Strategies and Approaches on Explainable Artificial Intelligence

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Software Engineeringrigor

2022

Ilia Ershov

Examiner(s): Kari Smolander, Professor

Hasan Mahmud, Junior Researcher (Doctoral Student)

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Software Engineering

Ilia Ershov

Strategies and Approaches on Explainable Artificial Intelligence

Master's thesis

2022

54 pages, 10 figures, and 3 tables

Examiner(s): Professor Kari Smolander and Hasan Mahmud

Keywords: XAI, Interpretability, Explainability, LIME, SHAP.

This research thesis focuses on the problem of AI explainability. In recent years, the field of Explicable AI (XAI) has developed an extensive collection of algorithms that provide a useful set of tools for developers to create XAI applications. Explainability is regarded to have exceeded the demand for academics or scientists to comprehend the models they develop and has become a crucial requirement for consumers to embrace and trust AI in a variety of fields. This study describes existing XAI methods, pointing out their features and specific applications. In addition, it shows practical examples of the application of the described methods with real-world examples, pointing out the benefits of the results obtained. As a valuable outcome, a summary of the described methods with recommendations for their use is presented, as well as instructions on the basic steps necessary to integrate XAI into an existing business or a project based on the AI model.

ACKNOWLEDGEMENTS

The research presented in this thesis was conducted at the LUT School of Software Engineering at Lappeenranta–Lahti University of Technology LUT, Finland, between 2020 and 2022.

I am deeply grateful to my supervisor Professor Kari Smolander for his guidance and support in the challenges I faced during my work on this research.

LUT University has offered a fantastic chance to make new acquaintances from all over the world because it is an international community. I am grateful for this chance and would want to express my gratitude to all of my friends who have helped me during my study and life in Lappeenranta.

Eventually, I would like to express my heartfelt appreciation to my parents, Irina and Aleksei, for their inestimable support and assistance, which has allowed me to overcome all challenges in my path.

Abbreviations

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
ML	Machine Learning
LIME	Local Interpretable Model-Agnostic Explanations
SOM	Self-organizing maps
SHAP	SHapley Additive exPlanations
t-SNE	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoders
RMSE	Root Mean Square Error
PDP	Partial Dependence Plot
MAE	Mean Absolute Error
GLMs	Generalized Linear Models
GAMs	Generalized Additive Models
ICE	Individual Conditional Expectation
PCA	Principal Component Analysis
ALE	Accumulation Local Effects

Table of contents

1. Introduction

This thesis aimed to research how Artificial Intelligence could be more interpretable and easier to understand for the key stakeholder. AI has changed dramatically over the last decades. Starting from the academic research it became an integral part of human lives spreading in plenty of fields from science, retail, and healthcare to entertainment and advertising. Instead of laboratory experiments Data Science and Machine Learning has become a powerful tool to solve real-world problems to make daily routine better and easier. Usually, the basic strategies of machine learning, statistical or deep learning models do not change. Therefore, the paramount importance of data science is to solve complex problems by applying models on the right data. The key priority of modern AI findings is to apply the technology, but not to research the theoretical aspects. Usually, complex problems require complex models that are still often perceived as a black box. It means that finding connections between inputs and outputs and defining their relations is extremely difficult. This problem stands for the opportunity for the particular AI model to be interpreted. When ML technologies are applied to business needs or integrated into our everyday life, it frequently requires an explanation of why the decision was made some particular way and not another. The black box does not show how input parameters have influenced the final output. Nevertheless, the harsh reality is that real business projects rarely succeed without the possibility of data science pipelines or models to describe themselves and their solutions. Any real-world machine learning project typically has two aspects: business and technical aspects. In most cases, data scientists work on developing models and providing business solutions. However, there is a need of making AI interpretable and explainable also. The business doesn't have to be familiar with the complex details of how the model is actually working, but since models make a lot of decisions for them, they likely should have the ability to get answers to the question, "How can I trust your model?" or "How does your model really make its decisions?" [6]. This research represents the possible solutions on how to answer these questions.

For well over a decade, researchers have been trying to determine how AI models can provide descriptions to their decisions. In 2016 Ribeiro, Singh, and Guestrin

[1] made the first key steps toward this path and presented the concept of Local Interpretable Model-Agnostic Explanations (LIME). One of the key ideas of their study is worth mentioning.

> "Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: if the users do not trust a model or a prediction, they will not use it." [1]

Another vital work was done by Molnar in 2018 called «Interpretable Machine Learning, A Guide for Making Black Box Models Explainable» where he described the certain criteria that can be used to classify methods of interpreting models [2]. Finally, Arrieta and Alejandro described the concept of Responsible Artificial Intelligence in 2020. They tried to represent the large-scale implementation methodology for introducing AI methods in real-world businesses and projects taking into account its explainability, fairness, and accountability [3].

The research goal of this thesis is to make literature research on what is the essence of XAI and how AI models could be interpreted for stakeholders' needs.

This thesis explores two research questions through the process of overviewing existing literature and eliciting the key aspects that could be valuable for real businesses.

1. What is the essence of Explainable AI and why AI should be interpretive?
2. What are the strategies for organizations for integrating AI into their processes?

The main focus of this literature research is to clearly outline the need for AI for the business and provide possible approaches on how to effectively use it.

This thesis is organized into six chapters. Chapter two consists of reviewing related works in the areas of explainable artificial intelligence, key definitions will be described. In Chapter three, traditional methods for interpretability will be described as well as XAI methods. Chapter four will provide several real-world examples of XAI methods usage in the transport and logistic field with the results of predictions. Chapter V will include the limitations of the research, and some recommendations for the business

users and model developers. Finally, chapter six will include a summary of the investigation.

2. The essence of Explainable AI

Explainable Artificial Intelligence (XAI) is a concept of AI models that states explainability as an essential feature for AI users to trust, understand and effectively manage robust AI applications. Explainable AI is based on explainability — a clear description of how the various variables and weights of machine learning models generate their solutions. The ability of models to explain their outputs is determined by their interpretability. It could be either a mathematical or statistical understanding of the numerical outputs of decisions made by predictive models. The following chapter expands more on what is model interpretability, why is it important, gives criteria for the model's interpretation, and defines its scope and boundaries [5].

2.1. Importance of Model Interpretation

Machine Learning has become as popular as it is just in the last several years, so the model interpretation as a concept still is quite theoretical. The essence of any ML model is a response function that maps and explains patterns and relationships between the independent variables called input and the dependent variables called the response or target [6]. The model tries to predict or find somehow insights by taking certain decisions and choices. The key aspect of understanding and explaining the predictions that were made by the responsible function. To make it possible to answer the what, why, how e.t.c. questions it is needed to understand the process of making the decisions. Making the model transparent for asking these questions and getting concrete and precise answers would allow a model to be interpretable and make it human-understandable.

There are three key aspects of how to make a model interpretable were stated by Dipanjan [6]:

1) What impacts the model's predictions? We should have the ability to make queries to the model and get complex not obvious features' relations to understand what influences the predictions of the model the most. This establishes the model's *fairness*.

2) Why model made a certain decision? We should be able to know why the particular inputs impacted certain decisions. This establishes the model's *accountability* and reliability.

3) Why should the model's predictions be trusted? We should have the ability to evaluate and validate the model's decisions, thus decision-making process has to be trackable at its every step. This allows the key stakeholders to easily understand the logic of why a particular decision was made and check if the model works as expected. This establishes the model's *transparency*.

Overall, the interpretability of the model is a crucial feature that AI models should strive for. Even though the model has a high performance it also should be able to give answers to stakeholders' key questions. This provides transparency, accountability, and fairness and makes the model understandable for humans.

2.2. Concept of Interpretability and Background

Researchers commonly use the terms "interpretability" and «explainability" interchangeably. Indeed, the terms are very close in nature, but there are studies that highlight the differences between them. In fact, there is no precise mathematical definition of the terms explainability or interpretability, nor is there a metric to measure them. Despite the difficulties in defining these concepts, attempts have been made in [7-9] to clarify them. Despite this, the main shortcoming of the existing works is the lack of a mathematical basis for describing and measuring explainability. The definition of interpretability by Doshi-Weles and Kim is considered one of the most popular. They define explainability as follows:

> "The ability to explain or present something to a person in incomprehensible terms." [9]

An alternative viewpoint belongs to Miller and is formulated as follows:

> "The degree to which a human being can understand the cause of a decision." [10]

Despite the clarity of these definitions, they are not supported by a mathematical basis and rigor.

As mentioned above, it is mainly intuition that underlies interpretability in the analysis of model results. Determining the causal relationships between its inputs and outputs is easier when the machine learning system is the most interpretable. Thus, in image recognition tasks, some key patterns (input) may influence the system's decisions in identifying some object as part of the image (output). On the other hand, explainability is based on the logic of the machine learning system. A deep understanding of the mechanics of a model's internal processes, it's learning, and decision-making logic makes the system more explainable. When a model is said to be interpretable, it does not necessarily mean that users must understand the logic of the system and its internal processes. Consequently, an interpretable machine learning model does not necessarily mean an explainable model and vice versa. The claim that in addition to the concept of interpretability, it is also necessary to use the concept of explainability was supported by Gilpin et al [8]. For the aforementioned reasons, in this thesis interpretability is treated as a broader concept than explainability.

Doshi-Velez and Kim [9] proposed how evaluation methods for interpretability can be classified: *application-grounded, human-grounded,* and *functionally-grounded.*

- *Application-grounded* assessment is concerned with how the results of the interpretation process affect the individual, the subject matter expert, and the end-user in terms of a specific and well-defined application or task. Specific examples of this type of evaluation include whether the interpretation method results in better error detection or less discrimination.

- *Human-grounded* evaluation is similar to application-level evaluation, but there are two main differences: firstly, in this case, the tester need not be an expert in the field, they can be any end-user, and secondly, the ultimate goal is not to evaluate the created interpretation for its suitability for a particular application, but to test the quality of the created interpretation in a more general environment and to measure how well the general concepts are conveyed. An example of measuring how well an interpretation

captures an abstract concept of input would be when a person is presented with different interpretations of input and then chooses the one that they think best captures their essence.

- *Functionally-grounded* assessment uses specific and well-defined mathematical descriptions of interpretability to evaluate the quality of the method and does not require human experiments. This criterion is usually used after the class of models has already been tested by humans or by application-based experiments. Some mathematical evaluation can then be used to determine the quality of the interpreted models and to compare them. Feature-based evaluation is also applied when, for ethical or other reasons, experiments cannot be carried out with human participants or when the method is not sufficiently trained to be tested by humans. However, finding and eliciting relevant measurement criteria and metrics for every case is a complex task and still is an unsolved problem.

2.3. Scope of Machine Learning Model Interpretation

In this chapter, some of the aspects of scope and boundaries definition of the interpretability are considered. Several criteria can be elicited and used for classifying interpretation methods of AI models. Molnar provided a really useful guide to this question. Classification of Interpretability methods of Machine learning models is presented in Figure 1.

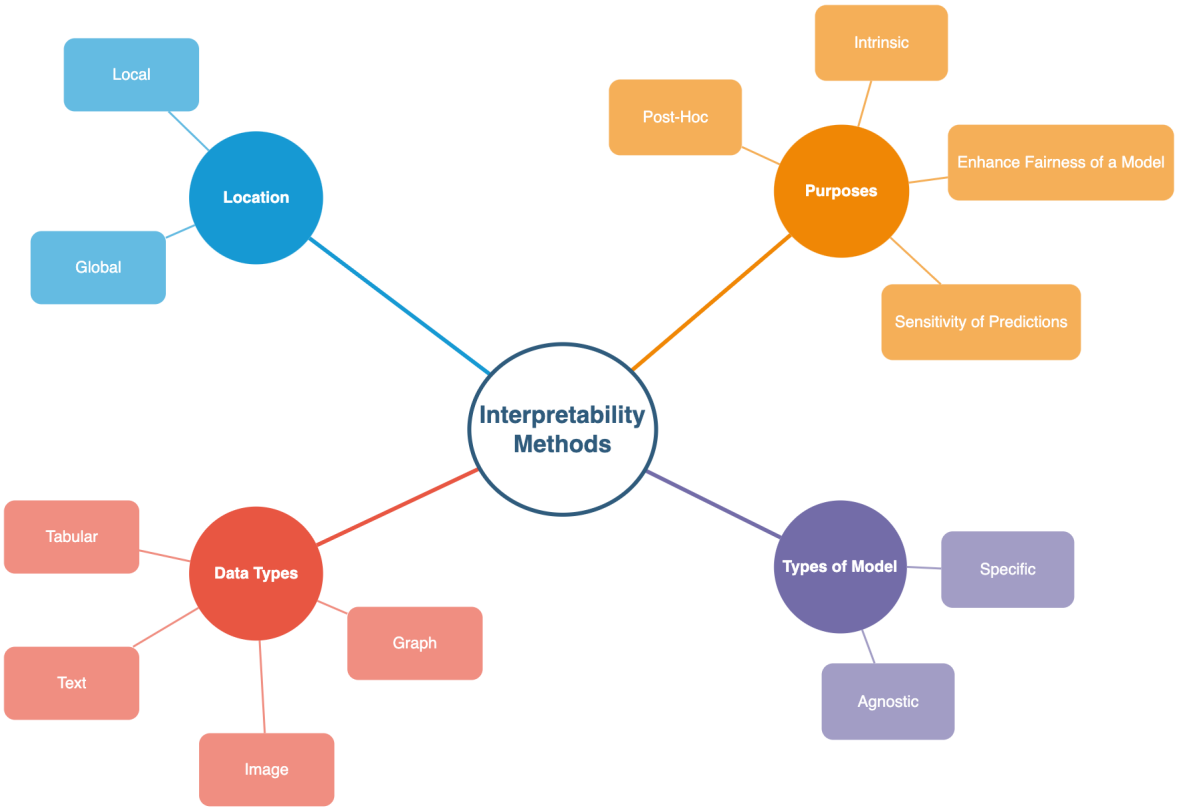


Figure 1. Classification of Interpretability methods [11].

2.3.1. Purpose of Interpretability

- *Intrinsic* interpretability is implied when considering a machine learning model that is inherently internally interpretable (linear/tree-based/parametric models).
- *Post-Hoc* interpretability is all about the training of a black box model as neural networks or ensemble methods and implying the same interpretability technics after the training.
- The *Sensitivity* of the model's predictions is an important part of the explainable AI field. Sensitivity determines how stable is the model depending on different inputs when a particular model is used in real-world processes.
- The *Enhance fairness* issues are part of the ethical concerns in XAI. The most prominent example is in medicine, the court system, and similar systems where a person has to remember a large set of rules (medical histories) and to be able to use them at any time in their work (conforming to personal data consent).

2.3.2. Type of a Model

- *Model-specific* interpretation tools coincide with intrinsic model interpretation methods well. These models highly depend on features and prospects on a per-model basis. These model-specific tools can be p-values, coefficients, rules from a decision tree, AIC scores about a regression model, e.t.c.
- *Model-agnostic* tools can be used for any ML model and they are generally specific for post hoc methods. Model-agnostic methods are defined as methods that do not have an access to any internals of the model as assumptions, constraints, or weights.

2.3.3. Location of a Model

- *The global interpretation* could allow us to realize how the AI model makes its predictions and how its subsets impact the results. To make the model's interpretation globally, global interpretability is needed. It helps to explain the complete model's results by analyzing relations between the independent features (predictors) and dependent variables (response) on the whole dataset. One of the possible approaches to comprehending feature interactions and importances is visualizing features. However, when the model becomes more complex and has more than two or three dimensions, it becomes tricky to visualize features' relations to analyze them. Therefore, frequently it is more effective to split the complete model into several separate parts and subsets of features, that might have an impact on a modal globally. Thus, Global interpretation requires an understanding of its stricture, constraint, and assumptions.
- *The local interpretation* allows us to understand how the models make decisions for one particular instance or a group of instances. It means that focus should be concentrated on local data points either than on the inherent structure or assumptions of a model. Understanding the model's decisions made at a single data point could be achieved by analyzing the local subarea in a feature space around that point. It basically uses the same approach as global interpretations but in the scope of a local feature. Nevertheless, feature spaces and local data distributions could perform completely differently

compared to global interpretations. It can be used to get more precise and more accurate explanations. For model-agnostic local interpretations, The Local Interpretable Model-Agnostic Explanation (LIME) framework can be used [1]. In addition, global and local interpretations can be used together to describe the model behavior for a group of instances.

2.3.4. Data used by a model

The difference in interpretation methods is easy to understand considering the types of data processed by machine learning models. It is possible to train a model to handle the most frequently used data types, such as images or tables, but it's required to take into account that models might handle textual and graphical data. Moreover, sometimes data could be converted from one type into another from the list mentioned above [12].

2.3.5.Conclusion

Interpreting the model in terms of transparency means understanding how the particular algorithms and features of decision-making work. Basically, the machine learning model is features that are used by an algorithm that tries to make predictions by mapping inputs to potential outputs (responses). Understanding what might influence the model's decisions and how a particular model is built can help to describe the model's transparency. This criterion can be represented in linear model coefficients, weights of a CNN filter, or neural network. However, as businesses rarely can have any knowledge of these technical details, usually the most valuable methods in representing model transparency are agnostic local and global interpretations.

The criteria mentioned above are not comprehensive for classifying interpretable methods due to the field of AI interpretability is still emerging.

3. Model Interpretation Strategies

This part aims to describe existing, traditional model interpretation methods. In addition, it also covers the classic model's interpretability vs. accuracy trade-off and overviews the XAI methods, their limitations, and challenges.

3.1. Traditional Model Interpretation Techniques

Bringing fairness, accountability, and transparency to the models would allow humans to be confident enough to apply these models to real business problems that have a dramatic impact on society. Therefore, there are plenty of existing techniques, that are considered below. They could be classified as visualization techniques and exploratory analysis (dimensionality reduction and clustering) and metrics for performance evaluation of the model (accuracy, recall, precision, mean absolute error for regression models, root mean-square error).

3.1.1. Traditional Methods Review

Exploratory analysis has been around for quite some time and for many years one of the most accessible and useful tools has been specifically data visualization to extract hidden information from data. Some of these methods reveal meaningful details and key features of the data. It is this information that helps to understand what influences the model in the decision-making process and to present this in a human-readable way.

In addition, if the curse of dimensionality is a problem, dimensionality reduction techniques can also help to reduce the feature space (the curse of dimensionality) and understand which features affect the model's decision-making process. Examples of these methods can be seen below

- Dimensionality reduction: Principal Component Analysis (PCA) [13], Self-organizing maps (SOM) [14], Latent Semantic Indexing [15].

- Variational autoencoders: An automated generative approach using variational autoencoders (VAE) [16].
- Manifold Learning: t-Distributed Stochastic Neighbor Embedding (t-SNE) [17].
- Clustering: Hierarchical Clustering [18]

An important research step in any data science to select a model is to evaluate its performance. Performance evaluation allows us to compare different performance measures of models in order to find the best one. In addition, it can allow individual model metrics to be improved, optimized, and tuned. The choice of metric for model evaluation is usually influenced by the type of problem to be solved. Some of the approaches are listed below [6]:

- Supervised learning - classification: In this type of task, it is necessary to predict a response variable. A number of useful metrics such as accuracy, precision, recall, and F1 score can be identified from the mixture matrix as shown in the following example. In addition, methods such as the AUC score [19] and the ROC-curve [20] can be used to solve classification problems.
- Unsupervised Learning — Clustering: For unsupervised learning problems based on clustering we can use metrics like the silhouette coefficient [21], homogeneity, completeness, V-measure, and the Calinski-Harabaz index.
- Supervised Learning — Regression: For regression problems, we can use standard metrics like the coefficient of determination (R-square), the root mean-square error (RMSE), and the mean absolute error (MAE).

3.1.2. Challenges and Limitations of Traditional Techniques

The aforementioned methods certainly allow us to evaluate AI models and reveal additional information about the data used, and their characteristics in order to improve the model. However, these techniques are not effective in explaining models for human interpretation. As a rule, any data science problem requires a model with an accountable

data set and an objective function (an optimized loss function). This function must satisfy some criteria, namely to satisfy business requirements and to be based on the performance of the model. Typically, performance metrics and exploratory analysis techniques are used to estimate the overall performance of the model using specific data. In reality, however, model performance often degrades and stabilizes after deployment due to variability in data characteristics, additional noise, and constraints [22]. This can combine things like changes in features, changes in the environment, and additional constraints. Therefore, it is not sufficient to re-train the model with the same features but to continually check how essential the features are in determining the results of the model and how well they perform with new data points.

In addition, models can often be biased due to the nature of the data we use, as in the problem of predicting rare classes (intrusion detection or fraud). The true history of a model's predicted decisions cannot be justified by metrics. The traditional model interpretation techniques described above can be easily understood by e.g. the data analyst as they are inherently theoretical and mathematics-based. At the same time, it is very difficult to explain the results of interpreting traditional methods to business stakeholders who do not have a technical background. Consequently, these methods are not suitable for project success assessment.

It is not sufficient to simply provide the business side with information about the high-fidelity predictions of the model in order for them to trust the model when it is implemented in real projects. Models must be able to provide intuitive inputs and outputs for human interpretation of model acceptance policies. This would make it easy to share in-depth information with colleagues (data engineers, data scientists, managers, and analysts). The use of that kind of form of explanation that can be analyzed on the basis of inputs and outputs can improve communication and cooperation, allowing businesses to make decisions more confidently when evaluating risks in financial institutions).

Overall, model interpretation definition is the ability to take into account the fairness (unbiased/non-discriminatory), transparency (ability to request and confirm predictive decisions), and accountability (reliable results) of the predictive model.

3.1.3.The Accuracy vs. Interpretability Trade-off

As we have a typical Bias vs. Variance Trade-off in machine learning, we also have a standard Trade-off between Model Performance and Interpretability. Typically, easy-to-understand models such as trees and linear models (linear-logistic regression) are preferred among industry stakeholders. They are easily interpretable and intuitive even for the non-technical person in the field of data science. Due to these models are easier to understand, the door to them is much higher for stakeholders. However, the harsh reality is that in order to solve real problems, it is often necessary to apply much more complex models, which are inherent in the problem of high data dimensionality. These models (neural networks and ensembles) are inherently non-linear, which makes their interpretation impossible. traditional methods. for these reasons, developers strive to improve the performance of models and preserve their ability to interpret. Finding a balance between these two requirements can take significant time for data scientists.

3.2. XAI Methods Review

Typically, most machine learning models are black boxes. However, it is very important to understand the decision-making process of the model. To assess the reliability of forecasts, it may be necessary to be able to know what parameters affect them. One of the goals of XAI is to develop innovative explanatory algorithms that promise to provide new insights into current black-box models of machine learning and thereby help the user better understand and trust the artificial intelligence system [24].

The following factors contributing to the proliferation of algorithmic decision-making based on artificial intelligence can be mentioned:

1) The availability of powerful computing resources (e.g. GPU computing, cloud computing)
2) The need to process various voluminous data
3) New and powerful algorithms

Legislative regulation is also introduced to control the implementation of the direction:

1) "Right to explanation" - European Union General Data Protection Regulation (GDPR) [25]

2) U.S. Department of Defense Ethical Guidelines for Artificial Intelligence [26] and addressing fairness, accountability, and transparency through automated decision-making systems.

3) U.S. Government Algorithmic Accountability Act of 2019 [27]

The most popular tools for explainable AI are the following: DeepVis Toolbox, LIME, Keras-vis, TreeInterpreter, MindsDB, SHAP, Microsoft InterpretML, Tensorflow Lucid, Tensorboard WhatIf, Cleverhans Tensorflow, etc.

At a general level, each of the proposed methods has similar strategies such as surrogate models, feature importance, feature interactions, explicit values, partial dependency, knowledge injection, etc. The renewed interest in XAI research after expert systems is due to recent advances in AI, concerns about unethical use, its integration into a wide range of fields, unwanted distortions in models, and lack of transparency. Moreover, various governments produce new laws that require more XAI research.

In 2019, Muller and colleagues provided a detailed overview of the approaches used by several types of "explanation systems" and classify them into three generations [29]:

1) First Generation - Expert Systems
The expert systems of the early 1970s which try to explicitly express the internal system’s workflow by embedding specialists’ knowledge in rules. Usually, it was collected directly from experts.

2) Second Generation - From Expert Systems to Knowledge-Based Tutors
They are considered as a human-computer system, built based on human experience and knowledge and the ability to provide cognitive support. For example, organize the interface so that it complements the knowledge that the user lacks.

3) Third-generation

Tools and methods that have recently emerged, starting in 2015. Third-generation systems are similar to first-generation systems in terms that they also attempt to clarify the inner systems' workflow. However, they, in contrast, are usually black boxes (e.g., ensemble approaches, deep networks).

Broadly speaking, methods of explanation are classified into three types: *internal interpretive methods*, *model-agnostic explanations*, and *example-based explanations*.

3.2.1. Interpretable Methods

Models that can be interpreted «out of the box» are interpretable by definition and they are a good and easiest start. These usually include conventional parametric models such as rule-based fits, linear regression, tree-based models, logistic regression, and even k-nearest neighbors and Naive Bayes type models.

In linear regression, the predicted target consists of a weighted sum of the input features. Thus, the weight or coefficient of a linear equation can be used as a means of explaining a prediction when there is a lack of features.

Logistic regression is an extension of linear regression that considers the problem of classification, which models probabilities for classification problems. The main difference between linear regression and logit regression is that it provides a probability o¤er between 0 and 1, where this value can represent with some approximation the linear dependence on the predicted probability. However, the weight provides information on how strong the effect is between classes and its direction (negative or positive).

Decision tree models split the data several times using a cut-off threshold at each node until a final node is reached. The main advantage of decision tree models over models

is that it works even when there is no linear relationship between inputs and outputs, and when objects interact with each other (correlation between objects). However, this method does not reveal a linear relationship between inputs and outputs. Moreover, small changes in the input data sometimes influence a lot on the predicted outcome. In addition, there

may be several different trees for the same problem. As a rule, the more nodes or the deeper the tree, the more difficult it is to interpret.

Another way is decision rules (simple IF-THEN-ELSE conditions). (IF-THEN) rules often cannot solve the classification problem and give incorrect results when interpreting linear relationships. For example, the RuleFit algorithm [28] considers sparse linear models that can detect interaction effects in the form of decision rules and as a consequence allows for internal interpretation. At the same time, it learns some new features in addition to the original ones, which usually worsens interpretability as the number of features increases.

There are also other extensions of interpreted models, such as generalized additive models (GAM) and generalized linear models (GLM). They allow some assumptions of linear models to be taken into account. Consider the case where there is no interaction between functions due to the presence of a Gaussian distribution in the target outcome y and a given function. Besides the benefits, these extensions complicate the models, as they increase the number of interactions and consequently make them more difficult to interpret. Moreover, there is the Naive Bayesian classifiers and K-Nearest Neighbors classifier.

These models can be classified according to their main capabilities as follows:

- Linearity: Used if there is a linear relationship between the target and the attributes.
- Monotonicity: It establishes that the relationship between the target outcome and the attribute always has one consistent direction (increasing or decreasing) depending on the attribute (over its entire value range).
- Interaction: Non-linearity features and interaction to the model can always be added through manual feature development.

To summarize the aspects mentioned above Table 1 by Molnar's excellent book, 'Interpretable Machine Learning' [2] is presented.

Table 1. Interpretation methods classification [2].

Algorithm	**Linear**	**Monotone**	**Interaction**	**Task**
Linear models	Yes	Yes	No	Regr.
Logistic regression	No	Yes	No	Class.
Decision trees	No	Some	Yes	Class.
RuleFit	Yes	No	Yes	Class. + Regr.
Naive Bayes	Yes	Yes	No	Class.n
K-nearest neighbours	No	No	No	Class. + Regr.

It should be taken into account that better ways of constructing and interpreting more complex and more efficient models have to be found due to mentioned above models are quite simple.

3.2.2.Model-agnostic explanations

Model-independent methods separate explanation from model machine learning so that the explanation approach can be used with a variety of models. There are several clear benefits to this separation, including [1]:

- The interpretation method is able to work with several machine learning models learning
- Different results of interpretability can be provided (e.g., visualizing the importance of features) for a particular model
- There is room for flexibility in representation, e.g., a text classifier utilizes abstract word embedding to classify but real words to explain.

Some of the model-agnostic explanation tools are Feature Importance, Feature Interaction, Individual Conditional Expectation (ICE), the Partial Dependence Plot (PDP), Accumulation Local Effects (ALE) Plot, Local Surrogate, Global Surrogate, and Shapley Values.

The Partial Dependence Plot (PDP) [30] or The marginal influence of one or two characteristics (at most three features in 3-D) on the anticipated ML model result is depicted by the PD plot [28]. It is a global technique since it displays the model's general behavior and can display linear or complicated relationships between the target and the object(s). It offers a function that is solely dependent on the object(s) being mapped by choosing other objects and including their interactions. By displaying changes in the prediction as a result of changes in individual functions, PDP gives a clear causal interpretation. PDP, on the other hand, implies that items are unrelated to one another. Furthermore, there is a practical limit to the number of aspects that PD can clearly convey at the same time. It's also a global technique since it shows the average influence of (all instances of) the object/objects on the prediction rather than the effect of all objects in a single instance.

Individual Conditional Expectation is a term used to describe how users believed things to turn out Unlike PDP, ICE creates a single line per instance to indicate how the item influences forecast changes [30].

Accumulation Local Effects. ALE plots, like PD plots, show how factors impact the forecast on average. ALE, on the other hand, works fairly well with correlated features and is faster than PDP. Even though the ALE plot is not biased toward correlated data, when the features are strongly correlated and studied separately, it is difficult to discern changes in the forecast. Only graphs that display changes in both linked features at the same time make sense in this scenario for comprehending changes in the prediction.

Feature Interaction [31] is another way. Individual feature effects do not add up to the overall feature effects from all features combined when features interact. H-statistics (also known as Friedman's criteria) aid in the detection of many forms of interactions, even when three or more characteristics are present. The difference between the partial dependency functions for the two functions together and the sum of the partial dependency functions for each function individually reflects the strength of the interaction between the two functions.

The following approach is called **Feature Importance** [32]. When we shuffle the feature values to break the association between the feature and the result, the relevance of a feature usually increases the model's prediction inaccuracy. If the feature value errors rise after shuffling, the feature is essential. Later, for random forests, a significance function based on permutations was introduced. The work was eventually expanded to include a model-independent variant. The relevance of features gives a simple and comprehensive insight into the ML model's behavior. While the significance of features considers both their main effect and their relationships, Because it is included in the significance of associated traits, this is a drawback. As a result, when features interact, the relevance of each feature has no bearing on the total performance reduction. Furthermore, it is uncertain whether to evaluate the value of objects using a test set or a training set, as the latter reveals the variations between runs in a shuffled dataset. It's important to note that feature significance is also covered by global techniques.

Global Surrogate. Using the interpreted ML model, the global surrogate model aims to simulate the general behavior of the black-box model. Surrogate models, in other words, try to replicate the black-box model's prediction function as closely as possible using the interpreted model, providing that the prediction is interpretable. A metamodel, a response surface model, an approximation model, or an emulator are all terms used to describe it. The disadvantage is that there are several plausible explanations for the same black box, such as various decision trees with different architectures.

Shapley Values is another method of local explanation. Shapley introduced this value in 1953 [33]. It is based on the theory of cooperative games, which helps to fairly distribute the importance of the functions among the players involved. Each instance function value is supposed to be a player in the game, and the prediction is the total payout, which is dispersed among the players (i.e., functions) based on their contributions to the total payment (i.e., prediction). Shapely Value is the average prediction contribution across all conceivable item coalitions, making it computationally costly when there are a large number of objects, e.g., there will be 2k coalitions for k objects. Shapely Value, unlike LIME, is a theory-based explanation technique that provides entire explanations. It does, however, have difficulty with connected features. Furthermore, Shapely Value returns a

single value for each object, so there's nothing to explain if changes in the input produced changes in the output. The SHAP library is one way to implement Shapley Values.

3.2.3. Example-based explanations

Example-based explanation approaches agnostically describe model behavior and data distribution by using individual cases from the dataset. "X is identical to Y, and Y caused Z, thus the prediction indicates that X will cause Z," explains the author. Next, consider [2], several methods of explanation that fall into this category.

The Adversarial method identifies the input alterations that will result in a substantial change in the prediction/output. Individual predictions can be explained using counterfactual explanations. It can, for example, give an explanation detailing a situation's direct association, such as "If A didn't happen, B wouldn't have happened." Although counterfactual explanations are useful for humans, they suffer from the "Rashomon effect," in which each counterfactual explanation tells a different tale in order to get at the prediction. In other words, for each instance-level prediction, there are numerous accurate explanations (counterfactuals), and the challenge is deciding which one is the best. Counterfactual approaches don't need data or models to work, and they can even work with a system that doesn't employ machine learning. Furthermore, this strategy is ineffective for categorical variables with a large number of values.

By employing counterfactual instances to deceive machine learning, the adversarial method has the advantage of reversing the result (i.e., small deliberate perturbations in the input data to make false predictions). Adversarial instances, on the other hand, can assist expose hidden flaws and enhance the model.

Prototypes are composed of a small number of instances that accurately reflect the data. On the other hand, a group of examples that fails to accurately reflect the data receives criticism. Influential Instances are data items from the training set that have an impact on model parameter prediction and determination. Determining the correct cutoff point to

distinguish influential and non-influential instances is a tough issue, even though it aids in debugging and better understanding model behavior.

K-nearest Neighbors Model, the k-nearest neighbor model forecast can be described by k-nearest neighbor data points, according to the k-nearest neighbor model (neighbors that have been averaged for prediction). An understanding of why an instance is a member of a certain group or cluster may be obtained by seeing an individual cluster comprising comparable examples.

3.3.LIME

LIME belongs to the category of locally explained algorithms. In most cases, such a strategy focuses on a single data instance in order to develop explanations utilizing various data. We want to generate g to clarify the decision that was made f for a single input instance x in this case (Figure 2).

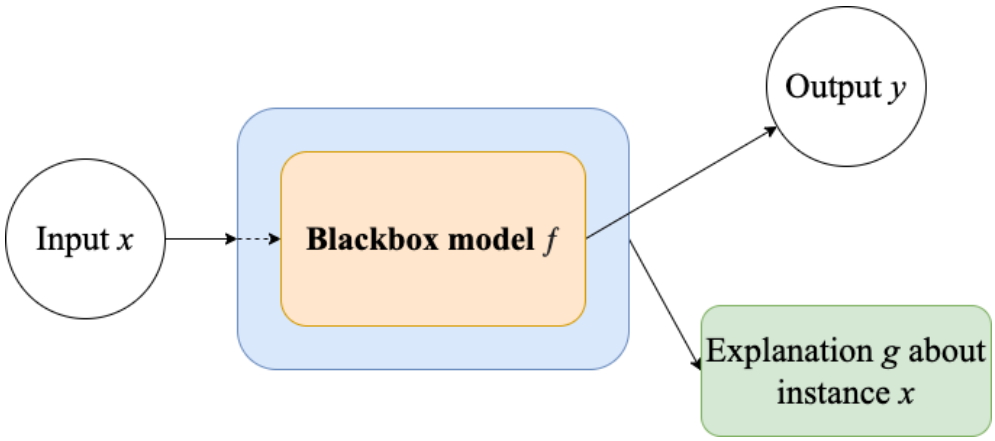


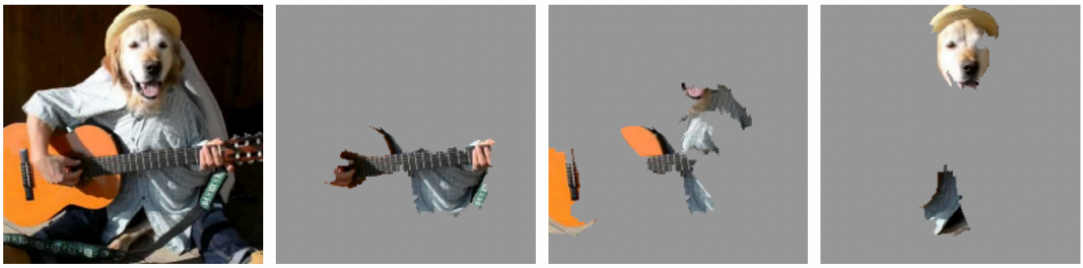
Figure 2. A diagram showing locally explainable models in general. In most cases, a single input instance is utilised to explain anything.

Bayesian approaches are used in foundational research in local explanations, and matrix functions are important for interpreting model output predictions. Positive real matrices or vectors have always been the output reasons. New research in local explanatory models improves on older techniques based on graphs and game theory, which provide a functional estimate of positive and negative correlations in output categorization. A

positive number leads a function to raise the output class's probability, whereas a negative value indicates that the function has decreased the output class's possibility.

Local interpretive model-agnostic explanations (LIME) were introduced by Ribeiro et al. in 2016 [1]. LIME seeks to discover the relevance of neighboring superpixels (pixel fragments) in the source picture in the output class to create a human-understandable representation. As a result, LIME finds a binary vector $x_0 \in 0, 1$ to signify the existence or absence of a continuous route or "superpixel" that provides the output class with the best representation. For single data input, this operates at the patch level. As a result, the procedure is classified as a local explanation. SP-LIME is a global explanation model based on LIME that is explained in the model's global explanation component.

An example visualization of the single-sample LIME algorithm is shown in Figure 3.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 3. Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24) and "Labrador" (p = 0.21) [1].

The procedures for describing the model for a single input instance and the broader LIME approach (Figure 4) are shown in Algorithm 1. We rearrange the data for the input case by locating a superpixel. The distance (similarity score) between the permutations and the original observations is then calculated. For the original input and the new "fake" data, we now receive different class scores. The f model may then be used to generate predictions based on the new "fake" data. The amount of superpixels picked from the original data determines this. The most essential aspect is that we may pick what enhanced the

rearranged data forecast. We may utilize feature weights, or coefficients, from a basic model, such as a locally weighted regression with permuted data, m functions, and similarity estimates as weights, to describe the local behavior of a complicated model.

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

Figure 4. Lime Algorithm [1].

In recent years, a lot of research has gathered steam, with the goal of enhancing and expanding the LIME algorithm to solve a variety of new issues. The summary of some of them is following:

- Mishra and colleagues [34] introduced an expanded LIME method for music that included content analysis via time segmentation, frequency and frequency-time segmentation of the input chalk spectrogram, and frequency and frequency-time segmentation of the input chalk spectrogram. Sound-LIME (SLIME) was their technique, and it was used to explain voice predictions.

- To clarify Bayesian prediction models, Peltola [35] defined LIME as KL-LIME, which is based on the Kullback-Leibler distance. The explanations are constructed using an interpretable model whose parameters are determined by minimizing the KL divergence and the prognosis model, similar to LIME. By projecting information from the forecast distribution into a simplified interpretable probabilistic explanation model, local interpretable explanations may be constructed.

- To replace the LIME random perturbation, Rehman and colleagues [36] utilized an agglomerative hierarchical clustering (HC) and K-nearest neighbor (KNN) method. The authors employ HC to cluster the training data in a similar way to clusters, and KNN to locate the instance under the test's nearest neighbor. Instead of a random perturbation, like in LIME, after KNN picks a cluster, that cluster passes a perturbation as input. The researchers state that this method produces more stable explanation models than the classic LIME technique.
- Using a quadratic approximation structure termed QuadraticLIME, Bramhall, and colleagues [37] have altered the linear LIME relations to incorporate nonlinear relations (QLIME). They arrive at these conclusions by taking into account linear approximations in the complex function. The results show that when QLIME is utilized, the mean square loss (MSE) of the LIME linear connection at the local level improves.
- Modified Perturbed Sampling for LIME was introduced by Shea and colleagues [38] as a substitute approach for selecting superpixel information for picture data (MPS-LIME). By transforming the superpixels into an undirected graph, the authors changed the usual superpixel selection procedure into a response set issue. In the MPS-LIME approach, the response operation reduces the number of affected samples, which improves running time. The authors compared their technique to LIME, which uses the Mean Absolute Error (MAE) and the R2 determination factor, and found that their method produced better results in terms of clarity, correctness, and timeliness.

3.4.SHAP

Shepley's additive explanations (SHAP) is a game-theoretic way to utilize Shepley values to explain the results of any machine learning model. Lundberg and colleagues [39] proposed it. SHAP calculates the contribution of various features to the output prediction to explain input x predictions. One may compute Shapley values to understand how to evenly distribute payoffs by treating the data functions as participants in a coalition game. Individual categories in tabular data or groupings of superpixels in images like LIME can

be used in the SHAP approach. The problem is then produced as a series of linear functions, with the explanation being a linear function as well [40].

Lundberg describes a number of variations of the basic SHAP method [44], including KernelSHAP, which reduces the number of estimates required for large input data in any machine learning model, LowOrder SHAP, which is effective for small maxima with coalition size M, LinearSHAP, which estimates SHAP values using model weights given independent input characteristics, and DeepSHAP, which adapts the DeepLIFT method to use deep neural networks to improve attribution. We discuss KernelSHAP in Algorithm 2 since it applies to all machine learning techniques (Figure 5). KernelSHAP's fundamental idea is to use coalition sampling to execute an additive feature attribution approach by deleting features from the input data and linearizing the model's effect using KernelSHAP.

Algorithm 2 KernelSHAP Algorithm for Local Explanations

Input: classifier f, input sample x
Output: explainable coefficients from the linear model
1: $z_k \leftarrow \text{SampleByRemovingFeature}(x)$
2: $z_k \leftarrow h_x(z_k)$ ▷ h_x is a feature transformation to reshape to x
3: $y_k \leftarrow f(z_k)$
4: $W_x \leftarrow SHAP(f, z_k, y_k)$
5: `LinearModel(`W_x`).fit()`
6: **Return** `LinearModel.coefficients()`

Figure 5. KernelSHAP Algorithm for Local Explanations [43].

SHAP is also widely utilized in the scientific community, has been directly implemented, and has advanced in a variety of ways. Let's look at some examples of how SHAP has been used in the medical industry to explain clinical decision-making, as well as some current work that is critical:

- To explain the autoencoders used to detect anomalies, Antvarg and colleagues [41] developed an expanded SHAP technique. The authors use the autoencoder to categorize anomalies by comparing real data instances to the reconstructed output. The authors think that explanations must be based on the reconstruction mistake because the final product

is a reconstruction. The most efficient features are assigned SHAP values, which are separated into contributing and compensating anomalies.

- Sundararajan and colleagues [42] pointed out that the SHAP technique has a number of flaws, including the ability to generate nonsensical explanations in circumstances when certain functions aren't necessary. The basic Shepley (BShap) approach improves this attribute of the "uniqueness" of the attribution method. Using integrated gradients for continuous areas, the authors improve the approach even further.

- By expanding the KernelSHAP technique to accommodate dependent features, Aac and colleagues [43] examined the relationship between SHAP values. The authors also demonstrated a method for calculating Shepley cluster values for dependent features. The KernelSHAP technique was thoroughly investigated utilizing four proposed approaches to replace the conditional distributions of the KernelSHAP method using an experimental method and either Gaussian.

- In a framework called TreeExplainer, Lundberg and colleagues [44] developed an extension of the SHAP technique for trees, where the structure of the global model is explained via local explanations. The authors presented a polynomial-time approach for calculating the local tree explanation using accurate Shepley values.

- A SHAP-based strategy for explaining time-series signal predictions incorporating long term short term memory (LSTM) networks were developed by Garcia and colleagues [48]. Individual cases in a test set were explained using the DeepSHAP algorithm based on the most essential attributes from the training set. The SHAP method, on the other hand, was left unchanged, and explanations were created for each input instance.

4. Overview of real-world cases of using XAI methods

In this chapter, several researches related to applying XAI methods to such smart cities' problems as transport are reviewed. The results are considered from the view of the explainability and effectiveness of the suggested methods. Some findings are proposed as suggestions for future related works.

4.1. Travel Time Prediction and Explanation

In 2021 Irfan, et al. [46] researched travel time prediction in the logistics domain and freeway travel. Authors have researched the topic and reviewed relevant literature to choose the most appropriate travel time prediction techniques.

4.1.1. What was done?

Their study aims to improve travel time forecasting for delivery vehicles, which as a result should improve delivery time forecasting. The work was supported by a company interested in obtaining valuable results for its activities. The aim of the study is to investigate existing problems and to provide a useful basis for research on the basis of travel time forecasting.

Several tests were carried out on several data models using three distinct datasets in order to obtain the optimal prediction model. Two ensemble learning methods, LightGBM and XGBoost, as well as three neural network models with LSTM, biLSTM, and one LinearSVR model, and, GRU layers have been chosen.

In addition, a single hybrid model architecture was tested using four alternative combinations of ensemble learning and neural network models. Finally, a comparison of the tested models' performance is shown. They found that ensemble tree-based learning approaches, such as XGBoost and LightGBM, perform the best across a variety of datasets. This indicates that they are the most reliable approaches for predicting path time.

For the PeMS dataset, there is no substantial difference in performance between ensemble techniques and neural networks, while there is for datasets 1 and 2. (NextUp Software dataset). More crucially for us, they used explicable AI (XAI) methodologies to clarify machine learning models' travel time forecasts.

For the training model three different datasets were used [46]:

- Dataset-1: This information was received from the case study firm. It comprises the location and temporal data of orders delivered between 22 January 2019 and 14 April 2021, a period of around 27 months. The deliveries take place six days a week, on weekdays and Saturdays. The data is taken from the case study company's transport management system (TMS).
- Dataset-2: It differs from dataset-1 only in that it contains data on the orders delivered between 5 March 2019 and 15 December 2020, approximately 22 months.
- Dataset-3: Caltrans Performance Measurement System provided this dataset (PeMS). PeMS is a user service for preserved data that contains roughly 10 years of data for historical study. PeMS collects traffic data from 39,000 individual detectors in all of California's main urban regions.

Because it employed non-interpretable pre-trained models, the model's explainability was termed posthoc explainability in that study [3]. Ensemble learning methods might employ model simplification or feature relevance strategies to explain the model. Feature relevance, model simplification, and visualization strategies are three forms of explanations for neural networks. Model simplification and local explanation strategies are both possible for reference vector machines. Furthermore, they adopted the feature relevance explanation for the regression problem since it makes the most sense.

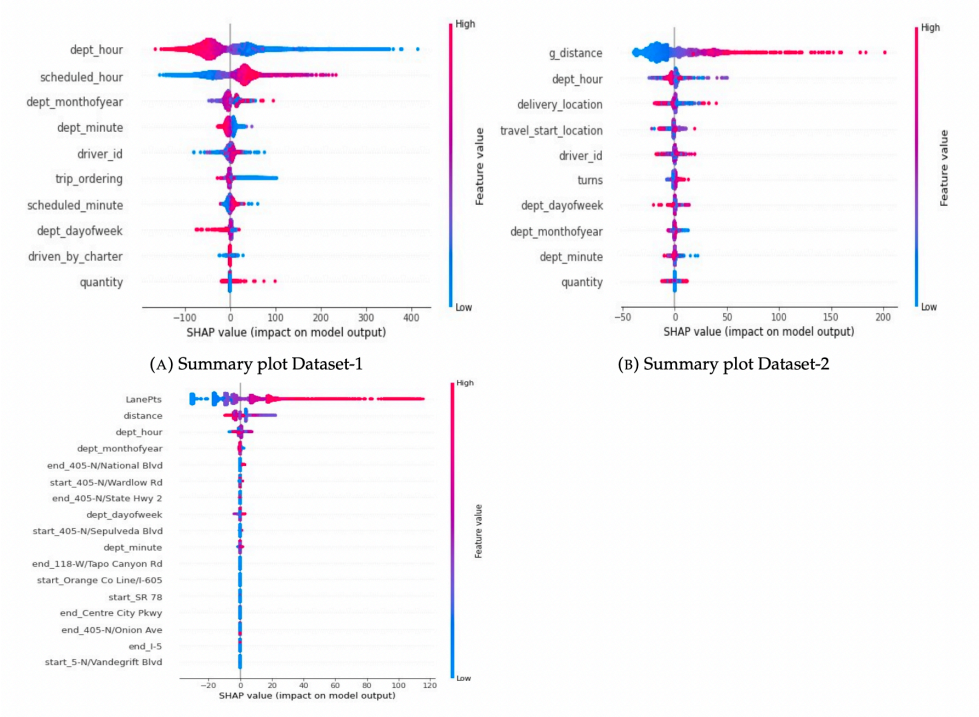
They employed the SHAP model explanation approach to determine the most essential elements for travel time prediction. It was done for each dataset considered in the research to understand what are the important features that influence the travel time forecast and if some particular features remain important across several datasets and ML models.

Once SHAP is implemented, there are a variety of approaches to describe the model, one of which is emphasizing the significance of the feature. The contribution of a feature can be determined by its significance, and the contribution might be positive or negative. This indicates that a feature can influence any prediction in either a positive or negative direction.

Moreover, explanations of the waterfall plot and power plot, as well as explanations of the contribution of the trait have been made using SHAP and LIME XAI methods.

4.1.2. Results

In this section, we provide the results of the important features explored for travel time forecast with help of the SHAP method applied to different ML models. As an example, Figure 6 and Figure 7 show the plots resulting from LightGMB and XGBoost models. They represent the overall impact of the features on the model output for different input datasets.

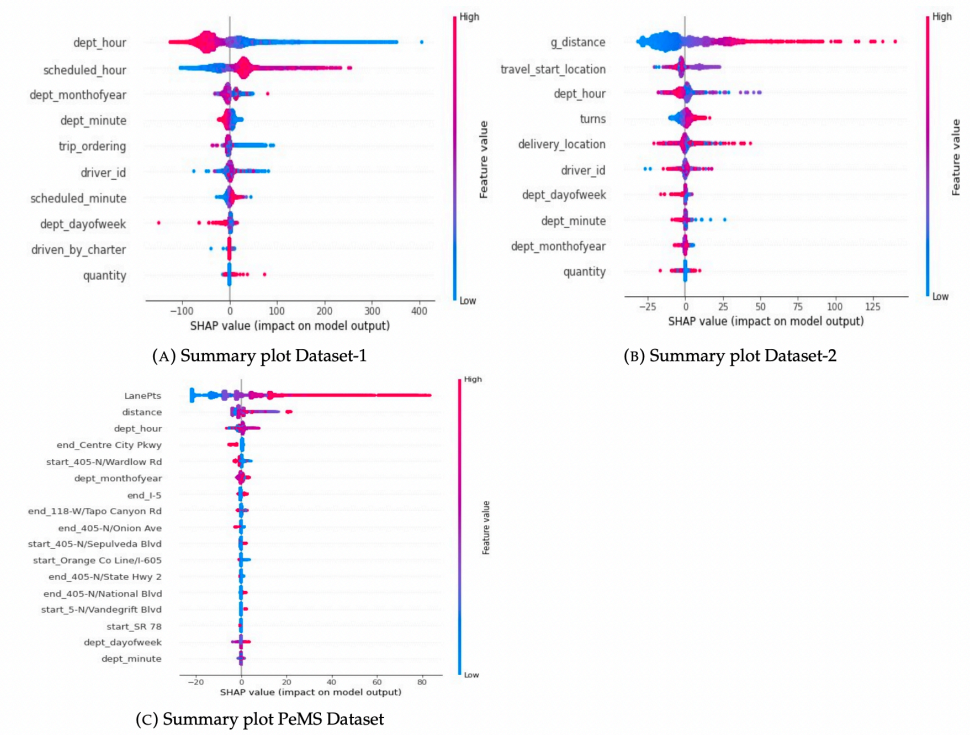


(A) Summary plot Dataset-1

(B) Summary plot Dataset-2

(C) Summary plot PeMS Dataset

Figure 6. Overall impact of features on model output using LightGBM model [46].



(A) Summary plot Dataset-1

(B) Summary plot Dataset-2

(C) Summary plot PeMS Dataset

Figure 7. Overall impact of features on model output using XGBoost model [46].

Important Features according to the SHAP summary plot remained the same for all models used: LightGMB, XGBoost, LSTM, bi-LSTM, GRU.

The feature list is as follows [46]:

dept_hour: Trip start hour is one of the most relevant variables in all three datasets, which suggests that trip start hour helps to travel time prediction regardless of whatever dataset or model we choose. Because the value of the feature is consistent across machine learning approaches, the explanatory AI method gains trust because it provides a reasonable justification for the prediction.

distance: Another element to consider is the trip distance. The connection between distance and travel time is linear, with travel time increasing as distance rises. As a result, this is one of the most significant features.

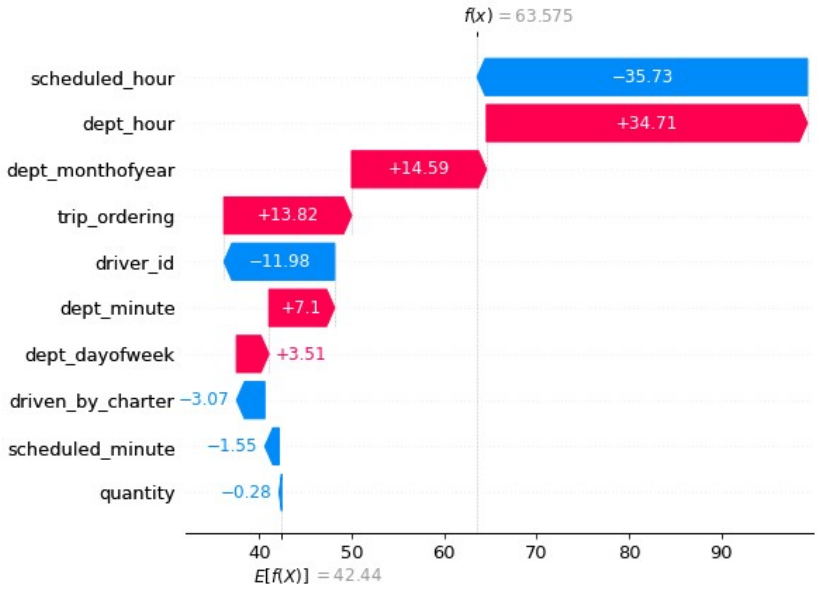
travel_start_location: The starting point is an important characteristic indeed, as different places can take longer or shorter to start, based on the locale and many other reasons, such as type of road, type of terrain, etc.

delivery_location: Similarly, the place of delivery is also just as important as where you start your journey.

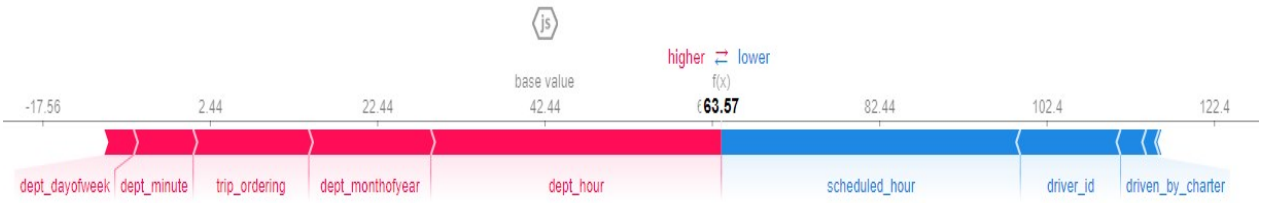
LanePts: This feature is only accessible in the PeMS dataset since it is gathered on highways. Furthermore, because motorway speeds differ from lane to lane, it may be assumed that a car traveling in the faster lane will take less time and vice versa.

Obviously, regardless of the machine learning model we utilize, the explained AI technique provides a comparable explanation, which is beneficial since it matches the resultant explanation, implying that the user can trust the XAI method's explanation.

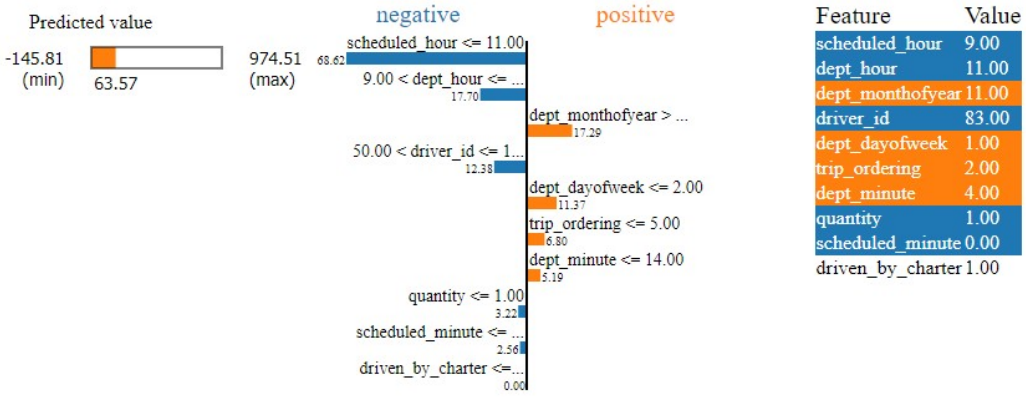
In addition, the SHAP and LIME methods were used to produce the waterfall graph, and the strength of the explanation graph is shown in Figure 8.



(A) Instance of SHAP waterfall plot for single prediction



(B) Instance of SHAP force plot for single prediction



(C) Instance of LIME feature contribution plot for single prediction

Figure 8. Plots describing the model's decisions [46].

Figure 8 shows three graphs for travel time forecast explanations using various methods. In plot A, we can observe that 'scheduled_hour' lowers the forecast by 35.73 minutes, whilst

’dept_hour’ raises it by 34.71 minutes. Plot B shows a similar picture of the explanation, but without the specific statistics that are displayed in the waterfall plot. Depending on the relevance of that element in that particular forecast, each feature pushes the prediction lower or higher. For the three random cases mentioned, it can also be observed that ’scheduled_hour’ and ’dept_hour’ are both more significant than any other feature.

The LIME approach may also be observed in action. The explanation in plot C is quite similar to the one in plot A. LIME is intrinsically local since it builds a local model around the unit of prediction value to explain, however, SHAP decomposes the prediction into contributions from each feature and adds the SHAP values for each feature to the final prediction, which is not the case with LIME.

4.2. Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain

Monje, Leticia, et al. made research on deep learning XAI for time series forecasting of bus passenger demand in Spain [47]. The goal of the study was to create a predictive and linguistically interpretable model that could be used to make decisions based on enormous amounts of data from various sources. They used a surrogate model and the 2-tuple fuzzy linguistic model to create an interpretable model from the LSTM neural network, which increases the linguistic interpretability of the resulting Explainable Artificial Intelligent (XAI) model without sacrificing precision.

4.2.1. What was done?

In their work, Monje et al. reviewed actual literature related to bus passenger flow prediction problems and concluded that most machine learning models used work as a black-box algorithm. After analyzing the existing XAI techniques, their purposes, advantages, and drawbacks they deduced to use a global model since the goal is to understand transport forecasting as a whole (at the bus line level). For the training, a black-box model dataset X was selected and predictions were made for that dataset. On dataset X and its predictions, the regression tree model was chosen and trained. The rules were

extracted from it, and the surrogate model's ability to replicate the black-box model's predictions was assessed. The variable was fuzzified in order to forecast it using the fuzzy 2-tuple linguistic model, and the surrogate fuzzy linguistic model was then interpreted.

The goal was to create a model that could be used to forecast and understand EMT's real passenger demand. They concentrated on the afternoon schedule of bus line 1 between January 2015 and February 2017. The following information was used:

- Date: The date is produced from the year, month, and day.
- Month: specifies the month (for example, January = 1).
- Holiday; 0: no, 1: yes.
- Day of week: denotes the weekday (e.g., Monday = 0).
- Passengers: the total number of travelers

After analyzing and comparing possible ML models to be used they chose LSTM as the most precise and effective one for the purpose of that research.

4.2.2. Results

LSTM model was built to predict passenger demand for buses with a high degree of precision. A surrogate tree was constructed and transformed into a 2-tuple. The following is the interpretation of surrogate rules produced for the LSTM model using 2-tuple linguistics [47]:

- Rule 1: Very high demand. Days that are neither weekends nor holidays, yet are not summer days. People go about their everyday routines on those days.
- Rule 2: High demand. There are a few days in September that aren't weekends or holidays. Because the majority of the population is now working and schools are beginning the academic year, passenger demand in Madrid rises in September.

For the following scenarios, passenger demand will be medium:

- Rule 3: In July, there are days that are neither weekends nor holidays. Students are not required to attend school on those days, while some employees begin their holidays from the second fortnight onwards. As a result, passenger demand begins to decline.
- Rule 4: Saturdays, when traveler demand was particularly strong on the eve. On Saturday afternoons, a considerable portion of the population does not work. Because public transportation is typically employed for leisure activities on certain days, passenger demand is lower than on eve. In the following scenarios, passenger demand will be low:
- Rule 5: In August, there are no weekends or days off. During the holidays, people appear to prefer to stay at home or go to different locations.
- Rule 6. Days in August that aren't weekends. Many businesses close in August, and employees are required to take mandatory vacations. Additionally, because students are not required to attend classes, the need for passengers decreases.
- Rule 7. Saturdays, when the day before's passenger demand was high/medium. Rule 4 has the same meaning.
- Rule 8. Sundays when passenger demand was low or very low the day before. On Saturdays and Sundays, people in Madrid do similar things, however, many street stores are closed on Sundays. As a result, passenger demand on these two days is equal, with Sundays having lower demand. Passenger demand will be very low in the following scenario:
- Rule 9. Sundays, when passenger demand was quite low the day before. Rule 8 has the same meaning.

The choice of using an interpretable surrogate model which was a regression tree model and a 2-tuple fuzzy linguistic model has allowed authors to get impressive results with a list of rules with explanations. The 2-tuple model, a fuzzy method, increased the language interpretability of the resulting XAI model without sacrificing accuracy.

5. Discussion

Overall, it is clear that XAI methods are highly demanded nowadays. Overviewing plenty of research papers and relevant literature allows us to make a summary of what user groups are there and what are the purposes of using AI explanations for them. Elicited groups of users are the following:

- Developers of models. Purpose: Fixing, improving, or debugging models.
- Business owners or administrators. Purpose: To evaluate the capability of AI applications, confirm the regulations, etc. to decide if they should be adopted and used.
- Decision-makers who are direct users of AI-based decision-making technologies. Purpose: To make informed decisions and build appropriate trust in the AI.
- Groups of people who can be influenced by AI and whose wellbeing could be impacted. Purpose: To go to court or challenge the AI.
- Regulatory bodies, auditing for compliance with legal or ethical standards such as fairness, security, confidentiality, etc.

To effectively implement XAI in your business or project you should define the key users and the purposes that XAI will suppose to manage. Figure 9 shows the key groups of XAI users.

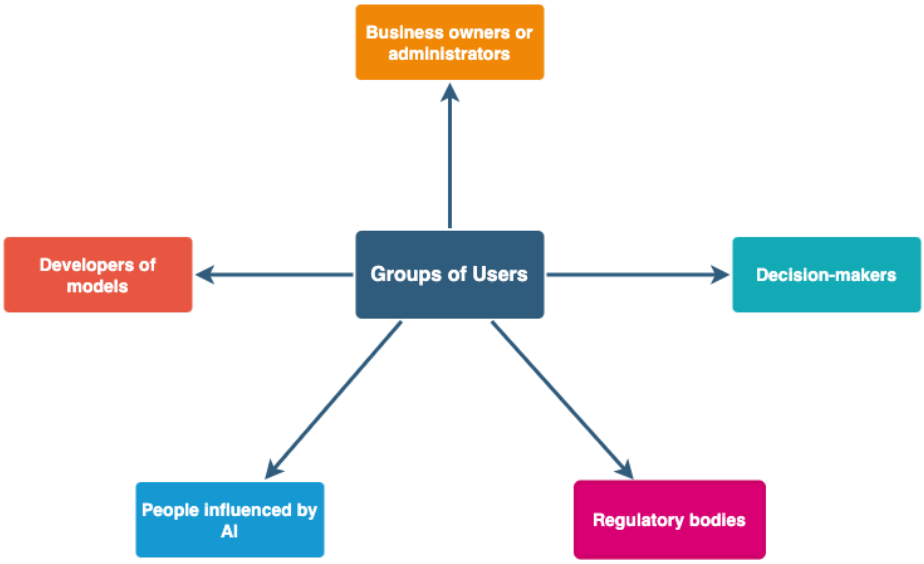


Figure 9. Groups of AI users

Different XAI methods are applicable in different cases. Not surprisingly, each method has its strengths and drawbacks. Thus, Table 2 shows the summary of XAI methods, its definition (Local or Global), and details about its advantages and disadvantages [48,49].

Table 2. Main features of different XAI techniques

Technique	Global or Local	Advantages	Disadvantages
Feature Importance	Global	Highly compressed global interpretation. Consider how features interact with one another.	It's unclear whether it can be applied to a training or a testing dataset.
Individual Conditional Expectation	Global	Intuitive and easy to understand	Plot can become too overcrowded to understand
Partial Dependence Plot	Global	Intuitive and clear interpretation	Assumption that characteristics are independent
Feature Interaction	Global	Detects all interactions been features	Computationally expensive
Global Surrogate Models	Global	Using the R-squared statistic, it's simple to assess the surrogate model's quality.	It's unclear what the appropriate R-squared cut-off is for trusting the surrogate model that results.

Local Surrogate Model (LIME)	Local	An easy-to-understand explanation. Explains the many forms of data.	It's possible that two highly similar points have very distinct interpretations.
Shapley Value Explanations	Local	Theorem of strong game theory is used to explain the problem.	Computationally very expensive

Another useful summary could be Table 3 provided by Liao et al. [50] which represents the XAI methods in compliance with the questions they could answer and also a brief description of what and how they explain. The list of the most relevant categories of questions is the following:

- How (global model-wide): what is the AI's basic logic or process it follows to get a global view.
- Why (a given prediction): what is the cause behind a certain prediction.
- Why Not (a different prediction): enquiring as to why the forecast differs from a desired or expected result
- How to be That (a different prediction): what are some options for changing the object to achieve a different result.
- How to Still Be This (the current prediction): what may be modified such that the object's prediction remains the same.
- What if: how do the input changes influence the prediction changes.
- Performance: What is AI's performance.
- Data: what is the training data.
- Output: what can be done or expected with the output of the AI.

The questions show that rather than focusing on defining the model's intrinsic properties, XAI should be described generally. More information regarding the model's performance, data, and output scope are frequently requested by users.

Table 3. A guide to showing the links between possible categories of user questions and examples of XAI methods for answering those questions [50].

Question	Ways to explain	Example XAI methods
How (global model-wide)	Describe the overall logic of the model in terms of feature effect, rules, or decision-trees+. If the user merely wants a high-level perspective, explain what the most important features or regulations are.	Prof Weight, Global feature importance, Global feature inspection plots, Tree surrogates
Why (a given prediction)	Detail how the instance's features, or crucial features, influence the model's forecast of it; or describe rules that the instance follows to ensure the prediction; or provide similar examples with the same anticipated result to support the model's prediction.	LIME, SHAP, LOCO, Anchors, ProtoDash+
Why Not (a different prediction)	Describe what features of the instance determine the present forecast and/or how the instance would get the alternative prediction with different adjustments. Alternatively, present paradigmatic cases with the opposite consequence.	CEM, Counterfactuals, ProtoDash (on alternative prediction)
How to Be That (a different prediction)	Highlight feature(s) that, if modified (increased, reduced, absent, or present), might shift the prediction to a different result with minimal effort. Show cases with little modifications that resulted in a different effect.	CEM, Counterfactuals, Counterfactual instances, DiCE
How to Still Be This (the current prediction)	Describe the features/feature ranges or rules that might ensure the same outcome. Show cases that aren't the same as the example but resulted in the same result.	CEM, Anchors
What if	Demonstrate how the prediction changes in response to the input change.	PDP, ALE, ICE
Performance	Provide information about the model's performance. Give each prediction a level of uncertainty. Describe the model's possible strengths and weaknesses.	Recall, Precision, Accuracy, AUC, F1; Communicate uncertainty of each prediction; See examples in Model Cards and FactSheets
Data	Give detailed information on the training data, including the source, origin, kind, size, population coverage, potential biases, and so forth.	Examples may be found in FactSheets and DataSheets.
Output	Describe the output or system functions' scope. Suggestions regarding how the output should be used for downstream activities or user workflow are encouraged if relevant.	Examples may be found in FactSheets and Model Cards.

The XAI techniques have been selected based on what is currently available in the open-source XAI library. The last three rows represent the wider needs of XAI, with no limits

for explaining model processes. This mapping guide can assist in identifying appropriate XAI methods based on users' questions.

Finally, the general steps for the average non-technical user to follow are presented in Figure 10 below.

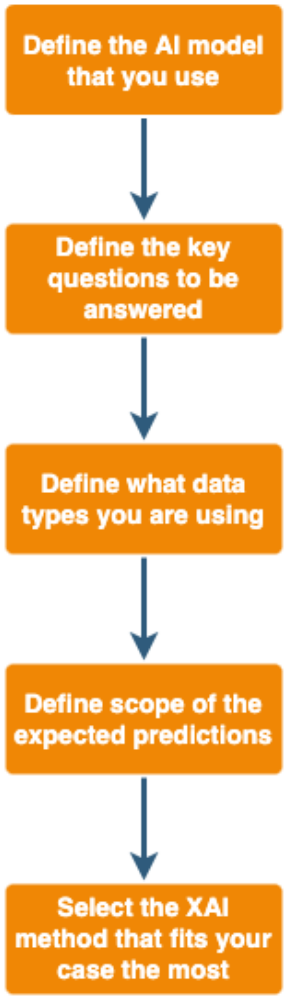


Figure 10. Proposed guidance on integrating the XAI into the existing project with AI model.

Following that guidance could help you to integrate the XAI into the existing project or business and improve the explainability of the AI results.

6. Conclusions

Building models and being able to make correct predictions are the most important challenges of today's world. In the long term, the current initiatives in artificial intelligence are aimed at assisting in the development, design, and implementation of human-centred AI systems whose agents would collaborate with humans in an interpretable and explainable manner to ensure fair decision-making as well as transparency.

In the first chapter, the introduction to the topic was done, highlighting the goal of the research and key research tasks.

In the second chapter, a detailed literature review was made to explore the essence of explainable artificial intelligence. The importance and motivation of its usage were shown. The metrics and strategies of model interpretations were overviewed and described in detail.

In the third chapter, popular XAI methods were overviewed in a detail.

The fourth chapter provides an overview of real-world cases of using XAI methods to solve transport problems. Real-world examples of XAI methods results were presented to highlight their differences and features.

Finally, in the discussion chapter, some suggestions and patterns were proposed to help potential AI users implement XAI methods.

Explainable AI methods can be extremely beneficial in a variety of industries. They can assist consumers in comprehending why they made a given decision. It can, for example, explain travel or delivery time projections, road or air traffic forecasts, fraud predictions, and so on. Explainable AI approaches might make prediction outputs more understandable by describing the influence of each input feature independently for every single prediction or even offering linguistic descriptions.

XAI is still a very modern and unstudied field that deserves great attention. It is very important to consider the need for the explainability of AI models so that the average user

or business can justifiably trust AI solutions. There are quite a few XAI methods available at the moment to address explainability, but nevertheless, this is an area that needs further development and improvement in the quality of explainability.

References

1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

2. Christoph Molnar «Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.» Springer; 1st ed. 2018.

3. Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI." Information fusion 58 (2020): 82-115.

4. Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.

5. The Essence of Explainable AI: Interpretability – 2019/ URL https://aibusiness.com/document.asp?doc_id=760945 (accessed 10.04.2022).

6. The Importance of Human Interpretable Machine Learning – 2018/URL https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476 (accessed 10.04.2022).

7. Lipton, Zachary C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.]

8. Gilpin, Leilani H., et al. "Explaining explanations: An overview of interpretability of machine learning." 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018.

9. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

10. Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

11.Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods." Entropy 23.1 (2020): 18.

12.Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai A review of machine learning interpretability methods." Entropy 23.1 (2020): 18.

13.Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

14.Van Hulle, Marc M. "Self-organizing Maps." (2012): 585-622.

15.Rosario, Barbara. "Latent semantic indexing: An overview." Techn. rep. INFOSYS 240 (2000): 1-16.

16.Doersch, Carl. "Tutorial on variational autoencoders." arXiv preprint arXiv:1606.05908 (2016).

17.Wattenberg, Martin, Fernanda Viégas, and Ian Johnson. "How to use t-SNE effectively." Distill 1.10 (2016): e2.

18.Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.1 (2012): 86-97.

19.Rosset, Saharon. "Model selection via the AUC." Proceedings of the twenty-first international conference on Machine learning. 2004.

20.Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

21.Aranganayagi, S., and Kȷ Thangavel. "Clustering categorical data using silhouette coefficient as a relocating measure." International conference on computational intelligence and multimedia applications (ICCIMA 2007). Vol. 2. IEEE, 2007.

22.Model Interpretation Strategies – 2018/URL https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

23.Predictive modeling: Striking a balance between accuracy and interpretability – 2016/ URL https://www.oreilly.com/content/predictive-modeling-striking-a-balance-between-accuracy-and-interpretability/ (accessed 12.05.2022)

24.Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, Elisabeth Andre “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. Journal on Multimodal User Interfaces volume 15, pages 87–98 (2021)

25.EU GDPR Recital 71 – 2018/URL https://www.privacy-regulation.eu/en/r71.htm (accessed 13.05.2022).

26.DOD Adopts Ethical Principles for Artificial Intelligence – 2020/URL https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

27.H.R.2231 - Algorithmic Accountability Act of 2019/URL https://www.congress.gov/bill/116th-congress/house-bill/2231/text

28.Mueller, Shane T., et al. "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI." arXiv preprint arXiv:1902.01876 (2019).

29.Friedman, Jerome H., and Bogdan E. Popescu. "Predictive learning via rule ensembles." The annals of applied statistics 2.3 (2008): 916-954.

30.Wright, Ray. "Interpreting black-box machine learning models using partial dependence and individual conditional expectation plots." *Exploring SAS® Enterprise Miner Special Collection* (2018): 1950-2018.

31.Oh, Sejong. "Feature interaction in terms of prediction performance." *Applied Sciences* 9.23 (2019): 5191.

32.Casalicchio, Giuseppe, Christoph Molnar, and Bernd Bischl. "Visualizing the feature importance for black box models." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2018.

33.Shapley, Lloyd S. "A value for n-person games, Contributions to the Theory of Games, 2, 307–317." (1953): 1953.

34.Mishra, Saumitra, Bob L. Sturm, and Simon Dixon. "Local Interpretable Model-Agnostic Explanations for Music Content Analysis." ISMIR. Vol. 53. 2017.

35.Peltola, Tomi. "Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback-Leibler projections." arXiv preprint arXiv:1810.02678 (2018).

36.Zafar, Muhammad Rehman, and Naimul Mefraz Khan. "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems." arXiv preprint arXiv:1906.10263 (2019).

37.Bramhall, Steven, et al. "Qlime-a quadratic local interpretable model-agnostic explanation approach." SMU Data Science Review 3.1 (2020): 4.

38.Shi, Sheng, Xinfeng Zhang, and Wei Fan. "A modified perturbed sampling method for local interpretable model-agnostic explanation." arXiv preprint arXiv:2002.07434 (2020).

39.Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

40.Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

41.Antwarg, Liat, et al. "Explaining anomalies detected by autoencoders using SHAP." arXiv preprint arXiv:1903.02407 (2019).

42.Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for a model explanation." International conference on machine learning. PMLR, 2020.

43.Aas, Kjersti, Martin Jullum, and Anders Løland. "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values." Artificial Intelligence 298 (2021): 103502.

44.Lundberg, Scott M., et al. "From local explanations to a global understanding with explainable AI for trees." Nature machine intelligence 2.1 (2020): 56-67.

45.García, María Vega, and José L. Aznarte. "Shapley additive explanations for NO2 forecasting." Ecological Informatics 56 (2020): 101039.

46.AHMED, Irfan, et al. "Travel Time Prediction and Explanation." (2021).

47.Monje, Leticia, et al. "Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain." Mathematics 10.9 (2022): 1428.

48.Elshawi, Radwa, Mouaz H. Al-Mallah, and Sherif Sakr. "On the interpretability of machine learning-based model for predicting hypertension." BMC medical informatics and decision making 19.1 (2019): 1-32.

49.Bueno, Itzcóatl, et al. "A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations." Information Sciences 589 (2022): 300-320.

50.Liao, Q. Vera, and Kush R. Varshney. "Human-Centered Explainable AI (XAI): From Algorithms to User Experiences." *arXiv preprint arXiv:2110.10790* (2021).