Computational Intelligence Methods and Applications

Rajanikanth Aluvalu
Mayuri Mehta
Patrick Siarry *Editors*

Explainable AI in Health Informatics

Springer

Computational Intelligence Methods and Applications

The monographs and textbooks in this series explain methods developed in computational intelligence (including evolutionary computing, neural networks, and fuzzy systems), soft computing, statistics, and artificial intelligence, and their applications in domains such as heuristics and optimization; bioinformatics, computational biology, and biomedical engineering; image and signal processing, VLSI, and embedded system design; network design; process engineering; social networking; and data mining.

Rajanikanth Aluvalu • Mayuri Mehta •
Patrick Siarry
Editors

Explainable AI in Health Informatics

Springer

Editors
Rajanikanth Aluvalu
Symbiosis Institute of Technology
Hyderabad Campus
Hyderabad, Telangana, India

Symbiosis International (Deemed University)
Pune, India

Mayuri Mehta
Department of Computer Engineering
Sarvajanik College of Engineering and Technology
Surat, Gujarat, India

Patrick Siarry
Lab. LiSSi
Universite Paris-Est Creteil
Vitry-sur-Seine, France

ISSN 2510-1765 ISSN 2510-1773 (electronic)
Computational Intelligence Methods and Applications
ISBN 978-981-97-3704-8 ISBN 978-981-97-3705-5 (eBook)
https://doi.org/10.1007/978-981-97-3705-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Artificial Intelligence (AI) has revolutionized the healthcare industry and has become an integral part of the system. AI-enabled autonomous systems are benefiting the healthcare industry. However, most machine learning and deep learning models are black models and are not explainable when the procedure goes wrong. A risk is still associated with unquestioningly trusting AI systems' recommendations, insights, or predictions. They operate as a black box, meaning users do not understand how such systems make decisions. Thus, the critical limitation of today's intelligent systems is their inability to explain their decisions and actions to human users. This issue is crucial for risk-sensitive healthcare applications, such as disease prediction, patient analytics, clinical decision support, surgery, etc. A lack of explainability hampers our capacity to fully trust AI systems.

This book is a collection of 12 chapters that provide an overview of recent advances in this area, that is, how explainable artificial intelligence techniques help provide trustworthy solutions in healthcare. The target audience of this book is researchers, practitioners, and students. A brief description of each chapter is given below.

Chapter 1 extensively discusses various fundamental underpinnings, methodologies, and implementation frameworks of XAI. The bedrock of XAI lies in the urgency to demystify the internal mechanisms of AI models, rendering their decision-making transparent for human stakeholders.

Chapter 2 addresses the capabilities of XAI frameworks to achieve accountability, transparency, result tracing, and model improvement. It discusses various XAI methods, use cases, and different application areas of XAI.

Chapter 3 discusses the use of deep learning in the real world and various datasets in the healthcare domain. It also discusses the transition from healthcare 1.0 to healthcare 6.0 and the use of XAI in various aspects of healthcare.

Chapter 4 introduces available XAI toolkits for experimental purposes and potential avenues for future developments in XAI. These aid researchers in exploring healthcare-oriented applications and contribute to advancing the healthcare 5.0 paradigm.

Chapter 5 explains the implementation of XAI methods using use cases on COVID-19 and cancer diagnosis. It discusses various post hoc methods in XAI and uses cases of using post hoc methods in XAI.

Chapter 6 discusses various applications of AI in disease diagnosis. It also discusses Artificial Intelligence-based design for automated drug z synthesizing. This chapter further stimulates additional research on developing and implementing artificial intelligence methods in drug discovery.

Chapter 7 discusses explainable AI and various big data control challenges. Further it discusses the post hoc explainable AI methods and the categories of explainability in XAI methods.

Chapter 8 casts a spotlight on the pivotal role of XAI within medical systems. It accomplishes this by delving into the essence of XAI, elucidating its various categories, exploring the algorithms harnessed to unveil concealed information within black-box systems, and addressing the challenges inherent to XAI. Furthermore, this chapter offers guidance to its readers on constructing intelligible deep learning models tailored for patient data analytics.

Chapter 9 discusses a proposal framework that will establish a standard for incorporating complex deep learning models into medical IoT devices. The implementation of this approach may offer a reliable and efficient diagnostic tool for the early detection of kidney abnormalities, which can lead to early interventions and better outcomes for patients. This research demonstrates the potential of AI-powered medical diagnostics to revolutionize medical care, particularly in detecting and treating kidney diseases.

Chapter 10 proposes an explainable DNN model for improved CRC detection utilizing stool-based microbiome data. The model employs a square root-based normalization method and a feature extension approach, incorporating customized normalization techniques to enhance prediction performance. These methods effectively address outliers, dominant features, and dimensionality challenges.

Chapter 11 proposes a deep convolutional neural network method called Decompose, Transfer, and Compose (DTC) that simultaneously estimates the present super- and fine-grained states. DTC can handle anomalies in the dataset by exploring its class limits using a class decay process.

Chapter 12 discusses how XAI is vital in diagnosing retinopathy, detecting skin cancer, and predicting ICU mortality. It discusses the complexities of deep learning algorithms that aid healthcare professionals in identifying retinal abnormalities and skin lesions, and forecasting ICU outcomes with utmost confidence. It also explains how XAI approaches illuminate the decision-making process, demystifying the "black box" and establishing a seamless link between AI and human expertise.

The editors are thankful to the authors who submitted their research work to this book and to all the anonymous reviewers for their insightful remarks and significant suggestions that helped enhance the book's quality. We trust that readers will find the book useful.

Hyderabad, Telangana, India — Rajanikanth Aluvalu
Surat, Gujarat, India — Mayuri Mehta
Vitry-sur-Seine, France — Patrick Siarry
18th November 2023

Contents

Abbreviations

16S rRNA	16S ribosomal RNA
2AVB-T1	Second degree atrioventricular block Mobitz type I with Wenckebach phenomenon
2AVB-T2	Second degree atrioventricular block Mobitz type II
2D	Two-dimensional
3D	Three-dimensional
abs-diff	Absolute difference
abs-shift	Absolute shift
ACC	Accuracy
ADASYN	Adaptive synthetic
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
AEHR	Advanced electronic health record
AF	Atrial fibrillation
AI	Artificial intelligence
AL	Accumulated local
AR	Augmented reality
AR/VR	Augmented reality/virtual reality
AUC	Area under the curve
AUROC	Area under the receiver operating characteristic
BB	Black-box models
BiLSTM	Bidirectional long short-term memory
BMI	Biomass index
CAM	Channel attention module
CAVB	Complete atrioventricular block
CD	Crohn's disease
CDSS	Clinical decision support system
CM	Counterfactual method
CNN	Convolutional neural network
CRC	Colorectal cancer
CT	Computed tomography
CV	Cross validation

CVD	Cardiovascular disease
CX-ToM	Counterfactual explanations with the theory-of-mind
DD	Disease diagnosis
DEC	Decompose, transfer, and compose
DL	Deep learning
DNN	Deep neural network
DRD	Deep radio mic descriptors
DT	Decision tree
DTC	Decision tree classifier
ECG	Electrocardiogram
EHR	Electronic health records
EWS	Early warning scores
F1	F1 score
Fig	Figure
FNIRS	Functional near-infrared spectroscopy
FPR	False positive rates
FSM	Finite state model
GAM	Generalized additive model
GBM	Gradient boosting machine
GLM	Generalized linear model
GPU	Graphical processing units
Grad-CAM	Gradient-weighted class activation mapping
GW-OLS	Geographically weighted OLS regression
HFSM	Hierarchical finite state machine
HFSM	Hierarchical fuzzy sets model
ICE	Individual CONDITIONAL EXPECTATION
ICU	Intensive care unit
IIM	Intrinsically interpretable methods
IoMT	Internet of military things
IoMT	Internet of medical things
IoT	Internet of things
ISIC	International Skin Imaging Collaboration
JR	Junctional rhythm
LBC	Lung and bronchus cancer
LD	Local dependence
LIME	Local interpretable model-agnostic explanations
LOS	Length of staying
LR	Logistic regression
LRP	Layer-wise relevance propagation
LSTM	Long short-term memory
MADN	Median absolute deviation normalization
MAE	Mean absolute error
MAM	Model agnostic methods
MELLODDY	Machine learning ledger orchestration for drug discovery
MI	Mutual information
min-max	Minimum-maximum

ML	Machine learning
MRI	Magnetic resonance imaging
MTL	Multi-task learning
NHAI	National Highway Authorities of India
NLP	Natural language processing
NN	Neural network
NSR	Normal sinus rhythm
OTU	Operational taxonomic unit
PDA	Patient data analytics
PDP	Partial dependence plot
PDP	Partial Dependency Plot
PEARS	Pregnancy Exercise And Nutrition Research Study
PM	Pacemaker rhythm
PRE	Precision
PROTACs	Proteolysis-targeting chimaeras
QSAR	Quantitative structure-activity relationship
RAM	Random access memory
RAS	Robot-assisted surgery
REC	Recall
RF	Random forest
RLE	Relative log expression
RMSE	Root mean squared error
RMSprop	Root mean square
RNN	Recurrent neural network
ROI	Region of interest
SAM	Spatial attention module
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic minority over-sampling technique
sqrt-prod	Square root product
sqrt-sum	Cube root-sum
SVM	Support vector machines
SVT	Supraventricular tachycardia
TB	Tuberculosis
TML	Transformational machine learning
TMM	Trimmed mean of *M*-values
TPR	True positive rates
UAV	Unmanned arial vehicles
UC	Ulcerative colitis
USMS	Universal patient-side manipulators
VT	Ventricular tachycardia
XAI	Explainable artificial intelligence
XDM	Explainable deep learning model
XML	Explainable ML
ZSM	Zero service management
ZSN	*Z*-score normalization

Introduction to Explainable AI

Amit Ganatra, Brijeshkumar Y. Panchal, Devarshi Doshi, Devanshi Bhatt, Jesal Desai, Bijal Talati, Neha Soni, and Apurva Shah

Abstract Explainable AI (XAI) has emerged as an essential realm aimed at tackling the opacity of intricate AI models and nurturing confidence in their judgments. This study extensively investigates the fundamental underpinnings, methodologies, and practical implementations of XAI. The bedrock of XAI lies in the urgency to demystify the internal mechanisms of AI models, rendering their decision-making transparent for human stakeholders. Within the domain of XAI, diverse methodologies encompass a spectrum of techniques such as interpretable models, scrutiny of feature significance, localized and holistic elucidations, visual representations, and explications in natural language. These methodologies collectively foster intelligibility and amplify the explicable nature of AI models. This research significantly enriches the expanding reservoir of scholarly exploration by clarifying the core tenets of XAI. This comprehensive survey unmistakably demonstrates that XAI assumes a pivotal role in bridging the chasm between intricate AI processes and human comprehension. Consequently, it clears the path for a more reliable and efficacious partnership between human intellect and mechanical ingenuity.

A. Ganatra
Parul University, Limda, Waghodia, Vadodara, Gujarat, India

B. Y. Panchal (✉) · N. Soni
Computer Engineering Department, Sardar Vallabhbhai Patel Institute of Technology (SVIT), Vasad, Anand, Gujarat Technological University (GTU), Ahmedabad, Gujarat, India

D. Doshi · D. Bhatt · J. Desai
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India

B. Talati
Department of Computer Science and Engineering, Parul Institute of Technology [PIT], Parul University, Limda, Waghodia, Vadodara, Gujarat, India

A. Shah
Computer Science and Engineering Department, Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

R. Aluvalu et al. (eds.), *Explainable AI in Health Informatics*, Computational Intelligence Methods and Applications,
https://doi.org/10.1007/978-981-97-3705-5_1

Keywords Explainable AI · Artificial intelligence · Deep learning · XAI · Healthcare · Automation

1 Introduction

AI stands for Artificial Intelligence. It is a division of computer science that helps in developing or designing a system that can show an intelligent behavior. At the core level, AI is the science and engineering of making intelligent machines, particularly computer programs or systems. In recent years, an AI technique has been successfully engaged to solve a wide variety of real-life problems related to health care, finance, transportation, defense, weather forecasting, etc.

With the advancements in Artificial Intelligence, Humans will have a harder time understanding and retracing the algorithm's steps to a decision. The entire calculating process is transformed into a "black box" that is impossible to understand. These black-box models are built from raw data. One of the major issues with the traditional AI approach lies in the implementation of machine learning techniques. That is, one cannot blindly trust the prediction or the output of the machine learning model as that might have drastic consequences.

What is a solution to this problem? Explainable AI (XAI). It addresses the challenge of establishing trust in machine learning models. XAI stands for Explainable Artificial Intelligence. Explainable Artificial Intelligence (XAI) is a collection of methods and tactics that enable human users to understand and trust the outcomes and productivity of machine learning algorithms. The term "Explainable AI" pertains to the foreseeable impact of a model and its potential biases. In AI-assisted decision-making, it contributes to the calculation of model correctness, fairness, transparency, and results. The dimensions of an organization are crucial when it comes to incorporating AI models into operational use. An organization's adoption of a responsible AI development approach is also aided by AI's explainability.

It is an emerging artificial intelligence approach. It is also known as transparent artificial intelligence. It indicates that in XAI, one must be able to understand how and why the algorithm makes decisions or predictions. In other words, the system can justify the result that it produces. Within the realm of explainable AI, outcomes or solutions are comprehensible to humans. This is in contrast to the opaque methodology of machine learning, where even the creator or developer of the model is unable to elucidate the rationale behind specific decisions made by the AI system. Explainable artificial intelligence delivers overall data about how an artificial intelligence program decides by disclosing the merits and demerits of the program or a model, the specific criteria that has been used by the program to produce the result. It also assists in understanding why a program produces a specific result as opposed to its substitutes, which induces a level of trust that's proper for many types of decision, what type of error the program is prone to, and how the error can be modified.

There are various benefits of understanding how an AI-enabled system has arrived at a certain conclusion. Explainability can help developers ensure that the system is working as planned, it may be necessary to meet regulatory standards, or it may be crucial in allowing those who are affected by a decision to challenge or change the decision. To provide clarity on this absence of consensus, it would be resourceful to cite D. Gunning's definition of the term Explainable Artificial Intelligence (XAI): "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

This interpretation amalgamates two concepts that require prior discussion. Nonetheless, it overlooks supplementary factors contributing to the requirement for interpretable AI models, such as causality, transferability, informative attributes, equity, and assurance.

Let's look at the existing and future scenarios. Currently, we are using an artificial intelligence technique in which the choice or suggestion generates several questions, such as why did the model do this? Why not try something new? When does the system produce 100% accurate results? When will the system fail? When can I put my faith in this machine learning model? How can I fix the system's erroneous result? However, unlike the current artificial intelligence method, explainable artificial intelligence will include an AI explanatory model in addition to an explanatory interface is employed to assist in understanding the aspect of artificial intelligence that pertains to both the reasons behind decisions and the reasons for certain decisions not being made. The explanation module and explanation interface will also assist in understanding when the system will succeed and when one must trust this system. In this book chapter, concepts such as black-box models, transparency, XAI tools and techniques, and many more topics will be covered in detail.

2 What Are Black-Box Models?

These black-box models are shaped by a machine learning algorithm directly from data, which implies that no one, even the developers, knows how variables are joint to produce forecasts [1]. Even if one knows a list of input variables, black box predictive models might be such intricate functions of the variables that no social can understand how the variables interact to produce a final forecast. Figure 1 shows the basic diagram of black box.

Interpretable models, which share the same mathematical equivalence as black-box models but can be more ethically sound, differ in their approach by constraining themselves to offer a more profound comprehension of prediction mechanisms. In certain instances, when a compact and valid logic encompasses only a handful of variables, or when utilizing a linear model where variables are assigned weights and combined, the connection between variables and the ultimate forecast can be exceptionally lucid. Decomposable models are frequently employed to generate easily understandable models, or additional limitations are introduced to impart a heightened level of insight. In contrast, many machine learning models prioritize high predictability on static datasets over readability [3] (Fig. 2).

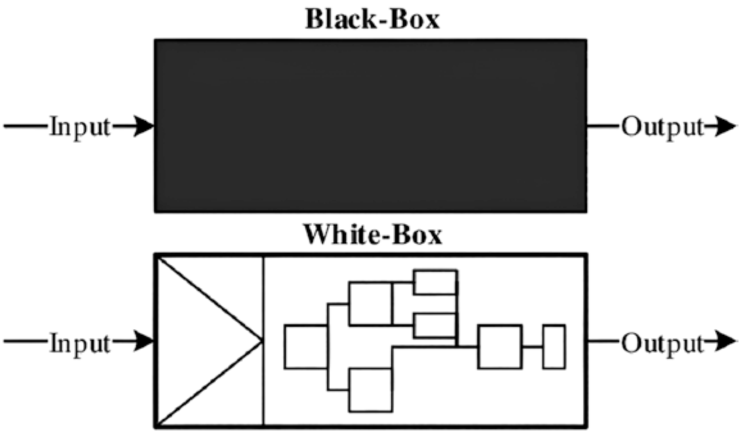

Fig. 2 Black box vs white box model [2]

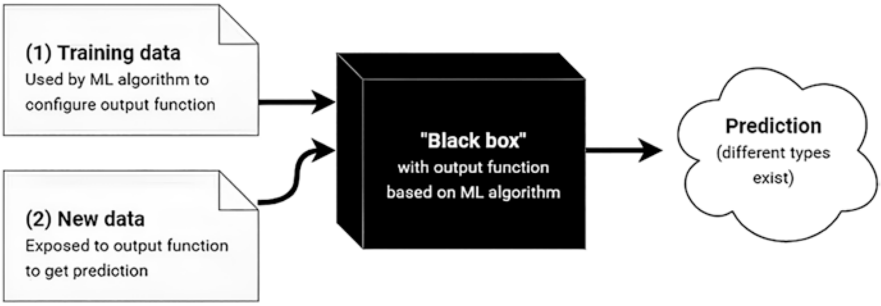

Fig. 1 What are black-box models? [2]

Conversely, certain machine learning models are not designed with the intention of overcoming interpretational challenges; their primary purpose is to deliver precise forecasts based on fixed data entries that might or might not mirror the model's practical application.

2.1 Why Is Model Interpretability Important?

The interpretability of a machine learning algorithm depends on how easy it is for humans to grasp the procedures that it takes to arrive at its results. Earlier, Artificial Intelligence (AI) algorithms were known as "black boxes," with no means of knowing what was going on inside and making it impossible to explain the results to regulators and stakeholders.

It is wrong to believe that accuracy must be compromised for interpretability. When extremely basic interpretable models exist for the same tasks, it has allowed corporations to promote and sell proprietary or sophisticated black-box models for high-stakes judgments. As a result, it permits the model's developers to profit while ignoring the negative repercussions for the people who are affected. A minority of individuals contest these models, as their developers hold the view that complexity is a prerequisite for accuracy. The Explainable Machine Learning Challenge of 2018 provides an illustrative example of evaluating the advantages and disadvantages of opaque models compared to transparent models.

When utilizing the results of an algorithm to make high-stakes judgments, it is critical to understand which factors the model took into consideration and which it did not. Furthermore, if a model is not easily interpretable, the company may not be able to use its insights to make process adjustments lawfully. In tightly regulated sectors like banking, insurance, and healthcare, understanding the factors that contribute to anticipated outcomes is critical in order to comply with regulations and industry best practices.

For a variety of other reasons, interpretability is essential. For example, if researchers do not grasp how a model works, they may have trouble translating their findings to a larger knowledge base. Interpretability is also necessary for avoiding embedded bias and debugging an algorithm. It also assists scholars in decisive the impact of trade-offs in a model. More concisely, as algorithms play a larger role in society, knowing how they arrive at their conclusions will become crucial gradually.

Currently, scholars must compensate for inadequate interpretability through judgment, expertise, observation, monitoring, and careful risk management, which includes a full grasp of the datasets they utilize. Regardless of the machine learning model, there are a number of strategies for improving interpretability.

2.2 Using Explainable AI to Decipher Black-Box Machine Learning Models

Machine Learning (ML) and Artificial Intelligence (AI) have surged in popularity, finding utility across diverse sectors. However, they have also encountered escalating critique due to concerns about the reliability of their decision-making. Certain ML systems, especially Deep Neural Networks (DNNs), are often labeled as enigmatic entities because comprehending their inner workings post-training proves arduous. This opacity impedes a full grasp and explication of a model's reasoning process. Nevertheless, the provision of explanations is indispensable to establish the dependability of a model's predictions. This assumes paramount importance when machine learning algorithms underpin decision support systems in sensitive domains. Explanations not only corroborate the precision of a model's prognoses but also play a pivotal role in preempting inadvertent errors and unearthing potential biases. Furthermore, they facilitate a comprehensive comprehension of a model, which is imperative for prospective enhancements and rectification of its limitations.

Explainable AI (XAI) tackles the quandary of furnishing explanations for models that surpass human understanding due to their intricacy. These explanations span from individual (local) explications elucidating specific outcomes of black-box models—such as unraveling the rationale behind a denied loan application or an erroneous image classification—to collective (global) explanations that unveil broader patterns within such opaque models. These comprehensive explanations can address queries like identifying the most influential risk factor for a certain type of cancer.

3 Transparency in Machine Learning Models

Transparent machine learning is introduced as a new kind of machine learning which explains itself. This means that it tells us how it works, its predictions, its insights—so that the user can understand and trust the outcome. If addressed, this technology might be the best-case scenario for AI system safety and security in the future [4].

Models created by current machine learning (ML) techniques are difficult or impossible to comprehend. Security, safety, and prejudice are all issues that these deployments face. Insight into the automated decision-making process is also difficult with opaque models.

Transparent machine learning aims to tackle these issues by creating understandable models and data. It would accomplish this by displaying and altering source code representations. Consequently, you would have a possibly self-contained executable that could be used right away.

It is critical that Transformational Machine Learning (TML) systems use well-known programming languages and data formats that are simple to comprehend. Furthermore, the source code and data it generates in those languages and formats must be clear enough for an engineer of acceptable competence to understand and modify it. This is a fundamental principle that should take precedence over all other factors, even if it means sacrificing model efficiency. Later, recommendations will be made on how to achieve both efficiency and readability without permanently abandoning either.

3.1 Long-Term Objectives

Transformational Machine Learning (TML) is primarily oriented toward enhancing one or more dependently typed programming languages that possess robust specification support. This would involve the incorporation of elements from the comprehensive deep specification initiative, which strives to meticulously validate the complete developmental continuum, spanning from applications to the operating system and extending down to the hardware level [5].

Fig. 3 Transparency in machine learning models [6]

Furthermore, the development of source code adheres to long-term quality aspirations, including:

- Support for multiple language targets
- Comprehensive and concise commenting
- Incorporation of high-level abstractions
- Mitigation of unnecessary complexity
- Utilization of accelerated hardware for improved performance

3.2 Short-Term Objectives

The immediate objectives are:

1. Develop a transparent machine learning method that works.
2. Has it generated readable source code?

Achieving a viable proof-of-concept poses challenges. Identifying Transparent Machine Learning (TML) systems that match or surpass the performance of leading Machine Learning (ML) models would serve as a positive initial step. This rationale is supported by the potential to invigorate research enthusiasm. Nonetheless, prioritizing comprehensibility remains paramount; disregarding this aspect would undermine the project's objective, as incomprehensible source code resembles another variation of an opaque model. As depicted in Fig. 3, the illustration portrays transparency within ML models.

3.3 Theoretical Limits

It is crucial to distinguish between program readability and program comprehension ease. This does not mean that the usual criteria of reasonable competence established in the TML definition is irrelevant. Its purpose is to promote discussion of TML source model theoretical restrictions at the intersection of maximal model complexity and perfect human understanding.

Two complementary definitions of AI's foundations will be examined to aid discussion:

- *Inexplicability*: There is no explanation that is both 100% correct and understandable to humans for some judgments made by an intelligent machine.
- *Incomprehensibility*: Certain intelligent system judgments will have a 100% true explanation that no human can fully comprehend.

To address these statements, we must first explore the difference between *opaque* and *transparent* machine learning. The explanation is inextricably linked to the model in TML because the model is the explanation. Also included in that model might be a description of the TML system.

3.3.1 Layer-Wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP) is a method that gives potentially complicated deep neural networks like explainability and scalability. It works by applying a set of specially developed propagation rules to propagate the prediction backward through the neural network.

3.3.2 Counterfactual Method

The counterfactual impact evaluation approach enables for determining how much of the observed real change (e.g., a rise in income) may be attributed to the intervention's influence (since such improvement might occur not only due to the intervention but also due to other factors, e.g., overall economic growth).

3.3.3 Local Interpretable Model-Agnostic Explanations (LIME)

The term LIME stands for Local Interpretable Model-Agnostic Explanations, representing key aspects of the explanation process. "Local fidelity" refers to the objective of ensuring that the explanation faithfully reflects the classifier's behavior in the vicinity of the instance under prediction.

3.3.4 Generalized Additive Model (GAM)

A Generalized Additive Model (GAM) in statistics extends the concept of a generalized linear model. Within a GAM, the linear predictor is intricately connected to smooth functions of specific predictor variables. The primary emphasis lies in making inferences regarding these smooth functions. GAMs were conceived by Trevor Hastie and Robert Tibshirani to amalgamate the advantages found in both generalized linear models and additive models.

$$g\left(E\left(Y\right)\right) = \beta_0 + f_1\left(x_1\right) + f_2\left(x_2\right) + \cdots + f_m\left(x_m\right) \tag{1}$$

In this framework, a solitary response variable (Y) is linked to particular predictor variables (x_i). The distribution governing Y belongs to the exponential family, encompassing distributions like normal, binomial, or Poisson. A link function (g), such as the identity or logarithmic function, establishes a connection between the predicted value of Y and the predictor variables. The model structure incorporates functions (f_i) that can be parametrically defined, perhaps as polynomials or unpenalized regression splines of a variable. Alternatively, these functions can be estimated non-parametrically, taking the form of smooth functions and relying on non-parametric techniques. To illustrate, within a conventional GAM, the function $f_1(x_1)$ might leverage techniques like scatterplot smoothing, such as locally weighted means. Conversely, $f_2(x_2)$ could involve a factor model associated with x_2. This adaptive nature enables non-parametric adjustments, minimizing assumptions about the genuine relationship between outcomes and predictors. While this adaptability can potentially enhance data fitting when compared to entirely parametric models, it does come at the expense of simplified interpretation.

3.3.5 Rationalization

AI rationalization involves the creation of explanations for the behavior of autonomous systems, simulating human-like reasoning. In this context, we introduce a rationalization technique that employs neural machine translation to convert an autonomous agent's internal state-action representations into everyday language. To validate our approach, we implement it within the Frogger gaming environment. Our objective is to train an autonomous game-playing agent to express its chosen actions using natural language. To build the training dataset, we utilize insights from human players who articulate their thoughts while playing.

We advocate for the adoption of rationalization as a strategy for generating explanations and present the outcomes of two studies that assess its effectiveness. The results underscore the efficacy of neural machine translation in generating rationalizations that faithfully capture agent behavior. Furthermore, the findings suggest that rationalizations are more favorably received by humans in comparison to alternative explanation methods.

3.4 Framework and Tools

Explainable AI is a new and developing discipline in the realm of artificial intelligence and machine learning. It's critical to establish human trust in AI models' choices. It's only conceivable if the dark box of machine learning models is made more transparent. Explainable AI frameworks are programmed that create reports on how a model works and attempt to explain how it works. Now we'll talk about six AI frameworks that are easy to understand.

3.4.1 SHAP

SHAPley Additive Explanations, commonly referred to as SHAP, is an abbreviation for SHapley Additive Explanations. It serves as a versatile tool for elucidating various machine learning algorithms, ranging from fundamental ones like linear regression, logistic regression, and tree-based models to more intricate models like deep learning models used in tasks such as image classification, captioning, and even in NLP tasks like sentiment analysis, translation, and text summarization. This approach is model-agnostic and harnesses Shapley values derived from game theory to illuminate model behaviors. Essentially, it unveils how diverse attributes impact the model's output and the role they play in shaping the ultimate outcome. This concept is visually represented in Fig. 4.

3.4.2 LIME

LIME is short for Local Interpretable Model-agnostic Explanations. While it shares similarities with SHAP, it boasts greater computational efficiency. LIME provides a collection of explanations that elucidate the contribution of each attribute in predicting outcomes for specific data samples, visualized in Fig. 5. Notably, LIME is versatile enough to handle any black-box classifier with two or more classes. The classifier simply needs to furnish a function capable of processing raw text or a numpy array, delivering the probabilities associated with each class. It's worth noting that Scikit-learn classifiers are already integrated with this capability.

3.4.3 ELI5

ELI5, an abbreviation for "Explain Like I'm 5," is a Python library crafted to simplify the troubleshooting and explication of machine learning classifiers. It extends its support to various machine learning frameworks, including but not limited to scikit-learn, Keras, XGBoost, LightGBM, and CatBoost.

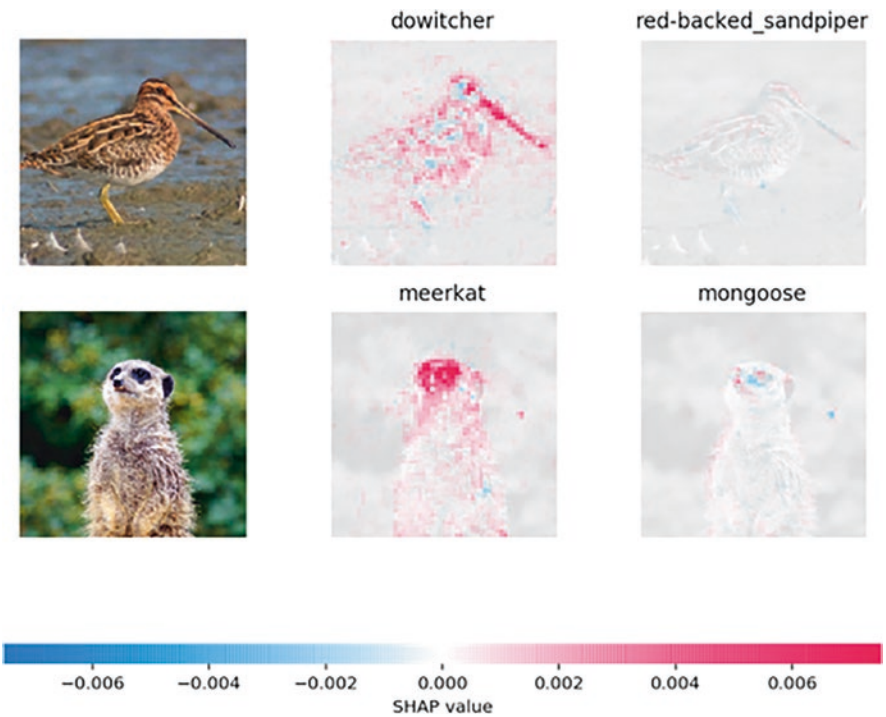

Fig. 4 Example of image classification [7]

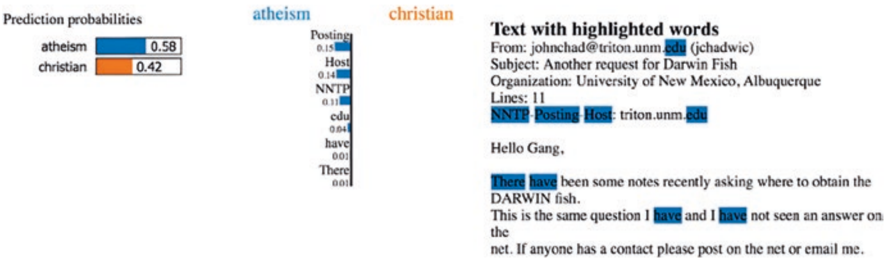

Fig. 5 Screenshot of explanations given by LIME [8]

3.4.4 What-if Tool

The What-if Tool (WIT), developed by Google, serves the purpose of enhancing the understanding of how machine learning models operate. WIT empowers users to simulate scenarios, assess the significance of distinct data attributes, and visually comprehend model behavior across diverse models and subsets of input data. This tool is also proficient in handling several machine learning fairness measures. Available as an extension within Jupyter, Colaboratory, and Cloud AI Platform notebooks, WIT covers tasks ranging from binary classification and multi-class