Explainable Artificial Intelligence for Interpretable Data Minimization

Maximilian Becker
Vision and Fusion Laboratory
Karlsruhe Institute of Technology
Karlsruhe, Germany

Emrah Toprak
Vision and Fusion Laboratory
Karlsruhe Institute of Technology
Karlsruhe, Germany

Jürgen Beyerer
Fraunhofer IOSB
Karlsruhe, Germany

***Abstract*—Black box models such as deep neural networks are increasingly being deployed in high-stakes fields, including justice, health, and finance. Furthermore, they require a huge amount of data, and such data often contains personal information. However, the principle of data minimization in the European Union's General Data Protection Regulation requires collecting only the data that is essential to fulfilling a particular purpose. Implementing data minimization for black box models can be difficult because it involves identifying the minimum set of variables that are relevant to the model's prediction, which may not be apparent without access to the model's inner workings. In addition, users are often reluctant to share all their personal information. We propose an interactive system to reduce the amount of personal data by determining the minimal set of features required for a correct prediction using explainable artificial intelligence techniques. Our proposed method can inform the user whether the provided variables contain enough information for the model to make accurate predictions or if additional variables are necessary. This human-centered approach can enable providers to minimize the amount of personal data collected for analysis and may increase the user's trust and acceptance of the system.**

***Index Terms*—XAI, Data Minimization, Counterfactual Explanations**

I. Introduction

Complex machine learning algorithms are used more and more because of the complexity of the problems that need to be solved [1]. These models act as black boxes and providing a clear explanation of the models' decisions and reasoning behind them is difficult [1]. Due to their high accuracy, black box models are increasingly being deployed in high-stakes fields, ranging from determining bail amounts [2] to the diagnosis and treatment of patients [3]. They are also prevalent in business activities, as prediction algorithms are considered valuable business assets, and companies want to keep them secret from users [4].

A significant limitation of complex machine learning models is their reliance on large amounts of data to make accurate predictions [5]. Therefore, a huge amount of data needs to be collected, processed, and stored. Because such data often involves personal and private information, concerns about data privacy and security have become increasingly important. To address these concerns, the use of personal data is strictly regulated by data protection laws, such as the European Union's General Data Protection Regulation (GDPR). One of the key data protection principles of the GDPR is data minimization, which allows the collection and processing of only as much data as absolutely necessary for a specified purpose [6].

The UK Information Commissioner's Office [7] distinguishes between data minimization methods used in the training phase, meaning the period during which the model is trained, and the inference phase, meaning the period during which the model is used to make a prediction. In this work, we will focus on reducing the amount of input data in a black box setting during the inference phase while preserving the accuracy as closely as possible to the accuracy of the original model. However, implementing data minimization without access to the model's internals can be challenging, as it requires precisely determining the effect of each input feature on each prediction or classification.

This is where Explainable Artificial Intelligence (XAI), particularly counterfactual explanations, can play a crucial role. In our approach, we use counterfactual explanations to identify the minimum set of features required for each individual accurate prediction while minimizing the collection and processing of input data.

Users may not want to share all their data. Therefore, we propose an interactive system, which forms an important component of many modern applications in industry and production. Based on the input data provided, our proposed method can explain to the user whether the provided data is sufficient for the model or which additional variables should be provided to make an accurate prediction. Besides technical and legal aspects of data minimization, such as storage and cost reduction, low liability [8], fairness, and privacy, this may also increase the trust and acceptance of the system, because the lack of transparency in machine learning models can undermine users' trust, especially in situations where the consequences of predictions are serious and may result in the rejection of the system [9].

As a result, we suggest a human-computer interaction system for data minimization that can reduce the amount of input data used to make predictions by machine learning models. Our proposed method focuses on minimizing the new data collected for analysis, and it does not require training data or model internals. Notably, the black box model does not need to be retrained, making it an efficient approach for data

minimization in existing systems.

To sum up, we make the following contributions:

- A new interactive system that utilizes counterfactual explanations to gather only the required user data, in accordance with the GDPR's data minimization principle, while maintaining accuracy.
- An investigation into the relationship between SHAP values and the frequency of occurrence of features in the minimal set of required features for the model to make accurate predictions.

The remainder of this work is structured as follows: Section II goes over important fundamentals in XAI and the legal concept of data minimization. Section III covers related work in data minimization, sections IV and V explain our proposed method for interpretable data minimization and the experiments conducted to evaluate our approach. Lastly, section VI gives a conclusion and an outlook on possible future work.

II. Fundamentals

With the advances in AI, increasingly more complicated models are developed to solve complex problems. However, the inherent nature of their architecture makes it harder to fully understand them [10]. Nevertheless, complex models often perform better than traditional AI techniques, such as linear models or trees, because of inherent problems of high bias or high variance [11]. The performance of a model is not always the most important aspect but explainability also plays a very important role in some domains [1]. The need for explainability is demanding if there is an "incompleteness" in the formalization of a problem [12]. Incompleteness refers to certain aspects of problems that cannot be adequately encoded into the system. It can stem from various sources, including ethics, safety, gaining new knowledge, and mismatched objectives [12].

In recent years, several XAI (Explainable Artificial Intelligence) methods have been proposed to facilitate users' understanding of machine learning and algorithms' decision-making processes. One well-known family of XAI methods is attribution-based explanations, such as SHAP [13] and LIME [14], which assign an importance score to each attribute. Another widely recognized family of methods is counterfactual explanations. Counterfactual explanations, such as the method by Wachter et al. [15], serve as an interpretation method that demonstrates the required changes to flip the model prediction to the desired output (a "what-if" explanation) by making the smallest perturbation to a data instance.

However, counterfactual explanations often do not comply with feature importance scores [16]. Mothilal et al. [16] show the relationship between them and find that attribution-based explanations emphasize the sufficiency of feature values for a given model, whereas counterfactual explanations highlight the necessity of feature values. However, according to the study, a good explanation should satisfy both of them.

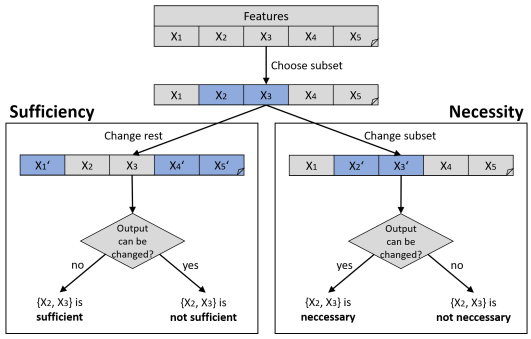


Fig. 1: Finding sufficiency and necessity of features

A. Sufficiency and Necessity of Features

Sufficiency may be rewritten as $\neg y \rightarrow \neg x$ and necessity as $\neg x \rightarrow \neg y$ [17]. According to Mothilal [16] sufficiency and necessity are two concepts that describe how important a set of features is for the decision of a model. Suppose we have an instance $X = \{x_1, x_2, x_3, \ldots, x_n\}$ and $f(X) = y^*$ is the output of the model f. $\{x_2, x_3\}$ represents the subset of features from X we want to investigate as shown in Figure 1. The sufficiency and necessity of a set of features can be determined by searching for counterfactuals. In our case, a counterfactual simply refers to changes to some features of an instance that lead to a changed prediction of the model f for that instance. It is also possible to constrain the criteria further by specifying the value ranges of each feature. For our work we determine the sufficiency and necessity of features as follows:

Sufficiency: By keeping the values of $\{x_2, x_3\}$ fixed and altering only other features, we can determine sufficiency. If no counterfactuals can be generated, then the set $\{x_2, x_3\}$ is sufficient for the model's output y^*.

Necessity: If counterfactuals can be generated by changing only the feature set $\{x_2, x_3\}$, then the set is necessary for the model's output y^*.

As a result, sufficiency refers to the ability of a feature subset to consistently generate a specific model output, regardless of the values of other features. On the other hand, the subset of features is necessary if changing the values of the subset leads to a change in the model's output [16].

B. A Legal Perspective

Data minimization is one of the key data protection principles of the European Union's GDPR [18]. In Article 5(1)(c), the GDPR [18] states that "personal data shall be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed". Firstly, only relevant personal data should be processed, and it does not allow for the processing of irrelevant details. The purpose of this principle is to prevent the collection of excessive data for constantly redefined and undisclosed purposes [19]. Secondly, it requires that personal data be adequate. The adequacy requirement may demand more data to be processed in some

cases, particularly where available data are insufficient to make inferences about underrepresented groups in the dataset, such as people with disabilities [8]. This principle is closely related to the fairness, transparency, and accuracy principles of the GDPR [19]. Finally, personal data should be limited, requiring only the minimum amount of personal information necessary to achieve the purpose [7]. Indeed, Article 25(2) of the GDPR [18] requires controllers to implement "appropriate technical and organizational measures" to ensure that only personal data necessary for each specific purpose of processing is processed.

Data minimization does not only cover the minimization of processed data but also the pseudonymization of personal data. The use of pseudonymization enhances the protection of data subjects' identity [20].

As a result, data minimization in either way promotes the development of machine learning by protecting data subjects and maintaining overall trust in models [20].

III. Related Work

There are many algorithmic techniques for data minimization. The Norwegian Data Protection Authority [20] distinguishes between breadth-based data minimization and depth-based data minimization. The former aims to minimize the number of features by excluding redundant and irrelevant features, while the latter tries to minimize the overall amount of data collected for the specific purpose.

Several algorithmic techniques have been developed to improve the quality of models and reduce the costs associated with data gathering. These techniques indirectly minimize data by utilizing outlier detection to identify and eliminate rare anomalies and noise, feature selection to remove redundant and irrelevant features, and active learning to gradually select data to be labeled or added to a model [21]. Additionally, feature abstraction can be used to reduce the dimensionality of input data by extracting the most important features while discarding irrelevant or redundant ones [22]. However, this technique is also employed with the goal of improving the model's performance, rather than solely for data minimization.

Various papers aim to minimize the number of features. Heuer and Breiter [23] explore the trade-off between big data and data minimization and demonstrate that it is possible to construct a successful prediction model for predicting student success without including private attributes such as gender, disability status, or the highest educational level. In addition, they also binarized the daily activity of clicks by encoding only whether a student was active or not and argued that the model with this simplification has better predictive performance than actual counts of clicks. Biega et al. [19] conduct an empirical study to check whether the original recommender performance can be preserved while limiting the number of known user ratings. The UK information commissioner's office's guidance on the AI auditing framework [7] suggests privacy-preserving methods such as perturbing or adding "noise" to the data in such a way that preserves structures of those features for data minimization in the training stage. Adding noise is a way of designing differential privacy. For example, consider an algorithm that examines a dataset and generates statistics such as mean, mode, median, etc. Such an algorithm is referred to as differentially private if the output does not reveal whether a specific individual's data was used in the original dataset or not [24].

Some papers focus on limiting overall data collection during model training. Hestness et al. [25] divide the learning curve, which shows the relationship between training data size and loss, into three phases. The phases are: (1) The small data region, where the data size is not representative enough and the performance is poor. (2) The power-law region, where a new training data set increases the performance of the model. (3) The irreducible error region, where a new data set is not able to improve the performance. Tae and Whang [26] adopt the power law relationship assumption during the data acquisition process and suggest a selective data acquisition framework that uses the learning curve to determine the optimal amount of data to acquire for each class in order to achieve similar error rates across all classes by optimizing model accuracy and fairness for each class. Similarly, Shanmugam et al. [8] also propose a framework to limit data collection. On the other hand, the framework offers a data collection stopping point by continuously updating an estimate of the learning curve during the data acquisition process. They also suggest random feature acquisition techniques instead of active feature acquisition techniques for data minimization because of the smoother learning curve, the success of data collection depending on initial conditions, and the potential excessive burden of data collection on certain users.

Our approach is closely related to papers that deal with limiting data collection at the inference stage. Since prediction algorithms are often valuable assets for businesses, they operate in a black box setting [4], which means that it is not known how the prediction model works, or one does not have access to the model's internals. Moreover, similar to our approach, they are only concerned with minimizing newly collected data for analysis or testing whether a black box model complies with the principle of data minimization as in our approach. Rastegarpanah et al. [4] proposed an audit method that uses feature imputation across all prediction instances to determine whether each of the input features employed in a particular model is essential for maintaining the desired level of predictive performance. The main difference to our work is that their approach evaluates the importance of features on a global level while we evaluate the importance of features for each instance given by a user. Goldsteen et al. [27] use a data generalization technique based on knowledge distillation, where a surrogate model such as a decision tree is trained to determine decision boundaries of the model, to suppress or generalize input features in classification while keeping the performance of the model at the desired level. As a result, it reduces the number and/or granularity of features collected for analysis and increases privacy protection [27]. This approach is functionally the most similar to ours but uses knowledge distillation rather than XAI techniques. However the user still has to provide all data and only some of the data

is generalized into equivalence classes.

IV. Interpretable Data Minimization

What if the user could choose to only provide the data that is necessary and sufficient for the model? Can we validate if the given information is enough for the model's intended purpose? For instance, can we ascertain if the data provided by the user is adequate to predict a disease? In case some data is missing, can we determine which features among the missing values are important to make an accurate prediction?

We propose a method that can reduce the amount of input data required for individual predictions made by machine learning models. Our key idea is to identify the input features that are necessary and sufficient for the prediction of a black box model. To achieve this, we use counterfactual explanations, as discussed in section II-A. Our focus is on minimizing the amount of newly collected data required during the inference stage, and we assume a black box setting where we cannot access the model's internals and have no access to the original training data. Therefore, our method can be applied to any machine learning algorithm without requiring any changes to the original model.

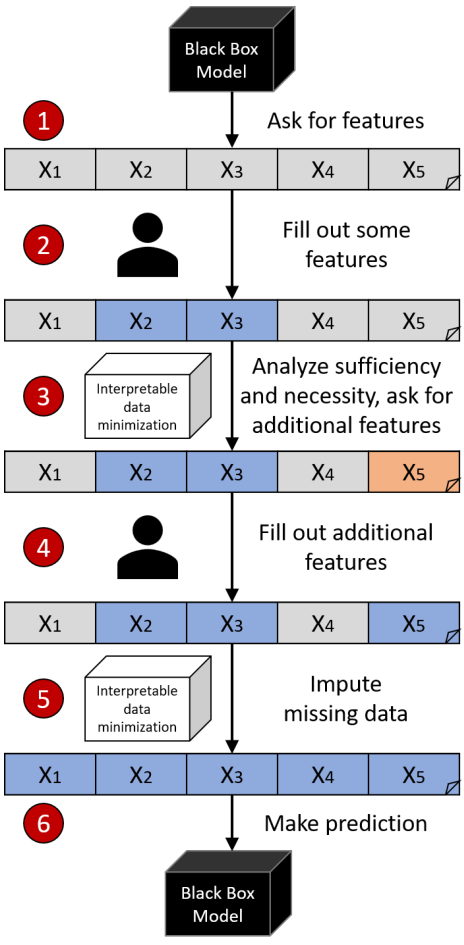


Fig. 2: Procedure of user interaction

The process we envision has 6 steps and is depicted in Figure 2:

1) The black box model requests the features it requires for its prediction from the user.
2) The user provides the features they are most comfortable with sharing.
3) Our system calculates the smallest set of sufficient and necessary features based on the users input and asks for additional features if needed.
4) The user fills in the missing features in the set of sufficient and necessary features.
5) Our system imputes the missing features.
6) The black box model makes its prediction using the features provided by the user and those imputed by our system.

The concepts of sufficiency and necessity were introduced in section II-A. Determining the sufficiency and necessity of features is done through the DiCE (Diverse Counterfactual Explanations) algorithm, which is proposed by Mothilal et al. [28] to generate counterfactual explanations for any machine learning model. For each subset of the feature set, if the model's output for an instance can be changed by changing only the values in that subset, it means that the features in that subset are necessary for the prediction of that instance. On the other hand, if the model's output for an instance cannot be changed by changing only values outside that subset, it shows that the features in the subset are sufficient for the prediction of that instance.

However, it may take a very long time if the number of features is high because this process is performed for all possible combinations of features for each instance. To decrease the cost of the processing, the process is optimized based on the idea that if one subset of features can change the output of the model, then every superset of the subset can also change the output of the model which means they don't have to be tested anymore.

After finding both sufficient and necessary features for each instance from the test dataset, the minimal subsets for sufficient and necessary features are extracted. For example, if A and B are two subsets for sufficient and necessary features for the same instance, and A is a subset of B, then only A is selected for the sufficient and necessary features.

Assuming we have received data from a user, but some values are missing, we can employ an imputation technique to fill in these missing values. For example, we can assign random values to impute the missing values. Imputation is necessary because most machine learning models are incapable of handling data with missing values.

If the provided features are sufficient and necessary, meaning that the missing values cannot change the prediction, the system confirms this to the user; however, if they are not, it informs the user about the additional features required. It's as if it dynamically specifies fields that need to be filled by users based on the provided data. Additionally, we can store only those adequate values provided by the user.

V. Evaluation

As illustrated in Figure 3, we evaluate our approach by following the following steps for different datasets:

1) We compare the performance of different models and select the best one that can generalize well.
2) We identify sufficient and necessary features using DiCE [28].
3) For the input features that are not sufficient and necessary, we remove values to simulate missing user data.
4) We impute the removed values using the MiceForest[1] imputation technique.
5) We evaluate our method by comparing the output of the model for the imputed test dataset with the output for the original test dataset.

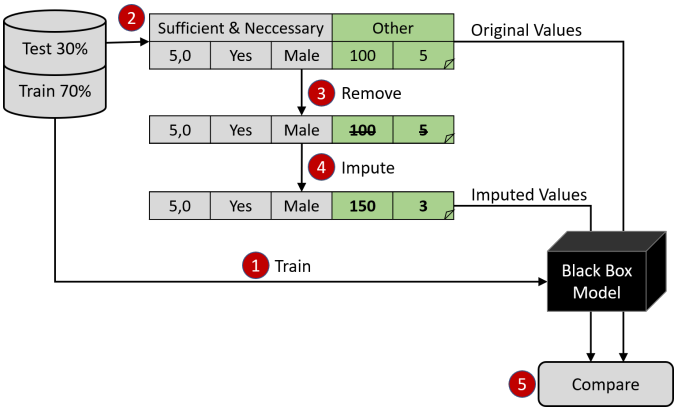


Fig. 3: Procedure of the experiment

We use three common datasets to evaluate our proposed method: German Credit [2], Adult [3], and Stroke [4]. These datasets are popular benchmarks and contain data from different fields.

German Credit: This dataset, from the UC Irvine Machine Learning repository [29], consists of 1000 observations (rows) with 21 features (columns), regarding people who took loans from a bank. Each person is classified as a good or bad credit risk according to the set of features.

Adult: This dataset from the UC Irvine Machine Learning repository [29] is based on the 1994 Census database and contains 48842 observations (rows) with 15 features (columns). The task is to predict whether an individual's annual income exceeds $50.000 using the variables in this data set.

Stroke: This dataset from Kaggle contains 5110 observations (rows) with 12 features (columns). Each observation corresponds to one patient and the attributes are variables related to the health status of each patient. These features are used to predict whether a patient is likely to have a stroke. However, out of the 916 participants who are 18 years old or younger, only two have experienced a stroke, resulting in highly imbalanced data. For our analysis, we will focus on participants who are over 18 years old, following the approach taken in prior work [30]. Consequently, the dataset with participants over 18 years old contains 4193 observations.

A. Data preprocessing and Training

We split each dataset into a training set of 70% and a test set of 30% using the stratified sampling method. Then we apply several types of classifiers to each dataset. The top three performing models for each dataset are shown in Table I. Additionally, more information about data preprocessing and model training can be found in the Appendix A.

B. Finding Sufficiency and Necessity of Features

As introduced before, in this step, the sufficiency and necessity of the feature values for each dataset are determined through the DiCE algorithm [28]. DiCE was configured to search for only one counterfactual example. For each subset, the necessary and sufficient features for each instance from the test dataset are determined.

C. Simulating by removing values

It is assumed that a user provides features that satisfy the sufficiency and necessity criteria, otherwise, the user would be asked to provide the missing values. To simulate this situation values other than those that provide sufficient and necessary features for each sample are replaced with NaN to indicate missing values. The majority of the instances have more than one subset of features that satisfies the sufficiency and necessity properties.

In Figures 4a, 4b, and 4c, the blue bar represents how often a feature was in the sets of sufficient and necessary features across all instances in the training dataset, while the yellow bar indicates how often the feature was absent in these sets. For instance, in Figure 4b the feature "ever_married" appears for 2050 out of the total 4079 instances in the sets of necessary and sufficient variables. This feature is considered important for the model's output in half of the cases, while it is considered insignificant in the other half. The variables "checking_status" in Figure 4a and "age" in Figure 4b are present in nearly all sets of necessary and sufficient variables.

D. Imputation

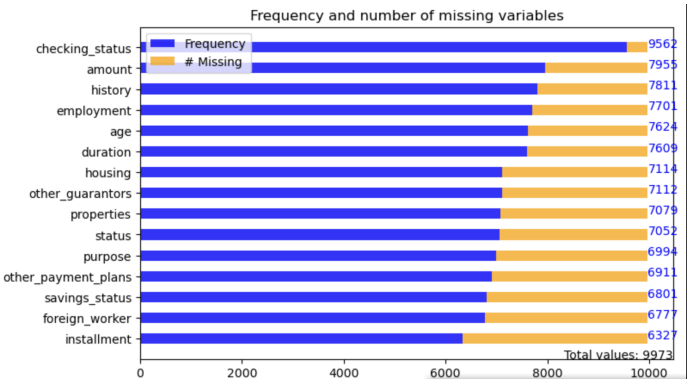
For the evaluation we utilized the MiceForest imputation technique [5] to fill in these missing values. MiceForest is an extension of the Multivariate Imputation by Chained Equations (MICE) algorithm. It employs an iterative process that models each variable as a function of the other variables to impute missing values in the dataset. The imputation is performed by training random forests, with each model trained on a different subset of the data, using the LightGBM model as the chaining function [31]. This technique is capable of handling both continuous and categorical features [31]. It was selected after comparing different imputation techniques because it performed the best.

[1]https://github.com/AnotherSamWilson/miceforest
[2]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[3]https://archive.ics.uci.edu/ml/datasets/adult
[4]https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
[5]https://github.com/AnotherSamWilson/miceforest

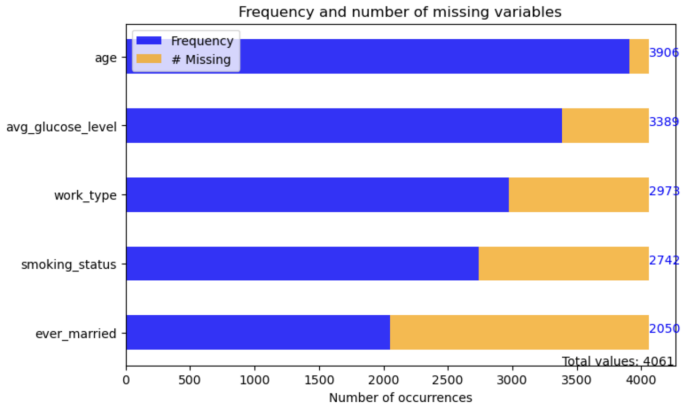
	German Credit			Stroke			Adult		
	XGB*	**SVM**	**RF**	**RF**	**Stacking***	**MLP**	**SVC**	**MLP**	**Gradient***
Accuracy	0.893	0.843	0.870	0.919	0.922	0.803	0.822	0.828	0.830
Precision WA	0.90	0.84	0.87	0.92	0.92	0.80	0.83	0.83	0.83
Recall WA	0.89	0.84	0.87	0.92	0.92	0.80	0.82	0.83	0.83
F1-score WA	0.89	0.84	0.86	0.92	0.92	0.80	0.83	0.83	0.83

TABLE I: Accuracy of Models on Datasets

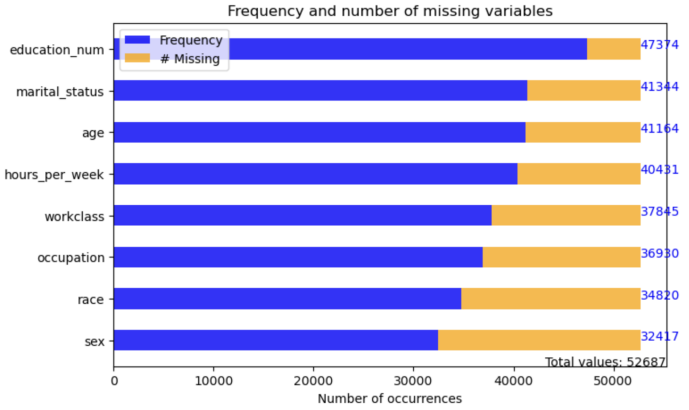
Note: Several classifiers were tested on all datasets but only the 3 best-performing are shown in the table. The classifiers selected for evaluation are indicated by an asterisk (*), Random Forest (RF), Weighted Average (WA)



(a) German Credit



(b) Stroke



(c) Adult

Fig. 4: The number of sufficient and necessary features in the datasets

E. Evaluating the Performance

We evaluate the performance of the model on each imputed dataset by comparing it to the model's performance on the original dataset. This was done by comparing the model's output on the original and imputed dataset. Moreover, our approach does not involve retraining the model to predict the imputed dataset. In all three datasets, the relative performance is 100%, which means that the model predicts exactly the same outputs on the imputed and original dataset. As a result, we can predict a dataset with missing values for all variables except the necessary and sufficient variables, with the same performance as if all variables were available. In cases where only sufficient and necessary features for the test datasets are provided, the model exhibits identical performance as shown in Table I. This implies that if we have the necessary and sufficient features, we can predict the dataset as if all variables are available.

	German Credit	Stroke	Adult
Relative performance	1.0	1.0	1.0
Number of instances	9973	4061	52687

TABLE II: Performance and number of instances for three datasets

F. Relation between SHAP Values and Frequency of Sufficiency and Necessity

SHAP is a variety of attribution explanations that answer the question: what are the most important features that contributed to a prediction [32]? SHAP was proposed by Lundberg and Lee [13] to explain the local predictions of any machine learning model by computing the contribution of each feature to the model's output.

	German Credit	Stroke	Adult
Pearson correlation	0.8153	0.8673	0.9061
p-value	0.0002	0.0568	0.0019

TABLE III: Pearson correlation values and p-values

We calculate the frequency of each feature occurring in the sets of sufficient and necessary features and then compute the Pearson correlation between the SHAP feature importance values and the frequency of each feature's occurrence. To calculate the SHAP feature importance values, we take the average of the absolute SHAP values for each feature across the entire dataset [33]. We use the training datasets as the background dataset for SHAP values, except for the Adult dataset, where we randomly selected 3000 training samples

due to the extremely slow run times that would be caused by using all training data samples. In table III, we describe the approximate correlation values and p-values, which indicate a strong positive correlation between these two values being measured for the German Credit and Adult datasets, but not for the Stroke dataset.

VI. Conclusion

XAI methods are generally used to increase transparency and fairness, which are key principles of the GDPR. However, in this work, we propose a new method that uses XAI to address another key principle of the GDPR, which is data minimization for machine learning models. We assume that the machine learning model is a black box, and we do not require any training data. Furthermore, the machine learning model does not need to be retrained. Therefore, our approach can be applied to any dataset and any machine learning model. As a result, the method described in this paper does not affect the training data or the trained model; rather, it only affects new data collected for analysis.

Our method utilizes counterfactual explanations to identify the sufficient and necessary features required for making an accurate prediction for each instance. With the three test datasets, we demonstrated that accurate predictions can be made even if only the sufficient and necessary features are provided, as if all variables were available. Although we achieved good results in terms of data minimization without compromising accuracy, the impact of our method on model accuracy may be minimal, depending on the counterfactual models and hyperparameters used.

Moreover, users may not want to share all of their data when providing it to a system. Our new approach proposes a human-computer interaction during the data input process. By using sufficient and necessary features, the system can explain to the user how their decision to share certain data or not influences the system's behavior. For instance, the user can be notified that the data they have entered is sufficient for predicting a cancer diagnosis, or what other data is required for an accurate prediction. This approach could increase the trust and acceptance of interactive systems, as users will have a better understanding of how their data is being used.

The main limitation of this work is that the time required to find the sufficient and necessary features can grow exponentially as the difference between the number of features provided by the user and the number of features in the set that contains sufficient and necessary features increases, especially if the number of given features is less than the number of features in this set.

SHAP is one of the most popular frameworks that explain individual predictions by computing the contribution of each feature to the prediction. We examined the relationship between the SHAP feature importance and the frequency of occurrence among sufficient and necessary features and found that there is a strong positive correlation between them for German Credit and Adult datasets, but not for the Stroke dataset.

A. Future Work

Our method can be applied to any machine learning model, as we assume that the model is a black box and we do not have access to the training data. However, the experiments presented in this thesis only focus on classification models. Therefore, other types of models, such as regression or generalization models, could be utilized to evaluate the performance of the model with only the sufficient and necessary features available.

Our approach does not affect training data; rather, it considers data minimization only for the new data collected for analysis. A promising area for future research could be exploring data minimization techniques for training machine learning models, using a similar approach to our model.

As mentioned previously the time needed to find the sufficient and necessary features is the main limitation of this approach. In section V-F we showed that there is a strong correlation between sufficient and necessary features and their SHAP values. This correlation could be used to accelerate our approach by prioritizing features based on their SHAP values.

Our method utilizes sufficient and necessary features to provide feedback to the user on how their decision to share certain data or not influences the system's behavior. However, we do not know how this feedback will affect the user's behavior in sharing data. A future study that investigates the impact of our proposed model on human behavior would also be valuable.

Acknowledgment

This work was supported by funding from the topic Engineering Secure Systems of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

Appendix

A. Data Preprocessing and Training

We processed the datasets and trained the models using the following techniques:

1) German Credit:

- **Preprocessing**
  - Features used from the dataset: *checking_status, history, purpose,*
    savings_status, employment_status, properties, other guarantors,
    other_payment_plans, housing, and *foreign_worker*
  - Features that were encoded through ordinal encoding: *foreign_worker* and *employment*
  - Features that were encoded through one-hot encoding: *checking_status, history, purpose, savings_status, status, properties, other_guarantors,*
    other_payment_plans, and *housing*
  - Data normalization was applied to the continuous features.
  - The *SMOTEENN*[6] method was applied to the training dataset.

[6]https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTEENN.html

- **Training**
  - We trained three types of classifiers: eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), and Random Forest (RF). The settings for these classifiers were as follows:
    - **XGB:** The maximum tree depth was set to *10*, the learning rate was set to *0.01*, the number of estimators was set to *200* and the evaluation metric was set to *logistic loss*.
    - **SVM:** The penalty parameter of the error term was set to *2* and the probability parameter was set to *true*.
    - **RF:** The maximum depth of the tree was set to *10*.

2) Stroke:

- **Preprocessing**
  - Features used in the dataset: *age, average_glucose_level, ever_married, work_type,* and *smoking_status*
  - A sample whose gender is listed as *other* was deleted
  - Features that were encoded through ordinal encoding: *ever_married*
  - Features that were encoded through one-hot encoding: *work_type* and *smoking_status*
  - Data standardization was applied to the continuous features.
  - The *SMOTEENN* method was applied to the dataset like in [34].
- **Training**
  - We trained three types of classifiers: RF, Stacking, and Multi-layer Perceptron classifier (MLP). These classifiers were configured as in the previous work [34].

3) Adult:

- **Preprocessing**
  - We process the dataset using techniques proposed by Zhu [30]. As a result, we obtained eight features: *age, education_num, hours_per_week, sex, workclass, marital_status, occupation,* and *race*
  - Features that were encoded through ordinal encoding: *sex*
  - Features that were encoded through one-hot encoding: *workclass, marital_status, occupation,* and *race*
  - Data standardization was applied to the continuous features.
  - The *SMOTEENN* method was applied to the training dataset.
- **Training**
  - We trained three types of classifiers: SVC, MLP, and Gradient Boosting Classifier (Gradient). The settings for these classifiers were as follows:
    - **SVC:** The probability parameter was set to *true*.
    - **MLP:** The hidden layer size was set to *10*, the maximum number of iterations was set to *500*, the solver for weight optimization was set to *stochastic gradient descent* and the initial learning rate was set to *0.3*.
    - **Gradient:** The default parameters of scikit-learn[7] were used.

REFERENCES

[1] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: Risk assessments in criminal sentencing," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] K. Basu, R. Sinha, A. Ong, and T. Basu, "Artificial intelligence: How is it changing medical sciences and its future?" *Indian journal of dermatology*, vol. 65, no. 5, p. 365, 2020.

[4] B. Rastegarpanah, K. P. Gummadi, and M. Crovella, "Auditing black-box prediction models for data minimization compliance," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 20 621–20 632. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/ac6b3cce8c74b2e23688c3e45532e2a7-Abstract.html

[5] A. Goldsteen, "The future of data and ai in the financial services industry," IBM, Jan 2022, accessed: 2023-May-05. [Online]. Available: https://developer.ibm.com/blogs/data-minimization-for-machine-learning/

[6] Wolford, Ben, "What is gdpr, the eu's new data protection law?" https://gdpr.eu/what-is-gdpr/, n.d., accessed: March 29, 2023.

[7] The UK Information Commissioner's Office (ICO), "Guidance on the ai auditing framework: Draft guidance for consultation," 2020, accessed: 2023-Apr-12. [Online]. Available: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

[8] D. Shanmugam, F. Diaz, S. Shabanian, M. Finck, and A. Biega, "Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 839–849.

[9] A. Rai, "Explainable ai: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137–141, 2020.

[10] V. Turri, "What is explainable ai?" Carnegie Mellon University, Software Engineering Institute's Insights (blog), Jan 2022, accessed: 2023-Apr-11. [Online]. Available: http://insights.sei.cmu.edu/blog/what-is-explainable-ai/

[11] D. Sarkar, "The importance of human interpretable machine learning," Dec 2018. [Online]. Available: https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

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

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[15] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[16] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2021, Conference Proceedings, pp. 652–663.

[7] https://scikit-learn.org/stable/

[17] D. S. Watson, L. Gultchin, A. Taly, and L. Floridi, "Local explanations via necessity and sufficiency: Unifying theory and practice," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 1382–1392.

[18] European Parliament and Council of the European Union, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, 2016, accessed on March 30, 2023.

[19] A. J. Biega, P. Potash, H. Daumé, F. Diaz, and M. Finck, "Operationalizing the legal principle of data minimization for personalization," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 399–408.

[20] Norwegian Data Protection Authority, "Artificial intelligence and privacy," Norwegian Data Protection Authority, Oslo, Norway, Report 2, 2018. [Online]. Available: https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf

[21] A. J. Biega and M. Finck, "Reviving purpose limitation and data minimisation in data-driven systems," *arXiv preprint arXiv:2101.06203*, 2021.

[22] E. M. Alkabawi, A. R. Hilal, and O. A. Basir, "Feature abstraction for early detection of multi-type of dementia with sparse auto-encoder," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 3471–3476.

[23] H. Heuer and A. Breiter, "Student success prediction and the trade-off between big data and data minimization," *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*, 2018.

[24] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. R. O'Brien, T. Steinke, and S. Vadhan, "Differential privacy: A primer for a non-technical audience," *Vanderbilt Journal of Entertainment and Technology Law*, vol. 21, no. 1, p. 209, 2018. [Online]. Available: https://scholarship.law.vanderbilt.edu/jetlaw/vol21/iss1/4

[25] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.

[26] K. H. Tae and S. E. Whang, "Slice tuner: A selective data acquisition framework for accurate and fair machine learning models," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 1771–1783.

[27] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash, "Data minimization for gdpr compliance in machine learning models," *AI and Ethics*, pp. 1–15, 2021.

[28] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual examples," in *ACM Conference on Fairness, Accountability, and Transparency*, January 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/explaining-machine-learning-classifiers-through-diverse-counterfactual-examples/

[29] D. Dua and C. Graff, "Uci machine learning repository," 2017. [Online]. Available: https://archive.ics.uci.edu

[30] H. Zhu, "Predicting earning potential using the adult dataset," RPubs, Dec 2016, accessed: 2023-Feb-15. [Online]. Available: https://rpubs.com/H_Zhu/235617

[31] S. V. Wilson, "Multiple imputation with lightgbm in python," Medium, Sep 2020, accessed: 2023-Apr-20. [Online]. Available: https://towardsdatascience.com/multiple-imputation-with-random-forests-in-python-dec83c0ac55b

[32] A. Shanbhag and A. Taly, "Counterfactual explanations vs. attribution based explanations," Fiddler, Feb 2020, accessed: 2023-May-10. [Online]. Available: https://www.fiddler.ai/blog/counterfactual-explainable-vs-attribution-based-explanations

[33] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Lulu. com, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[34] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/13/4670