: Adult Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/adult` (visited on 11/11/2021).

— (2021b). *UCI Machine Learning Repository: Automobile Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/automobile` (visited on 11/11/2021).

— (2021c). *UCI Machine Learning Repository: Bank Marketing Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing` (visited on 11/11/2021).

— (2021d). *UCI Machine Learning Repository: Bike Sharing Dataset Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset` (visited on 11/11/2021).

— (2021e). *UCI Machine Learning Repository: Communities and Crime Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime` (visited on 11/11/2021).

— (2021f). *UCI Machine Learning Repository: Student Performance Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Student+Performance` (visited on 11/11/2021).

Valaskova, K., T. Kliestik, L. Svabova, and P. Adamko (2018). "Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis." *Sustainability* 10 (7), 2144. DOI: `10.3390/su10072144`.

Vaughan, J., A. Sudjianto, E. Brahimi, J. Chen, and V. N. Nair (2018). "Explainable Neural Networks based on Additive Index Models." *arXiv:1806.01933 [cs, stat]*.

Wanner, J., K. Heinrich, C. Janiesch, and P. Zschech (2020a). "How Much AI Do You Require? Decision Factors for Adopting AI Technology." In: *Proceedings of the 41st International Conference on Information Systems (ICIS)*.

Wanner, J., L.-V. Herm, K. Heinrich, C. Janiesch, and P. Zschech (2020b). "White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems." In: *Proceedings of the 41st International Conference on Information Systems (ICIS)*.

Wastensteiner, J., T. M. Weiss, F. Haag, and K. Hopf (2021). "Explainable AI for tailored electricity consumption feedback–An experimental evaluation of visualizations." In: *Proceedings of the 29th European Conference on Information Systems (ECIS)*.

Yang, Z. (2021). *Enhanced Explainable-Neural-Networks*. URL: `https://github.com/ZebinYang/exnn`.

Yang, Z., A. Zhang, and A. Sudjianto (2021a). "Enhancing Explainability of Neural Networks Through Architecture Constraints." *IEEE Transactions on Neural Networks and Learning Systems* 32 (6), 2610–2621. DOI: `10.1109/TNNLS.2020.3007259`.

— (2021b). *GAMI-Net*. URL: `https://github.com/ZebinYang/gaminet`.

— (2021c). "GAMI-Net: An explainable neural network based on generalized additive models with structured interactions." *Pattern Recognition* 120, 108192.

Zhang, X., Q. Du, and Z. Zhang (2020). "An Explainable Machine Learning Framework for Fake Financial News Detection." In: *Proceedings of the 41st International Conference on Information Systems (ICIS)*.

Zinovyeva, E., W. K. Härdle, and S. Lessmann (2020). "Antisocial online behavior detection using deep learning." *Decision Support Systems* 138, 113362. DOI: `10.1016/j.dss.2020.113362`.

Zschech, P., K. Heinrich, R. Bink, and J. S. Neufeld (2019). "Prognostic Model Development with Missing Labels: A Condition-Based Maintenance Approach Using Machine Learning." *Business & Information Systems Engineering* 61 (3), 327–343. DOI: `10.1007/s12599-019-00596-1`.