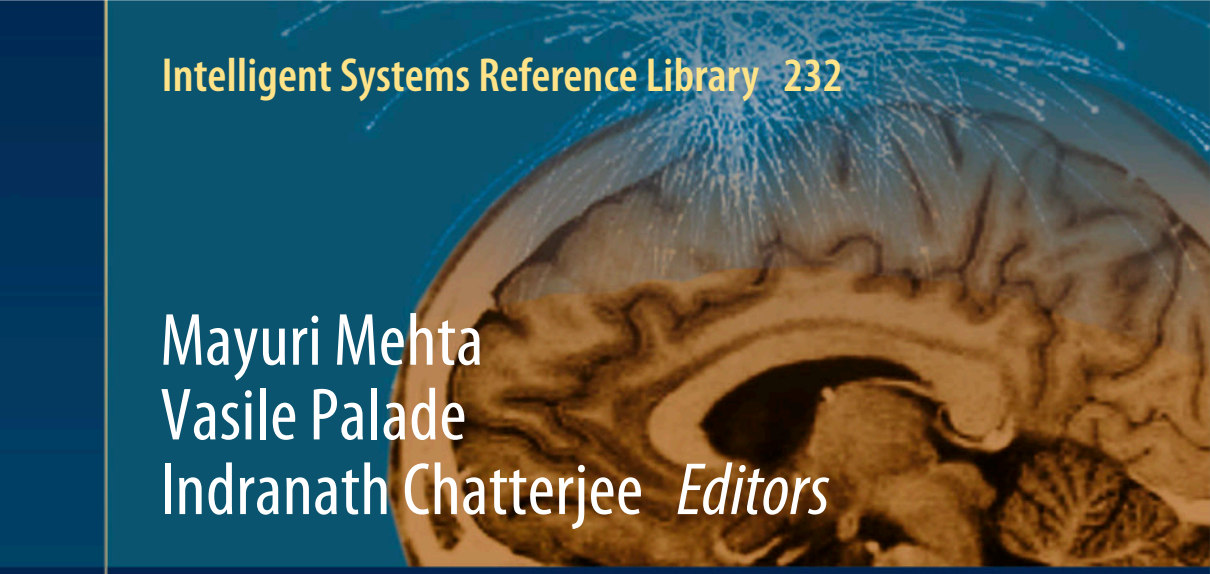
Intelligent Systems Reference Library 232



Mayuri Mehta
Vasile Palade
Indranath Chatterjee *Editors*

Explainable AI: Foundations, Methodologies and Applications

Springer

Intelligent Systems Reference Library

Volume 232

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

Indexed by SCOPUS, DBLP, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Mayuri Mehta · Vasile Palade · Indranath Chatterjee
Editors

Explainable AI: Foundations, Methodologies and Applications

Springer

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

Vasile Palade
Centre for Computational Science
and Mathematical Modelling
Coventry University
Coventry, UK

Indranath Chatterjee
Department of Computer Engineering
Tongmyong University
Busan, Korea (Republic of)

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-031-12806-6 ISBN 978-3-031-12807-3 (eBook)
https://doi.org/10.1007/978-3-031-12807-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial Intelligence (AI) has brought about a revolution in many real-world sectors and has become an integral part of our everyday lives. While AI-enabled systems are undoubtedly benefiting real-world sectors, there is still a risk in blindly trusting the recommendations, insights, or predictions provided by them. Many of these systems are often complex and opaque. They operate as a black box, meaning that users do not understand how decisions are being made by such systems. Thus, the key limitation of today's intelligent systems is their inability to explain their decisions and actions to human users. This issue is especially important for risk-sensitive applications, such as security, clinical decision support, or autonomous driving. A lack of explainability hampers our capacity to fully trust AI systems.

It is for this reason that AI techniques need to have explanatory capabilities, for users to understand why certain decisions are made. The methods developed to provide such capabilities have come to be known as explainable AI (XAI). Explainable AI contrasts with the so-called 'black box' machine learning. XAI helps present decisions being made with additional information about how and why the AI system arrived at a particular decision, including an interface to explain which features influenced its decision.

In this book, readers will learn about Explainable AI, including what it is, what the fundamentals of this area are, why it is needed, and how it is to be developed. Explainable AI offers a way to make decision-making more transparent and trustworthy. In other words, XAI aims to remove the so-called black box from the AI models being developed and explain the model decisions in an understandable form. It refers to an AI system's capacity to explain the logic behind its action to a human person. It can take two forms: explaining it to a computer scientist in a specialized language or explaining it to the system user in a human understandable form. It is critical because it is intimately related to human confidence in the AI system's usage and, more formally, whether that faith is well-placed by verifying things about the machine's behavior.

This book covers concepts related to model transparency, interpretable machine learning and explanations, various methods for Explainable AI, evaluation methods and metrics for XAI, ethical, legal, and social issues related to AI and XAI, as well

as a range of applications and examples of XAI in different real-life sectors, such as healthcare, autonomous driving, and law enforcement. The editors are thankful to the authors who submitted their research work to this book, as well as to all the anonymous reviewers for their insightful remarks and significant suggestions that helped enhance the quality of this book. We hope that readers will find the book useful.

Surat, India — Mayuri Mehta
Coventry, UK — Vasile Palade
Busan, Korea (Republic of) — Indranath Chatterjee
June 2022

Contents

Contributors

Aluvalu Rajanikanth CBIT, Hyderabad, India

Apostolopoulou Eleni School of Science, Technology and Health, York St John University, York, UK

Banerjee Puja Academy of Scientific and Innovative Research, Ghaziabad, India

Barnwal Rajesh P. AI & IoT Lab, IT Group, CSIR-Central Mechanical Engineering Research Institute, Durgapur, India

Bhattacharya Tanmay Department of IT, Techno Main, Kolkata, India

Brcic Luka Medical University Graz, Graz, Austria

Carvalho Rita Medical University Graz, Graz, Austria

Chennam Krishna Keerthi Vasavi College of Engineering, Hyderabad, India

Chopra Mayank Department of Computer Science and Informatics, Central University of Himachal Pradesh (H.P.), Dharamshala, India

Deshmukh Rashmi Department of Technology, Shivaji University, Kolhapur, India

Evans Theodore Medical University Graz, Graz, Austria

Eyo Eyo Umo Faculty of Environment and Technology, Civil Engineering Cluster, University of the West of England, Bristol, UK

Geißler Christian Medical University Graz, Graz, Austria

Goel Amit Kumar Delhi, India

Holzinger Andreas Medical University Graz, Graz, Austria

Hore Sirshendu Department of CSE, Hooghly Engineering and Technology College, Pipulpati, Hooghly, West Bengal, India

Jansen Christoph Medical University Graz, Graz, Austria

Kanarachos Stratis Faculty of Engineering and Computing, Coventry University, Coventry, UK

Kargl Michaela Medical University Graz, Graz, Austria

Klyushin D. A. Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Kumar Ajay Department of Computer Science and Informatics, Central University of Himachal Pradesh (H.P.), Dharamshala, India

Kumar Dheeraj Department of Information Technology, Parul Institute of Engineering and Technology, Parul University, Vadodara, India

Lu Yang School of Science, Technology and Health, York St John University, York, UK

Maheswari V. Uma KG Reddy College of Engineering, Hyderabad, India

Mehta Mayuri A. Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Sarvajanik University, Surat, India

Mudrakola Swapna Vasavi College of Engineering, Hyderabad, India;
Matrusri Engineering College, Hyderabad, India

Müller Heimo Medical University Graz, Graz, Austria

O. Chergykalo D. Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Onyekpe Uche School of Science, Technology and Health, York St John University, York, UK;
Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK

Palade Vasile Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK

Plass Markus Medical University Graz, Graz, Austria

Rao K. Gangadhara CBIT, Hyderabad, India

Regitnig Peter Medical University Graz, Graz, Austria

Sarkar Arjun Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Jena, Germany

Sarnayak Anshul Department of Technology, Shivaji University, Kolhapur, Maharashtra, India

Sharma Mukta Delhi, India

Shringare Yash Department of Technology, Shivaji University, Kolhapur, Maharashtra, India

Singhal Priyank Moradabad, India

Zerbe Norman Medical University Graz, Graz, Austria

Abbreviations

AAR	After-Action Review
ABS	Anti-lock Braking System
AdaBoost	Adaptive Boosting
AEPS	Average Error Per Second
AI	Artificial Intelligence
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
API	Application Programming Interface
Ar	Anger
ARDI	Actors, Resources, Dynamics, and Interactions
ASV	Asymmetric Shapley Values
AV	Autonomous Vehicle
Bm	Boredom
CAM	Class Activation Map
CBIR	Content Based Image Retrieval
CBR	Case-Based Reasoning
CDSS	Clinical Decision Support System
CEM	Contrastive Explanation Methods
CGLES	Common Ground Learning and Explanation System
Cm	Clam
CNN	Convolutional Neural Network
CNV	Choroidal NeoVascularization
CoD	EMODB + RAVDESS
COVID-19	Coronavirus Disease 2019
CRSE	Cumulative Root Squared Error
CSL	Classical
CT	Computed Tomography
CT Scan	Computed Tomography Scan
DARPA	Defense Advanced Research Projects Agency
DC	Direct Current

DeepLIFT	Deep Learning Important FeaTures
DeepSHAP	Deep Shapley Additive Explanations
DICOM	Digital Imaging and Communications in Medicine (standard)
DL	Deep Learning
DME	Diabetic Macular Edema
DNN	Deep Neural Networks
DR	Diabetic Retinopathy
Dt	Disgust
EC	European Commission
ECU	Electronic Control Unit
EG	Expressive Gradients
EHR	Electronic Health Record
ELI5	Explain Like I'm 5
EU	European Union
FAST	Fourier Amplitude Sensitivity Test
FCNN	Fully Convolutional Neural Networks
FDA	United States Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
FIS	Fuzzy Inference Systems
FMEA	Failure Mode and Effect Analysis
Fr	Fear
GBP	Guided BackPropagation
GLM	Generalized Linear Rule Models
GNN	Graph Neural Network
GNSS	Global Navigation Satellite System
GPIO	General Purpose Input/Output
Grad-CAM	Gradient weighted Class Activation Mapping
GSM	Global System for Mobile communication
GUI	Graphical User Interface
HAII	Human-AI Interaction
HCI	Human-Computer Interaction
HD	High Definition
HOG	Histogram of Oriented Gradients
HTML	HyperText Markup Language
HTTP	Hyper Text Transfer Protocol
Hy	Happy
I/O	Input-Output
ICE	Individual Conditional Expectation
IDNN	Input Delay Neural Network
IDS	Intrusion Detection Systems
IG	Integrated Gradient
IMP	Important
IMU	Inertial Measurement Unit
IoT	Internet of Things
IO-VNBD	Inertial Odometry Vehicle Navigation Benchmark Dataset

ISO	International Organization for Standardization
IT	Information Technology
IVDR	European In-Vitro Devices Regulation
KNN	K-Nearest Neighbors
LAN	Local Area Network
LIME	Local Interpretable Model-Agnostic Explanations
LIMS	Laboratory Information System
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
LT	Latest
MDR	European Medical Devices Regulation
MFNN	Multi Feedforward Neural Network
MIABIS	Minimum Information About Biobank Data Sharing (standard)
MIS	Minimally Invasive Surgery
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
NCD	Non-Communicable Diseases
NIDS	Network-based Intrusion Detection System
Nl	Neutral
NLP	Natural Language Processing
NN	Neural network
NumPy	Numerical Python
OAT	One-Step-At-A-Time
OCT	Optical Coherence Tomography
OEM	Original Equipment Manufacturer
OpenCV	Open-Source Computer Vision Library
PCA	Principal Component Analysis
PD	Partial Dependence
PDP	Partial Dependence Plots
Perm	Permutation
PIMP	Permutation Importance
PRRC	Person Responsible for Regulatory Compliance
QII	Quantitative Input Influence
RBD-FAST	Random Balance Designs-Fourier Amplitude Sensitivity
RBFNN	Radial Basis Function Neural Network
RBIA	Risk-based Internal Auditor
R-CNN	Region-Based Convolutional Neural Networks
RELU	Rectified Linear Unit
RGB	Red Green Blue
RNN	Recurrent Neural Network
ROI	Region of Interest
RT-PCR	Reverse Transcription-Polymerase Chain Reaction
sci-fi	Science fiction
SCS	System Causability Scale

Sd	Sad
SHAP	SHapley Additive exPlanations
SIM	Subscriber Identification Module
SQL	Structured Query Language
Su	Surprise
SUS	System Usability Scale
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TED	Teaching Explanations for Decisions
t-SNE	t-Distributed Stochastic Neighbor Embedding
UAV	Unmanned Aerial Vehicle
UI	User Interface
UI / UX	User Interface/User Experience
WhONet	Wheel Odometry neural Network
WKNN	Weighted K-Nearest Neighbors
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence
YOLO	You Only Look Once

Chapter 1
Black Box Models for eXplainable Artificial Intelligence

Krishna Keerthi Chennam, Swapna Mudrakola, V. Uma Maheswari, Rajanikanth Aluvalu, and K. Gangadhara Rao

Abstract Machine learning algorithms are becoming popular nowadays in cyber security applications like Intrusion Detection Systems (IDS). Most of these models are anticipated as a Black Box. Previously black box was a model where the user cannot see the internal logic. To reach the goal of overwhelming the crucial weakness, the cost may vary. This is related to both ethical and practical problems. Explainable Artificial Intelligence (XAI) is crucial to converting the machine learning algorithms to appreciate the management by accepting the human experts to understand the data evidence. Important role of trust management is to accept the impact of malicious data to identify the intrusions. This chapter addresses the XAI method to appreciate trust management using the decision tree models. Basic decision tree models are used to simulate a human contact to decision making by dividing the options into multiple small options for the IDS area. This chapter aims to implement the arrangement of issues labeled in the various black box methods. This survey helps the researcher to understand the classification of various black box models.

Keywords Black box · Cyber security · Decision trees · Intrusion detection system · Artificial intelligence

K. K. Chennam (✉) · S. Mudrakola
Vasavi College of Engineering, Hyderabad, India
e-mail: krishnakeerthich@gmail.com

S. Mudrakola
Matrusri Engineering College, Hyderabad, India

V. U. Maheswari
KG Reddy College of Engineering, Hyderabad, India

R. Aluvalu · K. G. Rao
CBIT, Hyderabad, India

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_1

1.1 Introduction to Machine Learning

There was a huge increase in artificial intelligence (AI) in a glimpse. Machine learning is a subset of AI. The main importance of machine learning is identifying the structure of data or format suitable data models used by the users. However, Machine learning is related to computer science and varies from former computational methods. Previously, Algorithms were written exclusively programmed instructions for computers to solve problems. Now machine learning (Othman et al. 2018) algorithms are used to educate the computers on data inputs and data statistics, analysis is used to produce output values within a range. Automatically decision is taken based on the sample data with the help of models and inputs. Many technologies are using machine learning (Gilpin et al. 2018) algorithms and get benefited. Facial recognition is one of the technologies which permit social media platforms like Facebook and Instagram's to help the users tag and share friends' photos (Logas et al. 2022). Movies or television shows using optical character recognition technology help to change images to text into movable (Jiang et al. 2022). Self-driving cars also depend on machine learning to map the routes (Saha and De 2022). Machine learning is consistently improving technology, which requires continuously improving methodologies for analyzing may affect the machine learning process (Pazzani et al. 2001). Supervised and unsupervised learning are two basic machine learning methods. Along with these two methods k-nearest neighbor algorithm, decision tree learning methods and deep learning are other important concepts in machine learning.

Firstly, supervised learning purpose is to learn by similar outputs by identifying errors and changing the models depending on the output (Cai et al. 2022). This model also uses the patterns to identify the labeled values and unlabeled data also. Supervised learning algorithms will make sure to identify the images and produce labels to the particular image by seeing the cat image, supervised learning will be able to identify and label it as an animal. Unsupervised learning is to identify the secret patterns in the data and automatically identify the classification of raw data. This is used for transactional data and complex data is more expansive and unrelated to organize properly (Kotenko et al. 2022). Example like unsupervised learning will be able to tag all cat images and group it.

Machine learning is based on statistics with basic knowledge by understanding and supporting machine learning algorithms. Correlation is used to identify the relation among two dependent or independent variables. Regression was used for identifying the relation among dependent and independent variable. When an independent variable is given and needs to identify the dependent variable, the regression statistics used to identify it is called regression enables prediction capabilities. To identify the pattern k-nearest neighbor algorithm is used for regression and classification. Small and positive integer is k value. Example of separating the square and circle shapes into two different classes, this classification is used.

Decision tree is a predictive algorithm based on the models, observations, analysis and gives target data values. This model is created to predict the target based input values. The data attributes identified based on the observation are branches

the conclusion of data target values is nothing but leaves. Deep learning is introduced based on neural networks with multiple layers in artificial neural networks based on hardware. The output is connected to an input to the next layer in the deep learning process. Computer vision and speech recognition have realized significant advances in deep learning approaches (Li et al. 2022). Humans can give biased decisions that lead to negative results, machine learning helps to overcome such issues and give unbiased decisions. Black box (Guo 2020; Perarasi et al. 2020a) systems exploit sophisticated machine learning models to identify separated secure data. Medical status, risk of insurances, eligibility score for credit cards acknowledge using machine learning algorithms construct predictive models and map the features into class in the learning phase (Svenmarck et al. 2018). The learning process is formed by the digital trances that are left after operating daily activities like social media activities, purchases, etc. Huge data may handle human biases and prejudices. Decision models are accomplished by inheriting biases, wrong decisions and illegal activities. Various scientific communities studied the issues of discussing machine learning decision models. Even though illustratable machine learning is the important case and accepted newly considering the situation, many ad-hoc distributed results.

The rest of the chapter is organized as follows. The First section discusses the importance of cyber security in XAI. Next section discusses Deep learning using XAI which follows the Intrusion Detection System (IDS). Section 1.5 is about applications of cyber security in XAI. Section 1.6 discusses the comparison of XAI using black box methods and finally about the conclusion.

1.1.1 Motivation

The unique aim of the chapter is to reach the novelty in research work using machine learning. AI understands different technologies under the same umbrella like machine learning to predict the results. Machine learning ultimately reaches the goal to reach for accurate results with training the model.

1.1.2 Scope of the Paper

Machine learning is one of the best options in career applications for smart systems to handle business attacks. Target is to calculate human intelligence and be able to make decisions more precisely under any situation. AI handles the different technologies that come under the same domain like pattern recognition, big data, machine learning, artificial intelligence and various other technologies. This is the reason AI is having much future scope in many applications.

1.2 Importance of Cyber Security in eXplainable Artificial Intelligence

Industries progressively improved with a better complex cyber security (Pienta et al. 2020) ecosystem depending on various types like users, technology and processes to functional roles. Cyber security is dependent on relations between users and groups, users, organizations and technology, technology and users. From the above trusting peers, cyber security prevents separately to defend against cyber attacks. AI models cite the knowledge from the gathered data. Actually, no human will believe the AI system for the possible and desirable quality of data, difficult methods and accountability, trained AI engineer. AI is trust related software that gives solutions to cyber-attacks. You may ask how to trust the AI models in cyber security, which are developed based on data analysis and predict the solutions from the data. The simple answer for this question is that XAI (Guo 2020; Arrieta et al. 2020) will justify reliability, ability, and trustworthiness. Main challenge for AI is the inability to understand and compare between transition models. A simple example is Autonomous vehicles (Perarasi et al. 2020b). Trustworthy AI should explain its decisions to allow the human expert to understand the underlying data evidence and causal reasoning.

Complex black box models study from machine learning and deep learning parameters. Based on the black boxes models, AI engineers identify direct models to make decisions and identify the behavior of models. Cyber security is liable for attacks and targets the trusted security in critical systems. Therefore XAI from AI plays an important role in developing the solution based AI with interpretability. Interpretability further assures uniformly in decision-making to detect the imbalanced dataset. Interpretability strengthens the powerful solution based AI using highlighting hidden could change the prediction. The decision tree model is developed based on the Intrusion detection system attacks (Svenmarck et al. 2018; Stampar and Fertalj 2015). The intrusion detection system developed fast in study and organization research in exchange for increasing cyber attacks on government and commercial enterprises internationally and action on cost is increased consistently (Lee et al. 2001). The main harmful cyber crimes are from vicious associates, denial of servers, web attacks, and organizations may lose the intellectual property related to vicious attacks in the system. Organizations install various firewalls, software like antivirus and intrusion detection systems against those attacks. Intrusion detection is a crucial role in cyber security, grants to determine vicious network activities previously compromises data connection, availability and opportunity. It is a method to identify security breaches by interrogating models in the data system.

Day-to-day, the digital system is adopted by the world. The network access leads to a lack of security issues that the Internet of Things devices (Lee et al. 2001; Chennam et al. 2022). Intrusion attacks with high possibilities on Internet of Things devices connected to the internet lead to network devices safely from intrusion. An IDS was developed to avoid important data from vicious acts. Important data with network access needs to be permanently protected from all pursuit to consume, expose, alter, disable, steal or gain unauthorized access. Traditional intrusion detection systems,

mainly signature-based, identify only popular attacks and may not identify new attacks. Machine learning is the best approach which is exclusively developed to maintain detection accuracy.

Artificial Intelligence (AI) has helped all the industries with effective results in deploying various applications to monitoring, Decision Making, Solving Complex problems, creative approaches, observation analysis, Language Recognition and Learning. Artificial intelligence has collaborated with additional technology like Machine Learning, Neural networks and Deep Learning. Artificial intelligence is used to compute the programs and prepare the system to behave like a human brain (Uma Maheswari et al. 2021; Deshpande et al. 2020). The AI has excelled in thinking, retrieving and taking decisions sometimes faster than the human brain. AI applications are used in medical Care, Teaching and Learning, Law, Commerce and public Departments etc. The above applications are intended to say that algorithms rule the world by AI, which is inevitable (Swapna et al. 2022).

XAI advantages are mainly concerned with ethics and continuous improvements. XAI required enough trust to handle the AI. For decades various AI models gave biased results or not perfect results which lead to ensuring the safety in AI decisions without any faults. To justify the final decisions taken by the AI required logical reasoning in decision making. AI helps to identify the malware weekly updated and all possibilities of pattern recognition, behavioral attacks of ransomware able to identify before entering into the system. Bots help to clear maximum chunks in internet networks. Stolen login details can create false account details, tampering data; bots can be the correct menace. Handling automatic threats is not possible alone. AI and machine learning will heal to construct the good bots to identify the engine crawlers, bad bots etc. AI starts to identify the data and accepts to provide cybersecurity to understand the strategy consistently.

1.2.1 Importance of Trustworthiness

The importance of Trustworthiness is an essential aspect to measure the safety, performance and reliability. The qualities requirements to say as trustworthy are the system must be accountable, fair, reliable behavior, reasonable and acceptable. The author Stephen Hawking says “AI can spread faster and can be violent if it is not controlled properly”. The AI systems need to be authorized and validated in each design and implementation phase. The AI systems need to be authorized and validated in each phase of the design and implementation. There are different algorithms used to predict the risk of the system, and it occur due to low-quality data training, narrow perception of the problem, technical issue management etc. can lead to unrecoverable loss of people, properties and loss the trust on AI practices. AI applications are used in important applications like facial recognition software, Tagging picture in television media, Health care practices and self-driving car are the high-risk applications, wrong decision may cause life. The author Davinder Kaur has raised some questions

Table 1.1 Questioner table states the importance of Trustworthy in AI

Research questionnaire	Proposed solution
Purpose of proving AI is trustworthy	Decisions taken by the AI system should be ethical practice, robust in nature, Lawful and acceptable
What protocols are used to work AI systems?	We can empower and help to maintain the AI system lawful practices
Why human control involved	AI systems need to collaborate with human intervention and machines in cognitive decision making
Reasons for AI acceptable	AI systems have proven to be trustworthy, fast and usable

to understand the requirements needed to conclude the AI system as worthy (Kaur et al. 2022) (Table 1.1).

The Black Box Model uses AI methods, the results are obtained, but its design will not help to justify the result. The explanations are required to extract the output function. We need to apply some techniques to find the reason to conclude (Zhang et al. 2022). The Post-Hoc Explainable is a reverse engineering process that starts to reach the initial state from the destination. Explainable algorithms like Support Vector Machine (SVM), Multi-Layer Neural Network, Convolution Neural Network and Recurrent Neural Network (Hermansa et al. 2022). XAI uses machine learning techniques to justify the results. The reasonable techniques are explained by simplifying the problem, Feature Connectivity, Local Reasoning, Visible Reasoning and Multi Classifier. The importance of AI is used to make better decisions, explain deep learning, Model Debugging, and build the latest model (Brito et al. 2022) (Fig. 1.1).

Machine Learning (ML) methods Contribution for XAI—The Machine Learning method works for limited data. The ML required defined features to the drive result. The complex problems will simplify and solve phase-wise, network designs are

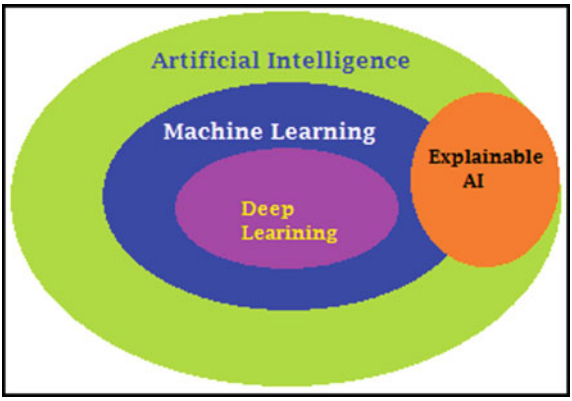


Fig. 1.1 Representation of AI, DL, ML, XAI Association

kept simpler, less trained data and good results are obtained on more and less data size (Lötsch et al. 2022). The ML concepts are classified into three classes: supervised, unsupervised, and Reinforcement Learning. The methods are Artificial Neural Network, SVM, Self Organizing Map, Model-based Reinforcement Learning, clustering, Dimension reduction, Regression, Classification, Transfer Learning, and NLP (Aliramezani et al. 2022).

1.3 Deep Learning (DL) Methods Contribute to XAI

The Deep Learning methods require a huge amount of training data. The feature extractions are undefined at the initial stage Based on the feature's importance, the feature is used for learning. The network training takes more time, based on hidden layers. As no of hidden layers increases, depth analysis is performed (Raza et al. 2022). The DL algorithms are CNN (Conventional Neural Network), MLP (Multi-Layer Perceptron), DNN (Deep Neural Network), and RNN (Recurrent Neural Network). All the methods extract the results using the above Neural Network. The above methods measure the performance using ROC, F1-Score, Accuracy, Recall and Precision metrics (Zhang et al. 2022).

Block Diagram of XAI is shown in Fig. 1.2, Data training is a machine learning process to teach the environment about the possible case studies and the latest and most possible cases, calculus applied to find the break-even point and threshold calculations, etc. Different Machine learning methods are used to extract the results. The Machine learning approaches discussed section I. The recommendation and conclusions are obtained based on the methods of ML Programs selected. The Explainable AI will strength up the decision by explaining the reason for obtained results. (Hoffman et al. 2018) Functionality difference between AI and XAI, ML functions will learn from training data and decisions are made based on learned function. The XAI will train the data, ML process will process the input to obtain the result. The Explainable AI will reason the result and explanation will be verified by the user to test the accuracy behavior. The conclusions and recommendations are assessed by the user (Hermansa et al. 2022) (Table 1.2).

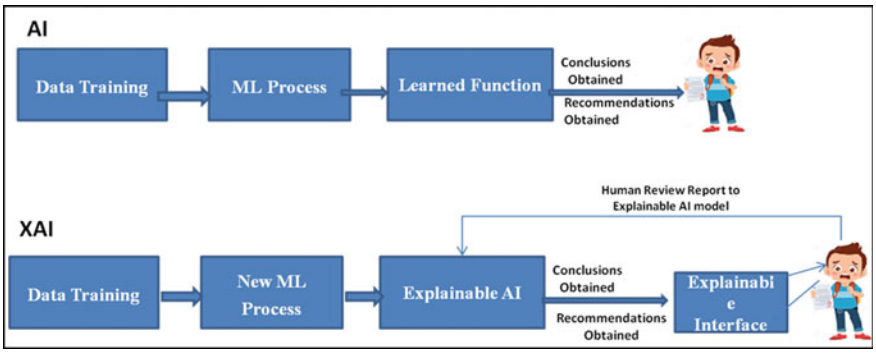


Fig. 1.2 Phases in AI and XAI model

Table 1.2 The details of the Explained AI and Non-XAI for User Perception

S. No.	User perception	AI/ML/DL algorithms	Explainable AI
1	Did the result obtained cause only this?	Unable to understand clearly	Understanding clarity
2	Alternate options for result opted	Unknown reason for not selecting another choice	Known reason for not selecting another choice
3	Success in result obtaining	Unknown for success	Known for success
4	Failure in result obtaining	Unknown for failure	Known for failure
5	Trust the system	Unpredictable	Predictable
6	Error in system	Correct based on the false positive	Reasoning for false positive

1.4 Intrusion Detection System

This section first explains the concept of IDS and then provides the details about the classification of IDS based on its deployment and the detection methodology. An IDS is the combination of two words, "intrusion" and "detection system." Intrusion refers to unauthorized access to the information within a computer or network system to compromise its integrity, confidentiality, or availability. The detection system is a security mechanism for detecting of such illegal activity. So, IDS is a security tool that constantly monitors the host and network traffic to detect any suspicious behavior that violates the security policy and compromises its confidentiality, integrity, and availability. The IDS will generate alerts about detected malicious behavior to the host or network administrators. Figure 1.3 depicts a passive deployment of NIDS, where it is connected to a network switch configured with the port mirroring technology. The task is to mirror all the incoming and outgoing network traffic to NIDS for performing traffic monitoring to detect intrusions. NIDS deploys in between the firewall and the network switch to allow all the traffic to pass through NIDS. IDS have grown

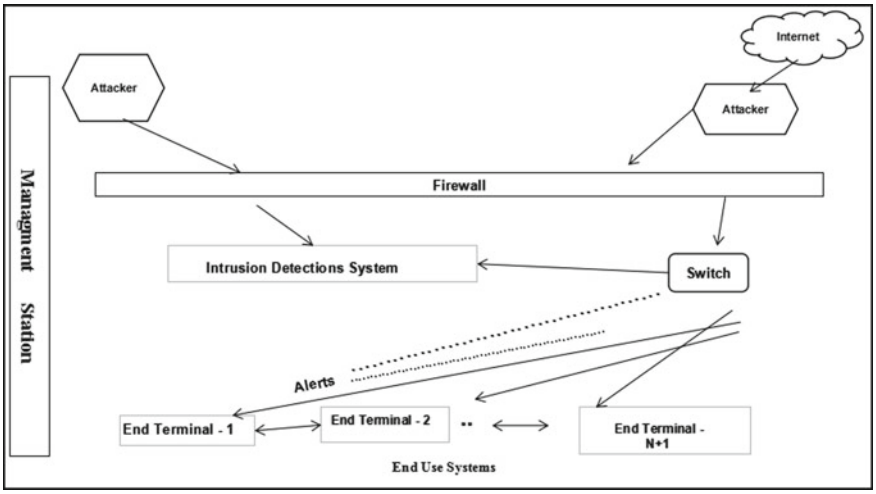


Fig. 1.3 Intrusion detections system network (Chebrolu et al. 2005)

quickly in exploration and industry in light of the expanding digital assaults against state run administrations and business ventures worldwide. The yearly expense of battling digital wrongdoing is constantly expanding (Stampar and Fertalj 2015). The most terrible digital wrongdoings are those brought about by noxious insiders, refusal of administrations, and online assaults. Businesses or organizations can lose their licensed innovation because of these pernicious assaults into the framework. To retaliate against such demonstrations, associations convey a firewall, antivirus programming, and an interruption discovery framework.

Recently detailed predictions, the deep neural network appropriately suitable for these various models are useful and difficult to identify. Autonomous vehicles need various parameters to deal with deep neural networks (Ye et al. 2004). Handling network administrators is difficult if deep neural network models from machine learning are implemented. Deep neural networks are also called a black box models. Decision making process issues are solved with black-box models using trial and error methods till they reaches for feasible solutions. Intrusion detection system implements machine learning methods to improve the accuracy of familiar attacks' analysis and identifies abnormal traffic issues and autonomous vehicle network issues. Machine learning algorithms can identify attacks to interpret the results. The main challenge is to combine the intrusion detection system with deep learning models to ensure security policies against attacks.

Various models proposed for intrusion detection systems like statistical methods (Lazarevic et al. 2003) proposed Markov model (Ye et al. 2004). Neural network (Novikov et al. 2006), fuzzy logic (Toosi and Kahani 2007) discussed. SVMs discussed the huge accuracy in implementing intrusion detection systems (Lahre et al. 2013; Zhang and Shen 2005; Ilgun et al. 1995). Training techniques designed to implement intrusion detection systems, experts considered the rules. Like, discussed

in Rudin (2019), need corporation rules against expertise in analytical models. The disadvantage of such analytical models is the extraction of huge rules, which leads to maximum difficulty in models. One of the critical aspects of a supervised classification model is feature selection. Identifying required features lessens the algorithm computation time. Intrusion detection systems expanse with selections (2017) with various feature. Author (Chebrolu et al. 2005) restricted fundamental features in composing an intrusion detection system is essential for recent detections. (Zaman and Karray 2009) Implements the selection models to construct a lightweight intrusion detection system.

As discussed in Vimalkumar and Radhika (2017), intrusion detection system models with a ratio profits as various techniques selections and two types of techniques like SVMs and endorsed maximum accurate levels denial of service attacks. One of the disadvantages of the model is that it is highly computational with a separate ratio and classification algorithm. Balakrishnan et al. (2014) examined different algorithms like k-means clustering algorithm, Naive Bayes, OneR algorithms for common traffic with accurate results and Denial of Service attacks. A genetic algorithm is implemented to ensure the identification of various models of intrusion with maximum efficiency as discussed by Farrahi and Ahmadzadeh (2015). Appliance different machine learning algorithms identify Denial of Service intrusions and establish the maximum efficiency with multilevel perception (Castelvecchi 2016). Peng et al. (2018) recommended an intrusion detection system planned on a decision tree to boost the accuracy detection. Naive Bayesian and KNN models are performing research to find an accuracy to identify the intrusions with better performance, maximum speed and less false alarm rates like in (Othman et al. 2018; Gilpin et al. 2018).

1.4.1 Classification of Intrusion Detection System

Intrusion detection systems are classified into various types. Firstly the IDS is classified into a network intrusion detection system that helps to identify the packet search, anonymous packets, inbound and outbound packers in firewalls and controls the network traffic. Next classifier is a host intrusion detection system that will monitor the host activities, clients can be ignored in IDS, able to identify the attacks and detected attacks can stop within the network. Another classifier is a protocol-based intrusion detection system used to analyse and monitor communication and monitor HTTP protocols and SQL protocols. Along with all the classifiers the hybrid intrusion detection systems works more efficiently by combining two or more IDS classifiers like host intrusion detection system and network-based intrusion detection system helps to monitor the system, application events and file systems. This hybrid IDS uses for critical server situations (Lin et al. 2022) (Fig. 1.4).

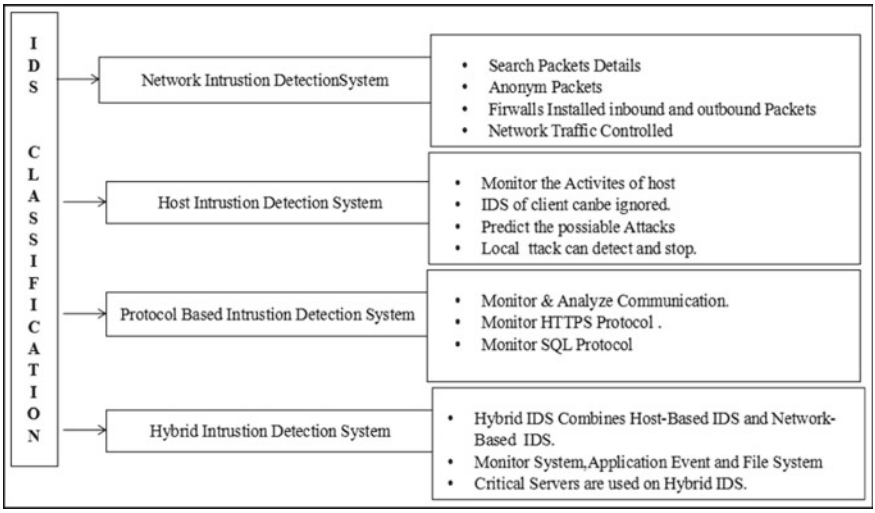


Fig. 1.4 Classification of Intrusion detections system

1.5 Applications of Cyber Security and XAI

It may be described as the method to relax the safety if you want to shield reputation casually, less business or monetary lack of a cluster Cyber security manifestly needs betters for safety with the notion to the business enterprise that regular customers use the network over the internet. Many intelligent tactics and strategies can be cast-off to install in it. The finest huge reality around safeguarding data is no longer a non-prevent process. The business enterprise owner must hold stuff modernized in mandate to keep the danger low (Bonfanti 2022). A business wants to appear a massive harm when they are no longer sincere about the protection in their online display. Nowadays, everyone connects from innovative cybercrime issues to single user issues like attacks, thefts, blackmails, illegal photographs. All are predicted at risk according to the financial supplier of businesses. Giving security to those various fields is vital to understanding general operations like handling information corresponding to credit cards and authorization information. Handling such sensitive data fails is one of the possible cyber attacks. Handling such sensitive data fails is one of the possible cyber attacks. To Gain knowledge on vicious emails, you need to study cyber security. Ransomware is another sort of vicious software program, considered to extract forex with the aid of gadgets or desktops and demand for money. Though after paying money to hacks, it's not guaranteed that the malware is removed from the gadgets or not (Urooj et al. 2022).

Social engineering is a tactic that fighters use to faux you into illuminating sensitive facts. They can importune a monetarist charge or development gets right of entry to your reserved information. Social engineering may be collective with a number of the pressures registered above to fashion you extra in all likelihood to attach on

links, switch malware, or perceive a malicious purpose. Most commercial enterprise operations goals run on the net, exposing their information and assets to diverse cyber threats. Since the information and gadget assets are the pillars upon which the agency operates, it drives the missing maxim that a threat to those people is surely a hazard to the institution itself. A hazard may be everywhere among a minor worm in a code to a complicated cloud hijacking liability. Risk evaluation and estimation of the reconstruction fee assist the agency in living organized and to appear beforehand for capacity losses (Rajanikanth 2021). Thus, understanding and formulating cyber security goals specific to each agency is important in protecting precious information. Cyber security is an exercise formulated for the project of complicated information on the net and on gadgets safeguarding them from attack, destruction, or unauthorized get right of entry to it. Cyber security aims to ensure threat-loose and stable surroundings for retaining the information, community, and gadgets guarded in opposition to cyber terrorizations. Cyber security has become a main challenge over the past 10–12 months within the IT world. All people are dealing with quite a few cybercrime issues in the existing world. As hackers are hacking principal touchy facts from authorities and a few corporation agencies, the people are very much involved as cyber security attacks can result in the whole thing from wholesale fraud to blackmail massive companies. There are many sorts of cyber-crimes rising wherein everybody wishes to be aware of the scams, and they're one of a kind measures and gear which may use for averting the cyber-crimes. Every agency desires to stabilize its private information from getting hacked. Getting hacked isn't always dropping the private information, but dropping the connection with clients within the market. The Internet is the modern day's fastest developing infrastructure. In modern-day technical surroundings, many new technologies are converting humanity. But because of technology, we're not able to guard our non-public facts in a green way, so cyber-crimes are significantly growing on day by day basis. The majority of the transactions in each business and private sector are achieved through an online transaction, so it's crucial to have knowledge that requires an excessive best of safety, retaining higher transparency to everybody and having more secure transactions. So cyber security is the present-day issue. Advanced technology like cloud services, mobiles, E-commerce, net banking and plenty of extra require excessive requirements and a more secure system of safety. All the gear and technology concerned for those transactions keep the maximum touchy and crucial consumer facts. So presenting vital safety to them may be very crucial. Cyber security and safeguarding touchy information and infrastructures are crucial to each nation's pinnacle precedence safety (Bendovschi and Ionescu 2015).

Organizations may see the loss if they are not transparent in security while stepping online. Nowadays, all know the progressive cyber defense agendas. Cyber security may lead to other results from natural theft, loss of photographs, and blackmail attempts at different levels. It all depends on the business levels, monetary services, infirmaries etc. Trusting and providing security in operations is compulsory. Cyber threat investigators train the users on position and recent susceptibilities also. Various kinds of cyber security phishing scam emails from different sources. The main aim of the data is to maintain the security and privacy of credit card details for user

logins which is the highest attack on the organization. Ransomware is one malicious software to blackmail the organization by leaking data or blocking the systems until the organization pays the demanded price. Malware is another kind of software to receive prohibited policy after using it to the system (Gazet 2010).

Enlightening sensitive data is another gimmick in social websites where malicious users can miss-use. The main aim of the huge business operations is to work on the online data and resources, leading to different kinds of cyber threats. The data and resources are an important support for any organization. If any organization lacks security for these two data and resources, it's a big threat from a small bug to difficult data seizing. The reconstruction cost of the organization is very high and may lose the customers due to a lack of trust in the organization. The objective of cyber security for any organization is securing the data and valuable customer information. Cyber security is one of the best practices for organizations to secure difficult online data and provide security from various attacks, destroying the data or checking the authorization or authentication for genuine users to access the data. The main achievement of cyber security is to make sure the surroundings are harmless surroundings for storing data, networks, and resources in contrast to cyber attacks. Among IT industry cyberattacks is one of the main issues from the past 10 years. Presently many common people are facing the cyber attacks problem with cybercrime. Hackers are able to get the sensitive data from the common people and from small organizations to huge organizations also. Different types of cyber crimes are increasing day to day life which gives alert to everyone to be understanding about the cybercrime now to avoid cyber crimes.

Every customer and organization tries to secure the data from the cyber-attacks discussed by Bendovschi in 2015. As infrastructure is emerging rapidly nowadays, the technologies are increasing and changing today. Due to this, we are losing hold on personal data and secure data, which regularly leads to increased cybercrimes (Hussain et al. 2021).

Non-commercial and commercial transactions mainly are happening through the internet. It's mandatory to ensure the experts provide high security and maintain transparency for more safety. New methods like cloud service providers, smartphones, E-commerce, net banking or mobile banking, telecom services need huge security in the implementation. Various tools and techniques tangled the critical sensitive client data.

XAI object detection software applications are designed to identify the objects on the image, video or online live streaming. The computer vision techniques are used to find objects, count similar objects, identify the object, identify the location, and read an image. The algorithms used are R-CNN, HOG, R-FCN, Single Short Detector, and YOLO (you only look once) (Kose et al. 2019). The Deep Neural Network is a 'black box' in behavior. The CNN algorithm will train the neural network; the reasoning for the output is evaluated using techniques Predicate Logic for self-monitoring methods (Floreano and Wood 2015).

Explaining Autonomous Drones application is specially designed to provide carrier service on mountains or hills. The uneven heights of the land and rocks have an issue in flight traffic plan, drone root plan, and physical location features. The

AI Drones are designed with Common Ground Learning and Explanation System (CGLES) with Explainable intelligence to what is comparable, why it is a drone, where explanation. Simultaneously performance prediction with explanations, Pre-Detection and Post-Detection Explanation used to determine the performance of the CGLE system. Pre-Detection is used to define the plan and execution performance, and Post-Detection is used to determine and find the betterment (Tseremoglou et al. 2022).

Explaining forecasting and packing for Air Cargo loading application will predict and decide to air cargo service to help accept or reject the booking based on the estimation of passenger aircraft belly. The cargos services are unpredictable due to the last hrs cargo service details being released. The training data and historical studies will help to predict the certainty. The author proposed a novel framework to provide process consideration based on the balance of aircraft capacity and dimensions of the ship (Han and Liu 2022).

Explaining Structure Health Monitoring through AI and ML applications are used to detect the extract the features and predict the patterns. Machine Learning is used for transparent processing, and some limiting features will undergo black-box execution. The XAI preprocessing problem uses applications in ML algorithms. An explanation is retrieved, interpretations and finally, build XAI model. The first ML algorithms are SHM systems. In this phase, it is required to remove the noise in the data and improper information. Supervised, unsupervised and reinforcement are the approaches. Four algorithms and ML algorithms are used to understand the problem, confirm the method, apply the method, explain the output, and interpret (Chou et al. 2022; Swapna et al. 2022; Kanaparthi and Swapna 2022).

Explainable AI for Deep Learning Models applications is specially used in applications for big data as it is difficult to handle for a complex task (Anders et al. 2021). Large scales of application are Speech Recognition, Text Analysis, Problem Solving and Image Classification. The above applications run very successfully for ML and AI concepts. But the decision extracted is not transparent. The author took an image classification problem in identifying the object and Interpreted the reason for classification. The process has been decomposed into classification steps. The input(x) was applied to a black box AI system and was predicted as a Rooster(x). The prediction has an AI explanation (we will verify the predictions, Identify the issues and differences, understand the problem, ensure the problem) (Han and Liu 2022).

Explaining AI for the Breast Cancer Detection Case based on reasoning applications designed in old patients or new patients, the general classification is the black-box approach. The measures are considered based on the terms of values and quality-wise. It has an automatic interpretation. The Databases are designed and retrieved using queries. The automatic case retrieval system will extract similar cases and retrieve preprocess queries stored in temporary memory. Automatic classification is used to classify the database, the reasoning based on Quantitative and Qualitative. The visual reasoning with query's and class classifications. The three algorithms are tested KNN, WKNN and RBIA to find the clinical validation to improve the study of CBR, and the third is to size the tumor. Finally, the knowledge discovered from the

medical dataset analyzer not only from CBR (Zhang et al. 2022; Swapna and Hegde 2021).

Various methods in XAI to predict the outputs are:

i. SHAP (Fidel et al. 2020)
ii. LIME (Visani et al. 2020)
iii. SHAPASH (Ghosh and Sanyal 2021)
iv. EXPLAINER DASHBOARD
v. DALEX (Baniecki et al. 2020)
vi. EXPLAINABLE BOOSTING MACHINES (Naser 2021).

SHAP (Shapley additive explanations) is a python tool used to visualize the model's output using Machine Learning implementation and helps visualize the output with more explanation. This algorithm will help explain the reason for the output of the prediction model. In the SHAP feature, importance is the first step to finding the important attribute from another attribute, and they are evaluated using slandered deviation and mean. It will help to remove the impurity in the decision. The SHAP will plot different plots like the SHAP Summary Plot used to combine the features and plot data points. SHAP Dependence plot will help plot the marginal effect of features, SHAP Force used for error analysis, explanation for findings for prediction (Kuzlu et al. 2020).

LIME (Local Interpretable Model-Agnostic Explanations) is another python package used to predict the classification and regression of the data. All the sample data will be extracted from the feature, and observe the results. The explainer will help to predict the result for each output. The linear regression is regularized. The difference between output and actual will use r2. The difference between actual and predicted output gave by linear regression function. It will explain the black-box model machine learning model. Local Interpretation will help to calculate trustiness and it also prove the untruth the model and visual explanation (Främling et al. 2021).

SHAPASH is a python library to Interpretable the Machine Learning results. It uses to build the Web Application to interpret the decision of the data scientist, users or customers, Stake holders of the business and Evaluators of the system. It provides the visualization explanation, to understand by the common users. The shapash has five step processes.

Step 1: Regression model is Build.
Step 2: SmartExplainer of Shapash is compiled and displayed on the webapp.
Step 3: Smart Predictor is predicted from the SmartExplainer.
Step 4: SmartPredictor can be saved in pickle File.
Step 5: Finally predictions can be made (Ghosh and Sanyal 2021).

EXPLAINER DASHBOARD is an interactive dashboard works for the machine learning models. It helps to analyze the reasons for the predictors and explanation on the working of the model. The Explainable Dashboard will help to build the unclouded machine learning model and Explainable. Classifier Dashboard has features like Classification Stats, Individual Predictions, and Own Condition prediction, Feature Dependence, Feature Interaction and Decision Tree. This can be

supported by Colab Programming. The dashboards are different types like single tab dashboard, Multi-TabDashboard, Documentation Dashboard etc. are the sample types of dashboards.

DALEX is a model Agnostic Language for Exploration and eXplanation is a Machine Learning analysis model helps to learn the behavior of the Model Predictor in Classification process and Applying Regression methods. This approach will also help to build the relation between the dependent variable to predict the outputs. It is also an interactive model for exploration of predictions. Explanatory Model Analysis. This model works with different levels of explanations like predict Level Explanation, Model Level Explanation, Save and Loading Explanation. Plot the graphs for user visualizations (Baniecki et al. 2020).

EXPLAINABLE BOOSTING MACHINES (EBM) are the cyclic Gradient boosting Adaptive model, tree-based classification and interactive model. This method for predication is said to be black box model, which said to be more accurate. Limitation is EBM will take long time to train the model but very fast at prediction. The process for EBM is train the model for classifiers, Visualization is explained in terms of local and Global. The Specific attribute analysis need to perform. All the Attribute Mean Absolute Score need to calculate (Naser 2021) (Table 1.3).

Table 1.3 Comparison of XAI techniques

References	Methods	Intrinsic	Post-hoc	Global logic	Local logic	Specific model	Agnostic model
Roth et al. (2021)	Decision trees	✓		✓		✓	
Das and Rad (2020)	Rule lists	✓		✓		✓	
Dieber and Kirrane (2020)	Lime		✓		✓		✓
Dhanorkar et al. (2021)	Sharply explanations		✓		✓		✓
Schlegel et al. (2019)	Saliency maps		✓		✓		✓
Fouladgar and Främling (2020)	Activation maximization		✓	✓			✓
Kłosok and Chlebus (2020)	Surrogate models		✓	✓			✓

(continued)

Table 1.3 (continued)

References	Methods	Intrinsic	Post-hoc	Global logic	Local logic	Specific model	Agnostic model
Ryo et al. (2021)	Partial dependence plot		✓	✓	✓		✓
Barbado et al. (2022)	Rule extraction		✓	✓	✓		✓
Adadi and Berrada (2018)	Model distillation		✓	✓			✓
Baur (2018)	Sensitive analysis		✓	✓	✓		✓
Keane et al. (2021)	Counterfactual explanations		✓		✓		✓
Heide et al. (2021)	Prototype and criticism		✓	✓	✓		✓
Anders et al. (2021)	Layer wise relevance program		✓	✓	✓		✓

1.6 Comparison of XAI Using Black Box Methods

The depth analysis of classifications of black-box models discusses in this section using XAI. The reverse engineering approach is used in the black box also familiar with black box predictors observing the input and output of the black box. Assigning decisions to black boxes is difficult to interpret may have differences and trust problems. Former datasets and training models handling human decisions could depend on (Pedreshi et al. 2008). These methods are acutely covered up inside the classified trainer. Improving the black-box model is a high risk, as discussed in Pasquale (2015), and carried by secret algorithms, legal protections, and differences consciously or unconsciously may lead to invisible or impossible. Automated differences are not new and not compulsorily due to the black box (Kuppa and Le-Khac 2020). Comparison methods for black-box models using XAI are discussed in the below table. The data types are used to analyze the black box models using XAI. General is the explanatory method to reach every black-box method. Random is the type indicated to randomly select any random perturbation of a data set. Code indicates the source code is available. The tabular method data set analyses the comparisons for black-box models using XAI (Table 1.4).

Table 1.4 Comparison methods for black box models using XAI

Reference	Explanator model	Black box types	Dataset technique
Fidel et al. (2020)	Decision tree	Neural network	General, dataset
Krishnan et al. (1999)	Decision tree	Neural network	General, dataset
Kuppa and Le-Khac (2020)	Decision tree	Neural network	General, random
Zhang and Shen (2005)	Decision tree	Neural network	General, random
Chipman et al. (1998)	Decision tree	Tree ensemble	Dataset
Peng et al. (2018)	Decision tree	Tree ensemble	General, random, dataset
Roth et al. (2021)	Decision tree	Tree ensemble	General, random
Hara and Hayashi (2016)	Decision tree	Tree ensemble	Random
Farrahi and Ahmadzadeh (2015)	Decision tree	Tree ensemble	Random
Främling et al. (2021)	Decision tree	Tree ensemble	General, random
Pasquale (2015)	Decision tree	Tree ensemble	Dataset
Adadi and Berrada (2018)	Decision rules	Neural network	Random
Adadi and Berrada (2018)	Decision rules	Neural network	General, random
Farrahi and Ahmadzadeh (2015)	Decision rules	Neural network	General, random, code, dataset
Ryo et al. (2021)	Decision rules	Neural network	Random, dataset
Chebrolu et al. (2005)	Decision rules	SVMs	Dataset
Balakrishnan et al. (2014)	Decision rules	SVMs	Dataset
Vimalkumar and Radhika (2017)	Features importance	AGNostic black box	General, code, dataset
Urooj et al. (2022)	Features importance	AGNostic black box	General, random, code, dataset
Rudin (2019)	Decision tree	AGNostic black box	General, dataset
Lahre et al. (2013)	Features importance	AGNostic black box	General, random, code, dataset

(continued)

Table 1.4 (continued)

Reference	Explanator model	Black box types	Dataset technique
Heide et al. (2021)	Features importance	AGNostic black box	General, random, code, dataset
Brito et al. (2022)	Features importance	Tree ensemble	Dataset
Barbado et al. (2022)	Features importance	SVMs	Code, dataset

1.7 Conclusion

AI takes final decisions with the help of difficult model analysis to calculate potentially secret patterns and low signals using huge data sets. Approaching the real-time application for trusting the AI-related solution. Analyzing and considering the AI-related solutions is required for trusting the applications in real-time. Cyber security systems are important applications sensitive to systems that are at risk in vicious attacks. Accordingly, decision tree algorithms for vicious nodes describe by handling the available datasets. Performance is calculated based on major tasks in the dataset like identifying ranks, decision tree extraction, and state of the art algorithms correlation. All the methods do not have the same level of improvement as the vicious node predictions. The various methods in XAI and the network traffic model are calculated with a double-time window as the important predictors in the decision tree and related to deep root node algorithms. The next second-highest rank is the feature-based network service for the personal TCP connection. This book chapter addresses problems like black-box model types, mapping the cyber security in XAI, and the role and importance of IDS in XAI. Black box working process and explaining decisions even without understanding the depth of opaque decision systems work regularly. Various approaches introduced in black boxes and a few scientific queries are still unable to answer. Research exercises neglected regular reading formalism by defining, describing, and identifying various types. XAI research is dependent on the application domains and affects the various huge applications. The evidence is the main drawback of formulation and unambiguous definitions. The work is affected by challenges with open problems in XAI. Finally, the further intriguing point is that clarifications are significant alone and indicators may be advanced straightforwardly from clarifications. AI is a powerful device that can be utilized in numerous areas of data security. There exist some vigorous enemies of phishing calculations and organization interruption identification frameworks. AI can be effectively utilized to create validation frameworks, assess the convention execution, survey the security of human collaboration verifications, brilliant meter information profiling, etc. Even though AI protects different frameworks, the AI classifiers are defenseless against vindictive assaults. Some work has been coordinated to work on the adequacy of XAI calculations and safeguarding them from different assaults.

References

Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S.A.: Applications of artificial intelligence in transport: an overview. Sustainability **11**(1), 189 (2019)

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Aliramezani, M., Koch, C.R., Shahbakhti, M.: Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: a review and future directions. Prog. Energy Combust. Sci. **88**, 100967 (2022)

Anders, C.J., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. arXiv preprint arXiv:2106.13200 (2021)

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)

Aseen, I.S., Kumar, C.A.: Intrusion detection model using fusion of chi-square feature selection and multi class SVM. J. King Saud Univ.-Comput. Inf. Sci. **29**(4), 462–472 (2017)

Balakrishnan, S., Venkatalakshmi, K., Arputharaj, K.: Intrusion detection system using feature selection and classification technique. Int. J. Comput. Sci. Appl. **3**(4), 145–151 (2014)

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: responsible machine learning with interactive explainability and fairness in Python. arXiv preprint arXiv:2012.14406 (2020)

Barbado, A., Corcho, Ó., Benjamins, R.: Rule extraction in unsupervised anomaly detection for model explainability: application to OneClass SVM. Expert Syst. Appl. **189**, 116100 (2022)

Baur, T.: Cooperative and transparent machine learning for the context-sensitive analysis of social interactions (2018)

Bendovschi, A.C., Ionescu, B.Ş.: The gap between cloud computing technology and the audit and information security. Audit Financ. **13**(125) (2015)

Bonfanti, M.E.: Artificial intelligence and the offence-defence balance in cyber security. In: Cyber Security: Socio-Technological Uncertainty and Political Fragmentation, pp. 64–79. Routledge, London (2022)

Brito, L.C., Susto, G.A., Brito, J.N., Duarte, M.A.: An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mech. Syst. Signal Process. **163**, 108105 (2022)

Cai, D., Wang, W., Li, M.: Incorporating visual information in audio based self-supervised speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2022)

Castelvecchi, D.: Can we open the black box of AI? Nature **538**(7623), 20 (2016)

Chebrolu, S., Abraham, A., Omas, J.P.: Feature deduction and ensemble design of intrusion detection systems. Comput. Secur. **24**(4), 295–307 (2005)

Chennam, K.K., Uma Maheshwari, V., Aluvalu, R.: Maintaining IoT healthcare records using cloud storage. In: IoT and IoE Driven Smart Cities, pp. 215–233. Springer, Cham (2022)

Chipman, H.A., George, E.I., McCulloh, R.E.: Making sense of a forest of trees. In: Weisberg, S. (ed.) Proceedings of the 30th Symposium on the Interface, pp. 84–92. Interface Foundation of North America, Fairfax Station, VA (1998)

Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J.: Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf. Fusion **81**, 59–83 (2022)

Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv preprint arXiv:2006.11371 (2020)

Deshpande, N.M., Gite, S.S., Aluvalu, R.: A brief bibliometric survey of leukemia detection by machine learning and deep learning approaches. Lib. Philo. Pract. 4569 (2020)

Dhanorkar, S., Wolf, C.T., Qian, K., Xu, A., Popa, L., Li, Y.: Who needs to know what, when?: broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. In: Designing Interactive Systems Conference 2021, pp. 1591–1602 (2021)
Dieber, J., Kirrane, S.: Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093 (2020)
Farrahi, S.V., Ahmadzadeh, M.: KCMC: a hybrid learning approach for network intrusion detection using k-means clustering and multiple classifiers. Int. J. Comput. Appl. **124**(9) (2015)
Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: detecting adversarial examples using SHAP signatures. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
Floreano, D., Wood, R.J.: Science, technology and the future of small autonomous drones. Nature **521**(7553), 460–466 (2015)
Fouladgar, N., Främling, K.: XAI-PT: a brief review of explainable artificial intelligence from practice to theory. arXiv preprint arXiv:2012.09636 (2020)
Främling, K., Westberg, M., Jullum, M., Madhikermi, M., Malhi, A.: Comparison of contextual importance and utility with LIME and Shapley values. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 39–54. Springer, Cham (2021)
Gazet, A.: Comparative analysis of various ransomware virii. J. Comput. Virol. **6**(1), 77–90 (2010)
Ghosh, I., Sanyal, M.K.: Introspecting predictability of market fear in Indian context during COVID-19 pandemic: an integrated approach of applied predictive modelling and explainable AI. Int. J. Inf. Manag. Data Insights **1**(2), 100039 (2021)
Gilpin, H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th International Conference on Data Science and advanced Analytics (DSAA), pp. 80–89. IEEE, Turin, Italy (2018)
Guo, W.: Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun. Mag. **58**(6), 39–45 (2020)
Han, H., Liu, X.: The challenges of explainable AI in biomedical data science. BMC Bioinform. **22**(12), 1–3 (2022)
Hara, S., Hayashi, K.: Making tree ensembles interpretable. arXiv preprint arXiv:1606.05390 (2016)
Heide, N.F., Müller, E., Petereit, J., Heizmann, M.: X 3 SEG: model-agnostic explanations for the semantic segmentation of 3D point clouds with prototypes and criticism. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 3687–3691. IEEE (2021)
Hermansa, M., Kozielski, M., Michalak, M., Szczyrba, K., Wróbel, Ł, Sikora, M.: Sensor based predictive maintenance with reduction of false alarms—a case study in heavy industry. Sensors **22**(1), 226 (2022)
Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
Hussain, F., Hussain, R., Hossain, E.: Explainable artificial intelligence (XAI): an engineering perspective. arXiv preprint arXiv:2101.03613 (2021)
Ilgun, K., Kemmerer, R.A., Porras, P.A.: State transition analysis: a rule-based intrusion detection approach. IEEE Trans. Softw. Eng. **21**(3), 181–199 (1995). In: Proceedings of the IEEE Symposium on Security and Privacy (1999)
Jiang, R., Wang, L., Tsai, S.B.: An empirical study on digital media technology in film and television animation design. Math. Probl. Eng. **2022** (2022)
Kanaparthi, S.H., Swapna, M.: A statistical review on Covid-19 pandemic and outbreak. Lecture Notes in Networks and Systems vol. 301, pp. 124–135 (2022)
Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Comput. Surv. (CSUR) **55**(2), 1–38 (2022)
Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B.: If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual XAI techniques. arXiv preprint arXiv:2103.01035 (2021)

Klesel, P.H.M., Wittmann, H.F.: Explain it to me and I will use it: a proposal on the impact of explainable AI

Kłosok, M., Chlebus, M.: Towards Better Understanding of Complex Machine Learning Models Using Explainable Artificial Intelligence (XAI): Case of Credit Scoring Modelling. University of Warsaw, Faculty of Economic Sciences, Warsaw (2020)

Kose, N., Kopuklu, O., Unnervik, A., Rigoll, G.: Real-time driver state monitoring using a CNN based spatio-temporal approach. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 3236–3242. IEEE (2019)

Kotenko, I., Izrailov, K., Buinevich, M.: Static analysis of information systems for IoT cyber security: a survey of machine learning approaches. Sensors **22**(4), 1335 (2022)

Krishnan, R., Sivakumar, G., Bhattacharya, P.: Extracting decision trees from trained neural networks. Pattern Recogn. **32**, 12 (1999)

Kuppa, A., Le-Khac, N.A.: Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

Kuzlu, M., Cali, U., Sharma, V., Güler, Ö.: Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access **8**, 187814–187823 (2020)

Lahre, M.K., Dhar, M.T., Suresh, D., Kashyap, K., Agrawal, P.: Analyze different approaches for ids using KDD 99 data set. Int. J. Recent Innov. Trends Comput. Commun. **1**(8), 645–651 (2013)

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the SIAM International Conference on Data Mining, pp. 25–36. SIAM, San Francisco, CA, USA (2003)

Lee, W., Stolfo, S.J., Chan, P.K., et al.: Real time data mining based intrusion detection. In: Proceedings of the DARPA Information Survivability Conference and Exposition II. DISCEX'01, pp. 89–100. IEEE, Anaheim, CA, USA (2001)

Li, J., Chen, J., Bai, H., Wang, H., Hao, S., Ding, Y., et al.: An overview of organs-on-chips based on deep learning. Research **2022** (2022)

Lin, I.C., Chang, C.C., Peng, C.H.: An anomaly-based IDS framework using centroid-based classification. Symmetry **14**(1), 105 (2022)

Logas, J., Schlesinger, A., Li, Z., Das, S.: Image DePO: towards gradual decentralization of online social networks using decentralized privacy overlays. In: Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), pp. 1–28 (2022)

Lötsch, J., Kringel, D., Ultsch, A.: Explainable artificial intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. BioMedInformatics **2**(1), 1–17 (2022)

Naser, M.Z.: An engineer's guide to explainable artificial intelligence and interpretable machine learning: navigating causality, forced goodness, and the false perception of inference. Autom. Constr. **129**, 103821 (2021)

Novikov, D., Yampolskiy, R.V., Reznik, L.: Anomaly detection based intrusion detection. In: Proceedings of the International Conference on Information Technology: New Generations (ITNG'06), pp. 420–425. IEEE, Las Vegas, NV, USA (2006)

Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T., Al-Hashida, A.Y.: Intrusion detection model using machine learning algorithm on big data environment. J. Big Data **5**(1), 34 (2018)

Pasquale, F.: The Black Box Society: The Secret Algorithms that Control Money and Information. Harvard University Press (2015)

Pazzani, M.J., Mani, S., Shankle, W.R., et al.: Acceptance of rules generated by machine learning among medical experts. Methods Inf. Med. **40**(5), 380–385 (2001)

Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568. ACM (2008)

Peng, K., Leung, V., Zheng, L., Wang, S., Huang, C., Lin, T.: Intrusion detection system based on decision tree over big data in fog environment. Wirel. Commun. Mob. Comput. **2018**, Article ID 4680867, 10 pages (2018)

Perarasi, T., Vidhya, S., Leeban Moses, M., Ramya, P.: Malicious vehicles identifying and trust management algorithm for enhance the security in 5G-VANET. In: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore (2020a)

Perarasi, T., Vidhya, S., Leeban Moses, M., Ramya, P.: Malicious vehicles identifying and trust management algorithm for enhance the security in 5G-VANET. In: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India (2020b)

Pienta, D., Tams, S., Atcher, J.: Can trust be trusted in cybersecurity? In: Proceedings of the 53rd Hawaii International Conference on System Sciences, Maui, HI, USA (2020)

Rajanikanth, A., et al.: Data security in cloud computing using ABE-based access control. In: Architectural Wireless Networks Solutions and Security Issues, pp. 47–61. Springer, Singapore (2021)

Raza, A., Tran, K.P., Koehl, L., Li, S.: Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. Knowl.-Based Syst. **236**, 107763 (2022)

Roth, A.M., Liang, J., Manocha, D.: XAI-N: sensor-based robot navigation using expert policies and decision trees. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2053–2060. IEEE (2021)

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F.: Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography **44**(2), 199–205 (2021)

Saha, D., De, S.: Practical self-driving cars: survey of the state-of-the-art (2022)

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of XAI methods on time series. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4197–4201. IEEE (2019)

Stampar, M., Fertalj, K.: Artificial intelligence in network intrusion detection. In: Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1318–1323. IEEE, Opatija, Croatia (2015)

Svenmarck, P., Luotsinen, L., Nilsson, M., Schubert, J.: Possibilities and challenges for artificial intelligence in military applications. In: Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting, Bordeaux, France (2018)

Swapna, M., Viswanadhula, U.M., Aluvalu, R., Vardharajan, V., Kotecha, K.: Bio-signals in medical applications and challenges using artificial intelligence. J. Sens. Actuator Netw. **11**(1), 17 (2022)

Swapna, M., Hegde, N.: A multifarious diagnosis of breast cancer using mammogram images—systematic review. In: IOP Conference Series: Materials Science and Engineering, vol. 1042, no. 1, p. 012012. IOP Publishing (2021)

Swapna, M., Uma Maheswari, V., Aluvalu, R., Vardharajan, V., Kotecha, K.: Bio-signals in medical applications and challenges using artificial intelligence. J. Sens. Actuator Netw. **11**(1), 17 (2022)

Toosi, A.N., Kahani, M.: A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers. Comput. Commun. **30**(10), 2201–2212 (2007)

Tseremoglou, I., Bombelli, A., Santos, B.F.: A combined forecasting and packing model for air cargo loading: a risk-averse framework. Transp. Res. Part E: Logist. Transp. Rev **158**, 102579 (2022)

Uma Maheswari, V., Aluvalu, R., Chennam, K.K.: Application of machine learning algorithms for facial expression analysis. Mach. Learn. Sustain. Dev. **9**, 77 (2021)

Urooj, U., Al-rimy, B.A.S., Zainal, A., Ghaleb, F.A., Rassam, M.A.: Ransomware detection using the dynamic analysis and machine learning: a survey and research directions. Appl. Sci. **12**(1), 172 (2022)

Vimalkumar, K., Radhika, N.: A big data framework for intrusion detection in smart grids using Apache spark. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 198–204. IEEE, Udupi, India (2017)

Visani, G., Bagli, E., Chesani, F.:. OptiLIME: optimized LIME explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 (2020)
Ye, N., Zhang, Y., Borror, C.M.: Robustness of the Markov-chain model for cyber-attack detection. IEEE Trans. Reliab. **53**(1), 116–123 (2004)
Zaman, S., Karray, F.: Lightweight ids based on features selection and ids classification scheme. In: Proceedings of the International Conference on Computational Science and Engineering, pp. 365–370. IEEE, Vancouver, BC, Canada (2009)
Zhang, Z., Shen, H.: Application of online-training SVMS for real-time intrusion detection with different considerations. Comput. Commun. **28**(12), 1428–1442 (2005)
Zhang, Y., Weng, Y., Lund, J.: Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics **12**(2), 237 (2022)

Chapter 2
Fundamental Fallacies in Definitions of Explainable AI: Explainable to Whom and Why?

D. O. Chergykalo and D. A. Klyushin

Abstract There are many articles that show a discrepancy between the various motives for the construction of XAI (Explanatory Artificial Intelligence), which is not surprising, since this area began to be actively centralized and actively developed only 6 years ago. But the strange thing is that the motives not only do not converge but may contradict each other. This indicates that there are fundamental errors in the very construction of different XAI concepts. These errors create not only contradictions between different visions of XAI, but also common to many concepts of error. The main one is the absence or incorrect answer to the question "For whom exactly should AI be explained?". Turning to human psychology and social processes that are accompanied by the exchange of explanations, we will try to consider what benefits the explanation brings to people and groups. Correcting fundamental errors in the construction of XAI concepts, we show that neural networks are no less explanatory AI than linear models and decision trees. Moreover, we will show what the neural network approach can do so that the explanation will not need to be exchanged for the quality of AI algorithms, and that they can even improve them.

Keywords Explainable artificial intelligence · Biomorphic artificial intelligence · Black-box models · Mental models

2.1 Introduction

2.1.1 A Short History of Explainable AI

The question of the explainability of AI did not arise due to the development of interest in AI bur rather the explanation of how the brain and psyche guided the development of AI. The earliest can be considered connectivism. Various psychologists have tried

D. O. Chergykalo · D. A. Klyushin (✉)
Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
e-mail: dmytroklyushin@knu.ua

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_2

to make psychology a more objective and constructive science, and therefore turned to neurobiology to link their vision of the psyche with the processes of the brain. An example is Sigmund Freud's "Project of Scientific Psychology", compiled in 1895 (Bob 2015). Obtaining information in this case is seen as a complication of the network, and learning is seen as the effective transfer of relationships in information to relationships in the network. And the vision of the brain as a multilevel distributed system was started in 1869 by neurologist John Hulingson Jackson (Greenblatt 1999).

These approaches were quite useful, based on them Friedrich Hayek built a model of learning Hebb's synapses (Hayek 1952) and hypothesized that the brain can be represented as a system of Hebb's networks (Hebb 1949). After that, this theory was developed and after some time, when computers became available to scientists, the first neural network model was programmed, that is the perceptron.

The very theory of networks as systems that can effectively adapt to external situations has proved its worth quite quickly. Pioneer of AI and cybernetics Joseph Licklider did not consider existing AI systems to be human-replacing, because as a psychologist he decided to improve the human–computer interface using mental models: ideas, strategies, ways of understanding that structure the existing experience (Licklider 1960). As early as 1962, in his work "Intergalactic Computer Network" (Licklider 1963), he described almost all the features of the modern Internet, and helped the Defense Advanced Research Projects Agency (DARPA) to create a prototype of the Internet: ARPANET. After that, improving the interaction between people and programs has finally become a long-term prospect for DARPA.

As Licklider identified, AIs at the time, that is, neural networks, could not replace some human functions. And when other researchers published articles explaining why this is so, investment in AI were frozen altogether. Researchers have tried to return to earlier ideas about the mind, which were started by Aristotle in ancient Greece. They began to build AI systems in such a way that they effectively used some rules of inference or symbolic manipulation, thus modeling some human dynamics within a single mental model. Although such systems were more understandable to end users, it quickly became clear that they were very fragile and could not evolve.

With the increase in the capacity of computers, neural networks began to develop again and gradually automate existing processes. But the expansion of the use of AI is gradually beginning to affect critical areas for people and humanity. At the same time, more advanced AI architectures are emerging, and it is becoming clear that rule-based approaches in mental models and neural networks do not contradict each other and can be effectively combined This was well noticed in DeepMind (Graves et al. 2014), and based on these views built their AI, which won other programs in various board games: AlphaGo, AplhaGo Zero, AlphaZero, etc.

DARPA, seeing changes in opportunities, in 2015 began to gather scientists to discuss and develope its future program "Explainable Artificial Intelligence". Using existing ideas about the effective interaction of man and computer and the consultation of scientists, they determined that the construction of AI requires a combination of three types of specialists: specialists in machine learning, specialists in human–computer interface and specialists in psychology of explanations.

Applications from scientists for the program began to be collected in 2016, from the same year the term "explainable AI" has become steadily popular. Of course, DARPA is not the only one who has thought about explanable AI, there are many articles reviewing various conferences on explainable AI (Adadi and Berrad 2018) as well as groups of scientists trying to act independently, the most influential of them is FAT ML (www.fatml.org). But DARPA can be singled out for their centralization in this topic providing a single key term for the problem, verifying existing statements for clarity, as well as developing a single system for evaluating explanations. Therefore, the analysis of their research is the best way to understand what problems are in the explainable AI.

Unfortunately, the structure of the DARPA's program "Explainable Artificial Intelligence" did not provide a way out of Licklider's ideas about human–machine symbiosis, and therefore did not take into account both the degree of AI autonomy and the need not only to transfer mental models but also to implement them effectively in AI. The report on their activities shows that they themselves understand that their initial idea of a "universal interface" between humans and AI was not very successful (Gunning et al. 2021). During and after their program, many articles analyze their results, but they do not go beyond DARPA's views, and most often repeat their mistakes.

Hereafter, we will discuss why this happens, and what contradictions in the perception of explainable AI hinder research.

2.1.2 Diversity of Motives for Creating Explainable AI

The paper (Lipton 2018) highlights the following motives for the creation of explainable artificial intelligence (XAI):

1. Trust, i.e. providing such an explanation of actions that will provide confidence in the algorithm.
2. Causality, i.e. obtaining causal relationships in the considerations of the algorithm.
3. Transferability, i.e. using of explanations of the algorithm for more correct application of the output of old models in new situations. For example, the model may deduce the probability of death from pneumonia, but interpreting this probability as a level of need for immediate treatment can be threatening and lead to even more deaths (Caruana et al. 2015).
4. Informativeness, i.e. obtaining useful information from the model.

There are also different motives for the use of XAI (Adadi, Amina):

1. Justification of the AI decision.
2. AI control.
3. AI improvement.
4. Obtaining from AI data that are useful for research in this area.

Some researchers (Doshi-Velez et al. 2017) consider these motives as part of some usage scenarios of explainable AI, each of which needs to be optimized separately, but, unfortunately, not everything is so simple.

In the following, the internal contradictions of each of these motives and their contradictions between each other will be presented.

2.1.3 Internal Inconsistency of Motives for Creating XAI

At first glance, the most truthful explanations are credible. But first of all, it is not possible to determine the correctness completely, especially for a non-expert. Information is perceived by a certain person, adjusted to the format of his ideas and "verified" with their help. Thus we have that for different people trust can be caused by different factors:

1. For non-experts: the use of their domestic intuition
2. For experts: the relevance of explanations to their experience and intuition
3. For scientists: the use to explain the facts of theories known to them.

Each of these cases requires explanations of different levels of complexity and different formats of information (for example, populism may inspire confidence in non-experts but may irritate and distrust academicians). If the goal of XAI is to increase trust, it will be able to make good explanations of not the best decisions, knowing what the person knows and how he "verifies" the information. Including using well-known human cognitive errors, stereotypes, etc. Other articles (Merry, M.) also point out that trust cannot be the main factor for explainable AI.

Causality. At first glance, if AI has found a pattern between two characteristics, it does not mean that he "believes" that there is a causal link between them. There may be a third factor influencing these, and AI tries to detect it using available information, but to guess about this factor will not work trying to dig into the internal structure of AI.

There are researchers who try to get explanations by extracting them from the local parts of the neural network (Guidotti et al. 2018), believing that in this way it is possible to isolate a part of the "neural network thinking process", which is incorrect due to the above. The problem with this approach is also emphasized by other researchers (Del Giudice 2021).

Tolerability. XAI is most useful for ordinary users and not for scientists (Gunning et al. 2021). The latter have methods of using old AI for new tasks and other methods of effective implementation of old models in new areas. They also know better how to use their models. And for practitioners implementing AI systems to know how to do it, scientists may need to create instructions for using this AI so that it can be adapted to needs. For example, the optimization goal might be to minimize the expected number of patient deaths and decide to change the old strategy to improve the target (expected number of patient deaths), knowing the number of staff required

for different levels of patient care and the previous distribution of their deaths, based on the old strategy of resource allocation.

Informativeness. At first glance, informativeness is the ability to get as much useful information from the model as possible. Informativeness can also be defined as the benefit of the information that a person receives. Although both of these definitions sound the same, they are not. According to research (Gunning et al. 2021), if the task is simple enough, then additional explanations will only annoy the person, and if the task is time-limited and/or already has a high cognitive load, then a large number of recommendations can even reduce the person's performance. In addition, the benefits of recommendations may change over time. That is, information should be useful for a particular person at a particular time in a particular situation.

As you can see, even within one motif we have different inconsistencies and contradictions. Of course, there are even more contradictions between the motives themselves.

2.1.4 The Contradiction Between the Motives for Creating Explainable AI

Trust is primarily a subjective feeling and it has nothing to do with understanding the work of the model or understanding its effectiveness. It has to do with how well something fits into a format that a person understands. And more precisely, it is related to our attitudes that work with data in formats that are clear to us and that tell us to trust or not to trust. Moreover, if the attitude of distrust and the attitude of trust is activated, then a person can both trust and distrust the subject. Some researchers even consider trust and distrust as independent and autonomous concepts (Van De Walle and Six 2014).

We can trust our loved ones in their judgments even more than experts. In addition, trust can be evoked by sympathy, positive associations, and how familiar you are with this object, using such factors, advertising can even lead to more trust without explaining how their mechanics or drugs work. There are many things that the advertising does not show as it is very difficult to fit into a format that is understood by the target audience. For example, drug advertising will not go into detail about the effectiveness and contraindications, which is the most useful information for the end user because its purpose is to interest society in their product and create some initial level of trust. In other words, trust is not related to causality (as, for example, the real reasons for the effectiveness of drugs can not be communicated to the end user), nor to tolerability (as in different areas of human life has its own settings), nor especially with information.

Causality can give trivial reasons that are already clear to man, and therefore can not help the person. In turn, identifying non-trivial causes is almost impossible, as the interdependencies in AI do not indicate a causal relationship. That is, limiting AI

so that it calculates everything due to the logic of causality can not only worsen the effectiveness of AI but also its informativeness.

Informativeness is usually understood as additional useful information, that is, one that is not obvious. But the model can learn to extract useful facts only from some area of application. Thus, with one-sided AI training to obtain information, tolerability is reduced. How to fix it we will talk in Sect. 2.3.

2.1.5 Paradigm Shift of Explainable Artificial Intelligence

We have considered many motives for the creation of AI and showed the contradictions in themselves and between them. These contradictions follow from the very procedure of constructing definitions for AI:

1. A human has his own needs to interact with AI
2. A human states the solution of these problems as an element of explicable AI
3. A human tries to generalize this problem, but due to limitations in its competences it does so only within its field of knowledge, while stating that this is XAI.
4. A human adjust other definitions of XAI for themselves, still highlighting from them only what makes sense within a fixed area of knowledge.

As already shown, different groups have different needs: trust, causality, tolerability, informativeness. They also have different formats in which it would be more convenient for them to interact with AI, as well as different types of information. The reason for the contradiction in the motives for the development of AI was stage 4: each considered how other needs relate to their needs without understanding the problems of other groups, and therefore not seeing obvious contradictions between these motives and the need for more correct generalization.

Some researchers say we just have to live with it, and the only thing we can do is try to structure all this variety of definitions of explainable AI in order to at least slightly reduce the formal and legal problems that arise when introducing AI in critical areas (Amann et al. 2022). But such an approach does not solve, but only slightly weakens the systematic problem. And in order to solve it and not come into conflict with the existing needs that are displayed in different definitions of explainable AI, it is necessary to understand the very motives for creating explainable AI.

There are good works that study the motivations for the use of AI in specific areas by interviewing employees of a particular company who specialize in the application of explainable AI in this area (Gerlings et al. 2021). But since they have not developed an understanding of the process of creating needs from the root cause, they often omit some groups of people who also play a key role in creating AI and also need explainability in accordance with their needs.

At the very beginning, the motive for the creation of XAI is the same as the motive for the creation of AI—to improve life through automation and optimization of existing processes. Since our interaction with AI has already become part of our

everyday processes (such as recommendations in Google, Facebook, etc.), it also fell under the need to be optimized, and XAI is the optimization of this process.

Considering XAI as an AI that optimizes the interaction of AI with a person or group of people, we have several possible directions for the development of XAI:

1. Learning an AI to interpret the internal processes of other AI.
2. Teaching an AI to get explanations based on the internal states of other AI.
3. Extending the old AI and training its multi-task abilities.
4. Learning a new multi-task AI that will display the necessary results and explanations.

It is known that multi-task contributes to more efficient internal representations, which can even improve the target output of AI. The fact that this method of learning is leading in the following models of XAI is indicated by some articles (Gunning et al. 2021).

2.2 Proposed AI Model

2.2.1 The Best Way to Optimize the Interaction Between Human and AI

To understand how to optimally adjust the interaction between AI and humans, it is necessary to look at how it is arranged and how human intelligences optimize interation with each other. By Piaget (2001) everybody from the beginning of the period of cognitive development named "preparation and organization of specific operations" (approximately 2 years) unconsciously or consciously receives a mental model of the world. Thanks to it, it can operate not only with external objects but also with concepts about them. From the age of 2, a child can listen to simple stories by ear and say simple sentences, perceive simple mental models and pass them on. In the future, a person learns to effectively express these elements of the internal symbolic model in a way that is understandable to others, that is to externalize them. Being in different social groups, a person gets different skills of externalization of this model, when talking: in the family, with educators in kindergarten, with peers, etc. The same goes for adults. For example, for groups of owners of certain resources (for example: water, land) for effective cooperation there are mental models such as ARDI (Actors, Resources, Dynamics, and Interactions) (Etienne et al. 2011). Theory of explanation is studied by various variants of externalization of internal symbolic models in the form of various mental models.

The idea of using the analogy of human communication and mutual learning to improve the explainability of AI is not new (Gallina et al. 2020). What these studies often don't include is that there are different explanatory needs at different stages of AI development and deployment, requiring vastly different mental models and explanatory skills.

In previous studies, we have written about how people's skills are organized like a tree, and how to transfer them to AI in general (Chergykalo and Klyushin 2021). In our case, we will focus as much as possible on the skills of explanations. This can be imagined as a branch of the skill tree with its branching into different skills. We will show in Sect. 2.3 how to optimally grow this branch. And before that, we need to show how explanation skills differ from forecasting skills.

2.2.2 Forecasts Are not Necessarily Useful Information

As we wrote in our previous research (Chergykalo and Klyushin 2021), people, when trying to think consciously, conduct a set of simulations of the future, through which we improve our function of choice. This principle is quite universal and can be used to improve the choice of AI. But this can only work when the AI knows the goal it is optimizing to make optimal decisions and has information about the external environment to predict its response to its actions.

In the case of systems where AI needs to cooperate with humans, to model events within AI, humans become part of the external environment that needs to be influenced so as to improve targets. If AI has the ability to influence the system, it, in the case of existing knowledge about the system, will build a set of possible scenarios for the future (the choice between which depends on its impact) and choose the most optimal of them. But when AI has no influence, or is modeled as if it does not have it, then AI can only predict one future—one that would be without its influence. AI cannot optimize anything if it has no effect on this system.

For example, a model can predict the probability of dying from pneumonia and indicate that the lowest mortality is from asthma. But this figure is maintained only by aggressive treatment. Therefore, if the prediction of the model affects physicians and their managers in such a way that they pay less attention to asthma and more to patients with symptoms that are predicted to be more associated with mortality, then mortality will only increase (Caruana et al. 2015). It is very interesting that when learning to predict the future, researchers do not teach the model to solve some problems, but teach it to model real processes.

For example, researchers may think that information about whether or not a person will be convicted in the future may help identify the most dangerous people to re-educate them. But when giving preliminary information for AI about a person, such researchers often begin to wonder why AI begins to use stereotypes about people to predict a person's future sentence. Picking up these results, publicists begin by accusing AI systems of racism or other prejudices. But, not surprisingly, their predictions of the existing AI systems show the problems of people who do not know how to fairly (even within the same legislation) to judge other people.

It is easy to cite examples of the fact that many characteristics of a person that he has throughout life and that are not related to his behavior subconsciously influence the decisions of others, including the decisions of judges and juries. The simplest such characteristic is attractiveness. There is good literature showing that people with

good looks has more chances for leniency (Castellow et al. 1990; Downs and Lyons 1990). Some studies (Stewart 1980; Beaver et al. 2019) show that attractiveness can halve the likelihood of imprisonment. Also, attractive people on average pay twice less fines in court for negligent damages (Kulka and Kessler 1978). But these are just some of the multipliers that reduce the penalties for attractive people. Attractive persons have many more privileges, such as reducing the filing of lawsuits themselves, supporting attractive people during difficult times, etc. (Benson et al. 1976).

There are many psychological factors that only reinforce this arrangement, such as the crowd effect. If you know that the majority of people vote for a decision, then subconsciously you also want to vote for it. For example, in the case of severe punishment, it will be more likely to be imposed when jurors vote by show of hands than by secret ballot (Kerr and MacCoun 1985). Probably the most interesting thing is that in all the above cases, the jury will be fully confident that they have made independent and fair decisions and generate plausible explanations for their decisions.

Quite often, advocates of ethics want AI, even if it has some idea of the person, not to take into account some factors. This requirement is very interesting, as these people represent that people may also, if necessary, ignore some factors. Of course, this is not the case, because if a judge requires jurors not to take into account some information that puts the defendant in a negative light, the jury's decision will take this information into account even more (Lieberman and Arndt 2000).

It is clear that in order for AI to interact effectively with people for a common goal, it needs to understand how to properly influence other people's actions through the explanations and recommendations it makes for them. He needs to understand how the whole system is designed to anticipate the actions of different people at different "stages of its life" to help researchers better tweak it, to help programmers integrate AI systems into our daily processes, to help users better adapt and help to extract long-term useful information during their interaction. In the following, we will talk about this system and how the old measures of explained AI are related to the effectiveness of AI–human interaction and practical goals.

2.2.3 Criteria for Evaluating Explanations

Researchers have already developed several explanation scoring systems, the most studied being the explanation scoring system (ESS) (Gunning et al. 2021). It consists of three main blocks:

1. Functionality measures:
    1.1. Speed generation of explanation and its assimilation by a human)
    1.2. Type of modality (visual, textual, etc.)
    1.3. The content of the explanation (justification, examples, reasons, connections that influenced the result (effect relations)
    1.4. Request for additional explanations (natural language, multiple choice, drill-down, etc.)

2. Learning performance measures. Checking the model on a test sample
3. Explanatory effectiveness measures:

    3.1. Explanation satisfaction
    3.2. Explanation goodness
    3.3. Mental model understanding
    3.4. User-machine task performance
    3.5. Trust assessment.

But this model does not reflect a comprehensive assessment of how the processes of AI interaction with humans and groups of people are optimized. To obtain this assessment, we will divide this process into several stages:

1. Stage of system development
2. Stage of system implementation
3. Stage of individual adaptation of people under AI and AI under people
4. Stage of joint practical activities
5. Stage of obtaining new practical knowledge from AI solutions.

Next it is necessary to choose a criterion that fits the area in which AI interacts with humans. For example, we will take such an area as medicine. The criterion of its success is the acceleration of the reduction in patient mortality. This criterion dictates what we need as soon as possible and at a sufficient level of quality to pass the (1), (2), (3) stage and optimize the transitions between (4) and (5) stage. Thus, even such a subjective feeling as trust in AI becomes an important problem if we consider it as a barrier (at the stage of individual adaptation) to the introduction of effective AI methods in medicine. We can say that at this stage it is really desirable to take into account human psychology (explanation satisfaction, trust assessment) so that a person begins to see in AI his colleague. And then, when it is necessary to make practical decisions we must be adjusting to the level of additional cognitive load that a person can afford to focus on the effectiveness of the transfer of mental models. Also, by analyzing the results (how much better/worse the patient ended up), it is vecessary constantly adapt the interaction with fellow doctors for greater efficiency (discussing existing situations, taking recommendations from doctors and giving recommendations to themselves).

If doctors do not perform surgeries or surgeries and, for example, only consult patients, the algorithm can not only give their view of the situation but also try to identify errors in the doctor's explanations or point out potential errors. Of course, this requires a large dataset that should show which doctors with what level of experience make which mistakes. This is necessary so that the amendments made by XAI are not meaningless for this doctor. Surgeons, if they have a difficult choice of methods that they want to use during surgery, can consult with AI before surgery.

How to create the necessary dataset and choose the architecture of XAI, so as to teach him to give explanations and recommendations at each of the above stages, we will tell in Sect. 2.3. And before that, let's show why understanding to whom and why explanations are so important, and what mistakes previous researchers made.

2.2.4 Explainable to Whom and Why?

Recently, researchers (Ribeiro et al. 2016; Gunning et al. 2021) define and work with explanations as with a certain universal interface. But as has already been shown for different groups of people, XAI needs to give different explanations, this is necessary so that XAI can not only explain to doctors or other end users, but also to integrate new technologies more quickly and painlessly into vital areas. Therefore, the approach with a universal interface is not correct in the sense that the interface is not universal but only for the end user.

However, one interface is useful just be the fact that it open the way for further development. Of course, this is only the first step in XAI, where a researcher who studies how best to create an interface plays the role of a UI/UX designer in an area such as human–computer interaction (HCI). In this area, there are so-called user tests that designers use to determine how information is perceived by users, how they cluster it. Using this information, they determine the optimal interface that minimizes the cognitive load on the user. To do this, they display information so that it already has a typical clustering and shows information that does not exceed a certain limit of cognitive load.

Researchers of the explanation interface faced a problem that is not so pronounced in HCI-designers, even basic explanations from the black box are quite difficult to obtain. To do this, they used a number of non-experts to select the interpreted indicators and give them an explanation. But even at this level it is already possible to identify problems such as "Explainable to whom and why?" For example, there are articles on XAI (Ribeiro et al. 2016) that select visualization and explanation techniques to help AI professionals find feedback loops or data leaks by taxing this interface so that even a non-expert can navigate. This is certainly useful for programmers who are trying to integrate AI into existing systems in order to improve the performance of a particular AI model and know its strengths and weaknesses to better integrate. But this is of no use to researchers who are creating new AI architectures themselves.

The same paper (Ribeiro et al. 2016) gave an example of the erroneous inference of the CNN neural network and shows that the "attention" of the neural network at this time was not focused on the object. Similar information has been presented in this article in the form that it can be useful to both AI developers and physicians. But it will not help neither the developers nor physicians. It can only help in the second stage of AI implementation, for programmers who adjust the model to real processes. For AI researchers, i.e. those who select the optimal AI architecture, such information is of no use also. For example, using the Fast R-CNN architecture, which first defines the area and then determines the type of object, you can easily include attention of AI as a factor to control and optimize.

It is more obvious that unnecessary information will only strain the doctor and may confuse him. For example, the fact that the coefficients of attention are more in the background does not mean that the internal picture is completely ignored, some of its elements may still stand out. It will be difficult for the doctor to navigate in

such cases and what they should tell him. At best, the programmer who implemented the system can give their functions of assessing whether to trust the system or not, which may already be useful to the doctor.

Usually, this method is not optimal for the interaction of doctors with AI. First of all, some articles (Merry et al. 2021) point out the lack of literature on how to improve the team work of AI. Improving the exchange of mental models has already proven to be a practice that leads to increased efficiency of care (Page et al. 2016). Mental models transmitted in operating rooms are already well studied (Nakarada-Kordic et al. 2016), and therefore it is easy to understand in what format should be the explanation of XAI in these cases, which may reduce retraining for AI.

2.3 Proposed Architecture

2.3.1 Fitness Function for Explainable AI

Summing up Sects. 2.1 and 2.2 we see that:

1. The main purpose of XAI is to optimize the processes of human interaction and AI.
2. Explanations cannot be evaluated by themselves, it is always necessary to focus on a key criterion in each area of application (reduction of errors of judges, doctors and increase their effectiveness).
3. The task of optimizing the processes of human interaction and AI can be effectively divided into 5 subtasks related to the 5 stages of the life cycle of AI.
4. At each stage of the XAI should give different types of explanations for different groups of people and for different subtasks.
5. Some of the explanations can be given by experts at the previous stages of AI implementation, in this case the XAI will add its explanations “on top” of the basic ones.
6. Knowing whether experts will give explanations to experts at the next stages has a strong influence on the format of XAI explanations, which shows that the XAI implementation system should be immediately ready to be established.

Learning the XAI system to give explanations at each stage, and to different groups of people, causes a multitask with a large number of sub-tasks to explain each of which must be evaluated. Fitness function can be the sum of all errors from different subtasks with weights. Each weighting factor will be determined in advance by the extent to which the explanations from the experts in the previous stages satisfy the experts on this subtask. If the explanation from previous experts is completely satisfactory, you can not even teach XAI this subtask. The more XAI explanations on this subtask can be useful, especially for the target, the higher the coefficient.

2.3.2 *Deep Neural Network is Great for Explainable AI*

Often simple models are more trustworthy than others. For example, it is enough to deduce the coefficients of the linear model in the diagram and you can already have some idea of their work. But with so many variables, it's unclear how they relate to each other and how to single out abstract concepts that can help with interpretation. Some co-researchers additionally teach sparse linear models to obtain more perceptible learning outcomes. But there is a clear ineffectiveness of such approaches, which is also pointed out by other researchers (Ribeiro et al. 2016).

There are also many tools for visualizing decision trees, with the help of which it is possible to understand more abstract concepts. But although they are good at capturing monotonic properties of data, they do very poorly to non-monotonic abstract characteristics, which, in complex problems, arizes no less frequently than monotonic.

It is easy to understand that the way non-monotonic abstract characteristics are calculated is quite difficult to depict by any methods, and even more difficult to understand what they mean. The task arises to ensure that these characteristics are somehow interpreted in a way that is clear to us humans. And the most effective teaching methods that work best with information and most effectively display it in text form are neural networks.

Neural networks can be taught to interpret some of the characteristics of simpler AIs by explaining them. But the most linear models, decision trees, etc. do not know how to effectively extract abstract characteristics from the data. What we have is that if we have already agreed that the neural network is best taught to explain something, the best way to give it information is to use another neural network or give it initial data directly. Thus, we have a neural network that encodes information, and this information is processed by two other neural networks: one displays targeted decisions (classification, etc.) and the other displays comments to explain their decision and possibly some recommendations.

As has already been shown by other researchers (Ribeiro et al. 2016), that if there are coefficients of attention in the architecture of the neural network-decoder, then it is possible to know what plays what role no worse than with linear models. The fact that neural networks are less clear than linear models is a myth is also pointed out by other researchers (Lipton 2018) Talking about the so-called post-hoc explanations.

The fact that AI now needs to solve both the main task and the task of explaining their actions is multitasking. We will talk further about why this is even good for the effectiveness of the main task.

2.3.3 *The More Multitasking the Better*

In our previous research (Chergykalo and Klyushin 2021) we have shown that multitasking is part of the work of biological neural networks. In addition, we have shown

that it is useful. Even when the tasks are not related to each other, mutual learning can help to regulate the internal representations, which can even slightly increase the efficiency of each task separately (Romera-Paredes et al. 2012). As mentioned above, when tasks are related, neural networks can even better shape internal representations. And in order to form all the necessary set of ideas in the neural network, it is good to teach it to explain their actions for the maximum number of groups of people and for different cases.

Thus, it is desirable to teach the neural network during each of the five tasks and also slightly adjust the amount of text for explanations or simplify some parts of the mental models that it displays or, conversely, learn to focus on certain aspects of mental models. All this is necessary when a person is cognitively overloaded or, conversely, considers the task very simple and may not catch important details.

2.3.4 How to Collect Multitasking Datasets

The use of mental models due to the existing fixed format can reduce the need for the size of datasets. But the main problem is that the conclusion for different explanatory tasks must be consistent. Because it will be very strange and confusing to see a case where a model will tell one group one thing and another another. From the outside, it will even look like the XAI is lying and will activate people's distrust.

The simplest and most reliable way out of this situation is to form groups of experts from different fields, and when this group solves some tasks, it must give a comprehensive explanation of its decisions in different mental models. For example, how would they explain their decision to the ethics team, how would they explain their decision to a fellow doctor, how would they explain the possible problems of their decisions to AI specialists (who will then see these explanations from the AI itself), and so on.

It is best to have the best experts in these groups who will give the best solutions and the best explanations. XAI trained on such data will be of the greatest benefit. But most likely there will not be much data from them. Therefore, as another option, it is proposed to include experts with an intermediate level of qualification—the main thing is that the expert who is key in making the decision should be at a fairly high level. In this case, it may be useful to include in the training data on the extent to which a highly qualified team of experts has made their decisions.

2.3.5 Proposed Neural Network Architecture

Some of the five subtasks are divided into their subtasks. For example, at the implementation stage AI can be evaluated not only by programmers but also by ethics specialists. At this stage, the greatest focus is given to how AI works with data, so a good solution is to use the reciprocal part for both of these tasks. Other researchers

have also written that subtasks in multitasking should be broken down into an effective hierarchy (Zweig and Weinshall 2013). Below we demonstrate the hierarchy of skills of explanations which will serve as a basis for our tree-like architecture of a neural network (Fig. 2.1).

In order to better understand this architecture, we offer a simple example of what such a neural network might look like. Showing it in Fig. 2.2, we omit the details (sizes of convolution kernels, activation functions etc.).

For this neural network, we can take for example the following fitness function:

$$f = E^2 = {E_{pred}}^2 + \frac{1}{100}\left({E_{adv1}}^2 + {E_{adv2_1}}^2 + {E_{adv2_2}}^2 + {E_{adv3}}^2 + {E_{adv4}}^2 + {E_{adv5}}^2\right)$$

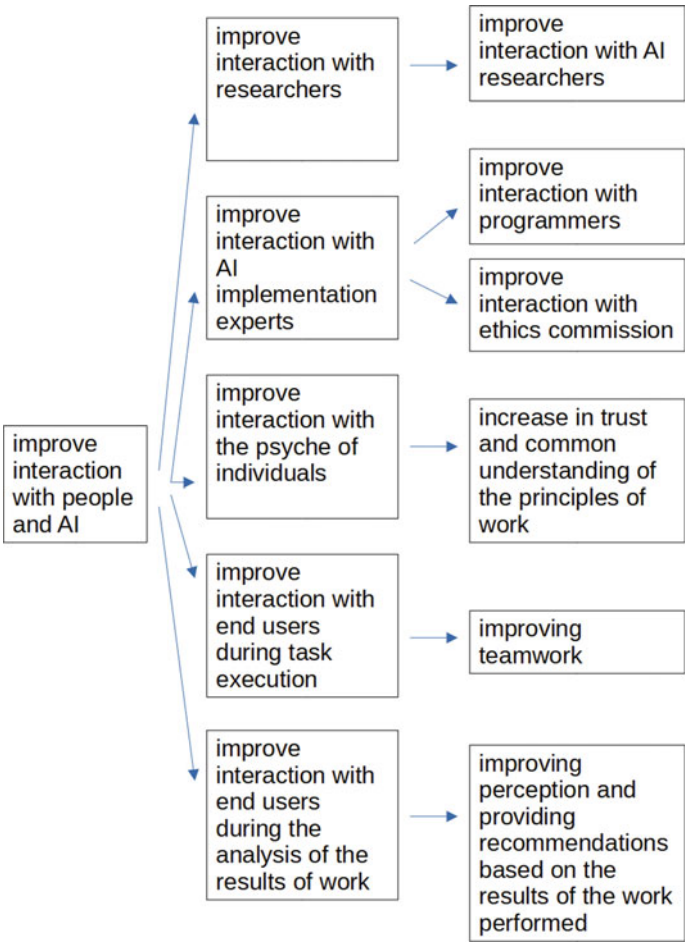


Fig. 2.1 Tree-like architecture of a neural network for XAI

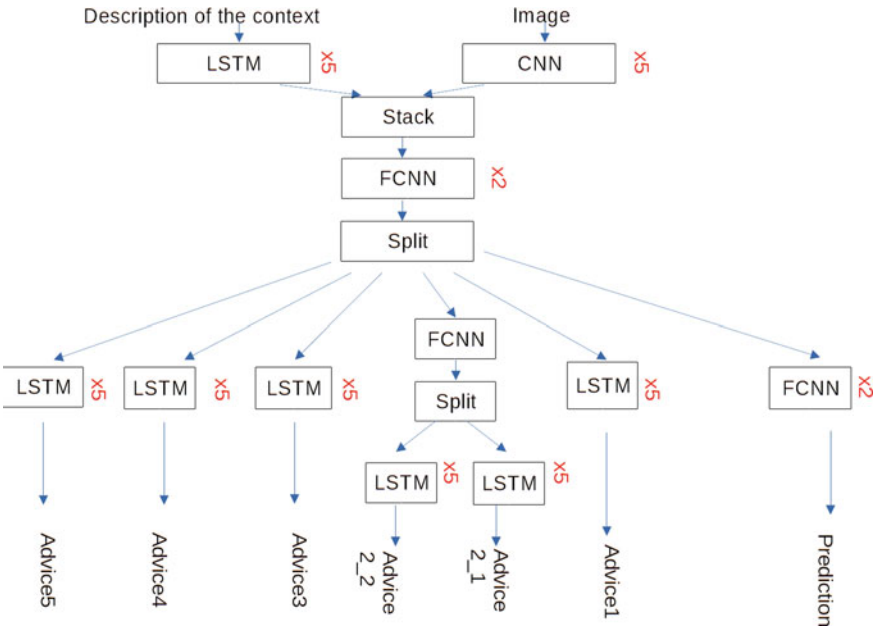


Fig. 2.2 A simple example of a neural network implementation. Xn—number of repetitions of layers

where—corresponding errors calculated for each neural network output.

We will also describe step by step how to adapt the one described in Fig. 2.1 architecture for solving domain problems on the example of medical image analysis:

1. Definition of versatile key data on the basis of which experts draw their conclusions. In the case of medicine, this is: a description of the context—a description of the mental state of the client and its background, a image—data obtained after the survey through medical devices.
2. Data collection as described in Sect. 2.3.4.
3. Choosing an appropriate architecture for encoding incoming data, transforming them and decoding the received pieces of information. For example encoding can be done for text data via Bert model and for images via ResNet convolutional layers.
4. Construct the fitness function formula in accordance with Sect. 2.3.1.
5. Train this neural network and implement its explanations in all areas described in Fig. 2.1. And thereby accelerate the creation, implementation, understanding and interaction with AI.

2.4 Conclusions

The effectiveness of XAI development is primarily determined by its implementation system. Therefore, there is a need for centralized and harmonized formats of mental models to be transmitted, as well as a centralized and harmonized XAI implementation policy. XAI optimizes the interaction between humans and AI, and the best and most experienced models of effective human-to-human interaction, and, by analogy, between human and AI, are mental models. We would also like to add that researchers avoid the phrase "learn from AI", but their needs and requirements, they show that this is one of the greatest values in these studies. Multitasking is the most promising area of XAI architecture. Taking into account this and other information, we have built a better and more comprehensive assessment of the effectiveness of XAI, which corrects errors in the purpose of XAI from past researchers. In addition, we have shown the general steps to implement a model that will satisfy this large set of requirements.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Amann, J., et al.: To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health **1**(2), e0000016 (2022)

Beaver, K.M., Boccio, C., Smith, S., Ferguson, C.J.: Physical attractiveness and criminal justice processing: results from a longitudinal sample of youth and young adults. Psych. Psychol. Law Interdiscip. J. Australian New Zealand Assoc. Psych. Psychol. Law **26**(4), 669–681 (2019)

Benson, P.L., Karabenic, S.A., Lerner, R.M.: Pretty pleases: The effects of physical attractiveness on race, sex, and receiving help. J. Exp. Soc. Psychol. **12**, 409–415 (1976)

Bob, P.: The brain and conscious unity: Freud's omega. Springer Science + Business Media (2015)

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, M.: Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, pp. 1721–1730 (2015)

Castellow, W.A., Wuensch, K.L., Moore, C.H.: Effects of physical attractiveness of the plaintiff and defendant in sexual harassment judgments. J. Soc. Behav. Pers. **5**, 547–562 (1990)

Chergykalo, D.O., Klyushin, D.A.: Biomorphic artificial intelligence: achievements and challenges. In: Hassanien A.E., Taha M.H.N., Khalifa N.E.M. (eds.) Enabling AI Applications in Data Science. Studies in Computational Intelligence (Springer, Cham), vol. 911, pp. 537–556 (2021)

Del Giudice, M.: The Prediction-Explanation Fallacy: A Pervasive Problem in Scientific Applications of Machine Learning. PsyArXiv. December 13 (2021). https://doi.org/10.31234/osf.io/4vq8f

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

Downs, A.C., Lyons, P.M.: Natural observations of the links between attractiveness and initial legal judgments. Pers. Soc. Psychol. Bull. **17**, 541–547 (1990)

Etienne, M., Du Toit, D.R., Pollard, S.: ARDI: a co-construction method for participatory modeling in natural resources management. Ecol. Soc. **16**(1), 44 (2011). https://www.ecologyandsociety.org/vol16/iss1/art44/. Accessed February 6, 2022

Gallina, B. et al.: Towards explainable, compliant and adaptive human-automation interaction. In: 3rd EXplainable AI in Law Workshop (XAILA 2020) co-located with 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020) (2020). http://ceur-ws.org/Vol-2891/

Gerlings, J., Jensen, M.S., Shollo, A.: Explainable AI, but explainable to whom? arXiv preprint arXiv:2106.05568 (2021)

Graves, A., Wayne, G., Danihelka, I.: Neural Turing Machines. ArXiv, arXiv:1410.5401 (2014)

Greenblatt, S.H.: (1999) John Hughlings Jackson and the conceptual foundations of the neurosciences. Physis Riv. Int. Stor. Sci. **36**(2), 367–386 (1999)

Gunning, D., Vorm, E., Wang, J.Y., Turek, M.: DARPA's explainable AI (XAI) program: a retrospective. Appl. AI Lett. 2: e61. (2021). https://onlinelibrary.wiley.com/doi/full/https://doi.org/10.1002/ail2.61. Accessed February 6, 2022

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 (2018)

Hayek, F.A.: The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology. University of Chicago Press (1952)

Hebb, D.O.: The Organization of Behavior. Wiley, New York (1949)

Kerr, N.L., MacCoun, R.J.: The effects of jury size and polling method on the process and product of jury deliberation. J. Pers. Soc. Psychol. **48**, 349–363 (1985)

Kulka, R.A., Kessler, J.R.: Is justice really blind? The effect of litigant physical attractiveness on judicial judgment. J. Appl. Soc. Psychol. **4**, 336–381 (1978)

Lieberman, J.D., Arndt, J.: Understanding the limits of limiting instructions. Psychol. Public Policy Law **6**, 677–711 (2000)

Licklider, J.C.R.: Man-computer symbiosis. IRE Trans. Human Factors Electron. HFE-1:4–11 (1960)

Licklider, J.C.R.: Memorandum for members and affiliates of the intergalactic computer network. Adv. Res. Projects Agency (1963)

Lipton, Z.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018)

Merry, M., Riddle, P., Warren, J.: A mental models approach for defining explainable artificial intelligence. BMC Med. Inf. Decision Making **21**, 344 (2021)

Nakarada-Kordic, I., Weller, J.M., Webster, C.S., Cumin, D., Frampton, C., Boyd, M., Merry, A.F.: Assessing the similarity of mental models of operating room team members and implications for patient safety: a prospective, replicated study. BMC Med. Educ. **16**(1), 229 (2016)

Page, J.S., Lederman, L., Kelly, J., Barry, M.M., James, T.A.: Teams and teamwork in cancer care delivery: shared mental models to improve planning for discharge and coordination of follow-up care. J. Oncol. Pract. **12**(11), 1053–1058 (2016)

Piaget, J.: The Psycology of Intelligence. Routledge, London (2001)

Ribeiro, M., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (San Diego, California), pp. 97–101 (2016)

Romera-Paredes, B., Argyriou, A., Bianchi-Berthouze, N., Pontil, M.: Exploiting unrelated tasks in multi-task learning. Proc. Mach. Learn. Res. **22**, 951–959 (2012)

Stewart, J.E.: Defendant's attractiveness as a factor in the outcome of criminal trials: an observational study. J. Appl. Soc. Psychol. **10**, 348–361 (1980)

Van De Walle, S., Six, F.: Trust and distrust as distinct concepts: why studying distrust in institutions is important. J. Compar. Policy Anal. Res. Pract. **16**(2), 158–174 (2014)

Zweig, A., Weinshall, D.: Hierarchical regularization cascade for joint learning. Proc. Mach. Learn. Res. **28**(3), 37–45 (2013)

Chapter 3
An Overview of Explainable AI Methods, Forms and Frameworks

Dheeraj Kumar and Mayuri A. Mehta

3.1 Introduction

Artificial Intelligence (AI) has become an integral part of many applications in recent years. AI has reached the masses with the availability of intelligent software to automate systems. Machine learning models like a decision tree and Bayesian network are easily interpreted by the realization of the importance of each feature to the output. However, models developed using a deep neural network (Shi Zhang and Zhu 2018; Ras et al. 2018; Gilpin et al. 2018) do not allow understanding its internal mechanism. Due to the lack of transparency in the deep neural networks, it is hard to justify their predictions to the end-users. In addition, the stochastic nature of predictions made by them is another obstacle to understandability and transparency for humans. Thus, explaining the model developed using a deep neural network (often referred to as the black box model) is essential for users to accept AI-based solutions (Inam et al. 2021).

A black box model does not disclose its internal design and inference mechanism while making a prediction. Moreover, explanations are also necessary to evaluate the ethical and moral standards of a machine (Islam et al. 2021; Doran et al. 2018). Explainable Artificial Intelligence (XAI) aims to create a suite of techniques and frameworks that explain and interpret predictions made by black box models. The explainability of a black box model is the ability to explain its prediction in an understandable form for end-users. The goal of XAI is to communicate to end-user why a

D. Kumar (✉)
Department of Information Technology, Parul Institute of Engineering and Technology, Parul University, Vadodara, India
e-mail: dhirajsingh66@gmail.com

M. A. Mehta
Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Sarvajanik University, Surat, India
e-mail: mayuri.mehta@scet.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_3

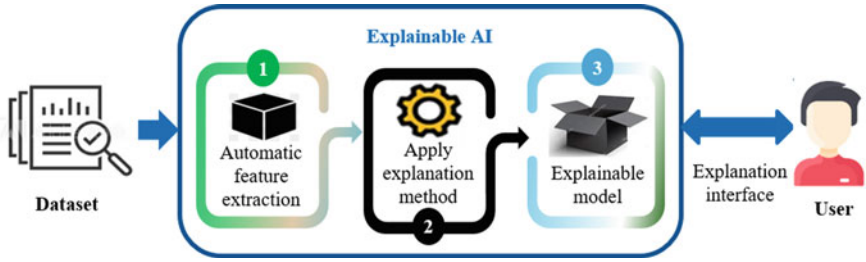


Fig. 3.1 Pipeline of an explainable AI method

black-box model made a particular decision. Moreover, it helps the researcher retrace the model functionality for deducing the inference mechanism. The methods that explain and interpret the internal mechanism and the decisions of machine learning and deep learning models are known as explanation methods. These methods explain the model decision and logic in different forms of explanation. Figure 3.1 illustrates a general pipeline of an explainable AI.

Interpretability is another term commonly used as a synonym for explainability. However, interpretability is the capability of a model to provide inference in an understandable form. An interpretable model enables end-user to prevent model bias, get the feature importance on model output, test reliability and ultimately assist in debugging the model. The complexity of a machine-learning model is directly related to its interpretability. Generally, complex models have higher accuracy, but they are more difficult to interpret and explain because they are developed with many hypermeters. Furthermore, the interpretability of machine learning and deep learning models often compromises the accuracy and overall performance of the models. Thus, the most straightforward way to get to an interpretable model is to design an intrinsically interpretable model.

The model explanation is critical when an AI-based system generates an unexpected result. The explanation method enables control over a model, which helps identify and rectify flaws and unknown biases. Moreover, the explanation method justifies the decision made by a model to answer why a particular decision was made (Failed 2022; Adadi and Berrada 2018). When the user understands how a model produces the result, they can easily improve the model to enhance its capability. The user understanding of the model boosts users' trust in critical decision-making using an AI-based system. It also enhances trust and human acceptability toward AI-based solutions (Inam et al. 2021; Linardatos et al. 2021). Additionally, the model explainability enables an AI-based system to add new knowledge to its knowledge base and learn new rules or behaviour for smart decision making. As per literature, the explainability of AI systems is needed mainly for four reasons: (1) to justify decisions, (2) to control the inner working of black box models, (3) to improve the model to produce expected results, and (4) to build the trust of human on model results. Figure 3.2 illustrates the reasons and needs for explainable AI.

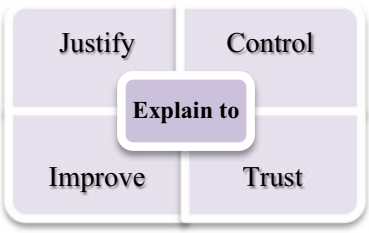


Fig. 3.2 Need for explanation in AI-based system

The research community introduced a variety of explanation methods and used them to produce understandable results from the black box model. This chapter aims to give a concise overview of different forms of explanation, existing XAI methods and frameworks that help researchers select the best possible method as per their application. The rest of the chapter is organized as follows: Sect. 3.2 presents the proposed classification of methods suitable for XAI and their brief description. Section 3.3 describes different forms of explanation useful for XAI. Section 3.4 presents a review of the six most popular XAI frameworks. Finally, the chapter is concluded in Sect. 3.5. In addition, our observations and future directions are discussed in this section.

3.2 XAI Methods and Their Classifications

This section discusses techniques and approaches used to explain black box models. It's important to incorporate explainability for better trust and understanding in black box models (Inam et al. 2021). Numerous XAI methods have been developed to explain the inner working of black-box models and their predictions (Linardatos et al. 2021). As shown in Fig. 3.3, we classify these methods in four different ways by considering different aspects of each method: (1) based on the scope of explainability, (2) based on implementation, (3) based on applicability, and (4) based on explanation level. This classification helps select suitable methods for explaining and interpreting black box models.

3.2.1 Based on the Scope of Explainability

The scope of explanation is one of the ways to classify explainable AI methods. It characterizes the extent of an explanation generated by an XAI method. An explanation can either describe the entire model or partially describe the model based on individual input instances (Mohseni et al. 2021). According to the scope of explainability, the explanation can be global or local (Ras et al. 2018; Doran et al. 2018; Adadi and Berrada 2018; Doshi-Velez and Kim 2017).

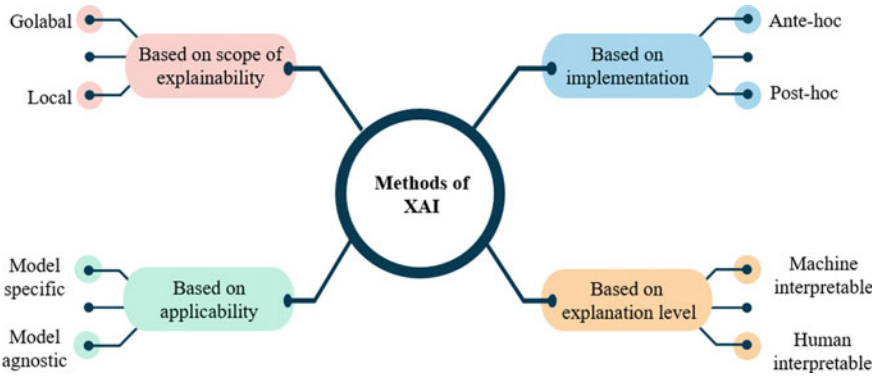


Fig. 3.3 Classification of explanation methods

The global methods are used for the explainability of the whole model. They provide a global explanation for the model's inner workings and decision-making (Ibrahim et al. 2019). A global method approximates the overall behaviour of a model to produce general representations of the model's relationship with its input instances. Symbolic representation is a commonly used technique to generate an interpretable representation of all predictions made by the model (Confalonieri et al. 2021). Bayesian rules and generalized additive models are a few examples of the global explanation method (Alicioglu and Sun 2022).

Unlike the global method, local methods are used to explain and understand individual predictions of a model. Explanations are built by generating local surrogate models considering individual prediction and input instances. The surrogate models are intrinsically interpretable models used to explain a complex model. They are trained on the predictions of the black box model to generate an explainable model. Local explanation methods often use saliency methods (Adadi and Berrada 2018; Linardatos et al. 2021; Angelov et al. 2021) to explain the relationship between specific input–output pairs (Mohseni et al. 2021). However, the local methods' explanations vary greatly depending on the instance considered (Confalonieri et al. 2021). Local methods like Local Interpretable Model-Agnostic Explanations (LIME) (Palatnik de Sousa et al. 2019), Shapley Additive Explanations (SHAP) (Failed 2019) and Deep Learning Important FeaTures (DeepLIFT) provide local explanations for an instance of a model. Generating model explanations using the local method is easier than explainability using the global method.

3.2.2 Based on Implementation

The explanation of the model can be generated during the training of a model, or it can be generated after model creation. Based on the way the explainability of the model

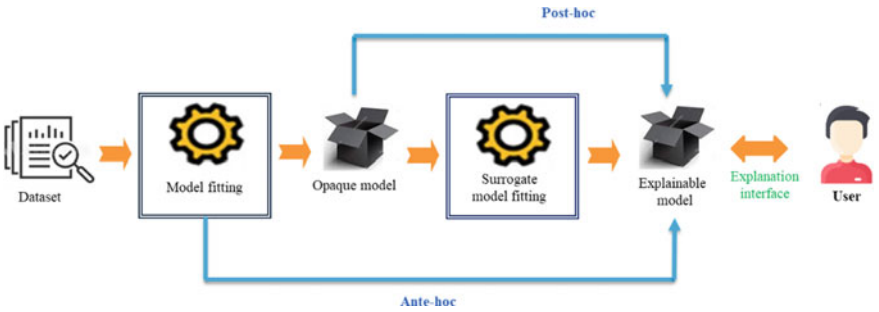


Fig. 3.4 Working of Ante-hoc and post-hoc methods

is implemented, XAI methods are categorized into two categories: ante-hoc and post-hoc methods. Figure 3.4 illustrates the working of ante-hoc methods and post-hoc methods. Ante-hoc methods are used to generate explanations from the beginning of the model training. They incorporate explanation directly into the structure of a model during design (Holzinger et al. 2017). Ante-hoc method, also known as the intrinsic method, often uses glass-box approaches for intrinsic explanations of models (Holzinger et al. 2019). Ante-hoc methods use decision trees or rules to explain how a prediction has been made through the model parameters. The models generated using ante-hoc methods are transparent and self-explanatory models such as bayesian rules and tree-based models (Islam et al. 2022).

Post-hoc explanation methods are used for models that are not readily explainable by their design (Holzinger et al. 2017; Barredo Arrieta et al. 2020a). They explain the inner working and inference mechanism of an already developed model or a newly created model after completing its training process. Generally, post-hoc methods mimic the behaviour of an already developed model to an external explainable model (Islam et al. 2022). Grad-CAM, Layer-wise Relevance Propagation (LRP), LIME, and Saliency Maps (Adadi and Berrada 2018; Linardatos et al. 2021; Angelov et al. 2021) are the most common examples of post-hoc methods.

3.2.3 Based on Applicability

Based on the application of explanation on different models, explanation methods are further categorized into two categories: model specific and model agnostic. The model agnostic methods are applied to any model, whereas the model-specific methods are restricted to particular models (Islam et al. 2022). Model-specific explanations are intrinsic methods where explanations are limited to a specific class of model. These methods aim to bring transparency to a particular type of model.

Model agnostic methods are used to interpret already developed models (Failed 2019). The advantage of these methods is that they do not impact the model's performance because they are independent of the inner working of the model (Linardatos

et al. 2021; Dosilovic et al. 2018). These methods have been often used due to their flexibility in the architecture of a model. Model agnostic methods also provide post-hoc explanations. LIME, LRP, and SHAP are popular examples of model-agnostic post-hoc explainers (Alicioglu and Sun 2022; Zhang et al. 2022). Following are some post-hoc model agnostic methods to achieve an understanding of a model.

(i) Attribution method: The attribution method analyzes the sensitivity of how the output is influenced by its input and/or weight perturbations. In other words, it measures the importance of each attribute or feature to the prediction. Sensitivity analysis, LRP, and feature importance are examples of the attribution method (Adadi and Berrada 2018; Chakraborty et al. 2018).
(ii) Visualization method: Visualization method explores the pattern hidden inside a learned model by visualizing its representations (Ras et al. 2018; Kim et al. 2019; Rajaraman et al. 2018). These techniques are essentially used for supervised learning models. Surrogate models, partial dependence plots and individual conditional expectations are examples of Visualization methods. Partial dependence plots are used to identify relationships between a set of features with the model outcome.
(iii) Knowledge extraction: The knowledge extraction method provides a comprehensible description of the knowledge by approximating the decision-making process using the input and output of the given model. Either rule extraction or model distillation can be used to gain insight into the model.

3.2.4 Based on Explanation Level

The explanations of XAI models are either analyzed by a machine (i.e. bot) or presented to a human. In machine-interpretable methods, users can mathematically analyze algorithmic mechanisms of predictions made by the system. This type of explanation is known as machine-to-machine explainability. Machine reasoning explainability uses techniques like compositionality, computational argumentation, and iterative contrastive explanations (Inam et al. 2021).

Despite the advantages of machine-interpretable methods, their explanations fail to generate human-understandable models (Confalonieri et al. 2021; Angelov et al. 2021). Human interpretable methods use linguistic variables and symbols to provide user-centric explanations of how a decision has been made (Holzinger et al. 2017; Chakraborty et al. 2018; Abdul et al. 2018). Rationalization is a form of explanation that attempts to explain a model prediction based on how a human would explain it. AI rationalization closely resembles explanations that are most likely given by a human (Ehsan et al. 2018).

3.3 Forms of Explanation

This section presents the proposed classification of different forms of explanation suitable for XAI. Many types of explanations are generated and used to explain the predictions made by machine learning and deep learning models. A broader range of end-users, such as naive users, data scientists and domain experts with different perspectives, demand the explanation in their understandable terms. Hence, selecting the best form of explanation suitable for a given model is a prime concern. Our literature study identified four different forms of explanations commonly used for deducing a decision (Ibrahim et al. 2019; Islam et al. 2022). Figure 3.5 illustrates these four forms of explanations: (1) analytical explanation, (2) visual explanation, (3) rule-based explanation, and (4) textual explanation. The analytical explanations are appropriate for domain experts and data scientists. Visual and rule-based explanations are found to be suitable for naive users. The textual explanations are presented in natural language for the general users.

3.3.1 Analytical Explanation

Analytical explanations are generated by measuring the contribution of the input features to the model's outcome. They are represented by various numeric metrics such as saliency, causal importance, feature importance, features confidence score, and mutual importance (Islam et al. 2022). Domain experts mostly utilize them to view and explore the data concerning their feature importance. They are suitable for post-hoc methods adopted for already developed models. Confident itemsets explanations presented in Moradi and Samwald (2021) are among the best examples of analytical explanation. The confident itemsets explanations utilize confidence score to find confident itemsets by considering each word as an item in the text record. Figure 3.6 illustrates an example of confident itemsets explanation for a text record "Where is mile high stadium?" from the TREC question classification dataset (Moradi and Samwald 2021).

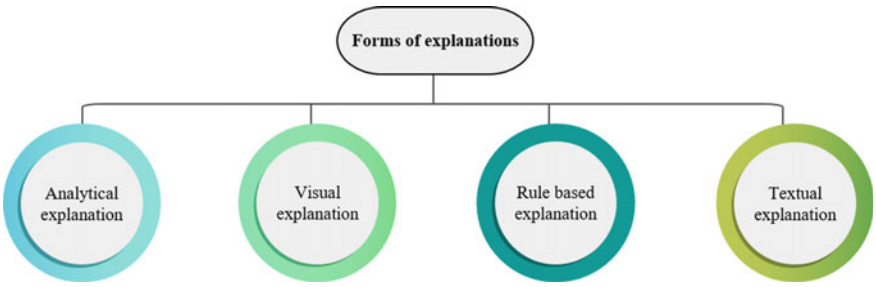


Fig. 3.5 Different forms of explanation useful for XAI

Fig. 3.6 Example of analytical explanation

Text record: Where is mile high stadium? **Prediction**: *LOC: other*			
Explanation using confident itemsets explanations: Minimum confidence threshold: 0.6			
Class: LOC: other **Score:** 2.554		**Class:** NUM: count **Score:** 0.666	
Itemset	**Confidence**	**Itemset**	**Confidence**
<where>	0.888	<mile>	0.666
<stadium>	0.666		
<where>, <stadium>	1.0		

3.3.2 Visual Explanation

Visual explanations are one of the most commonly used explanation forms for computer vision tasks. Visual explanations use visualization techniques such as class activation maps, gradient-based class activation maps (Abdul et al. 2018; Guidotti et al. 2018; Kim et al. 2018; Yasaka and Abe 2018) and attention maps to explain the model's prediction (Alicioglu and Sun 2022; Angelov et al. 2021). In these methods, the saliency heatmaps (Mohseni et al. 2021; Alicioglu and Sun 2022; Islam et al. 2022) using visual elements are generated to specify important regions of the input image. Visual explanations are post-hoc methods used for both the local and global explanations. Naive users can easily interpret visual explanations, which contain charts, trend lines, etc. An example of a visual explanation proposed by the authors of Sun et al. (2020) for fault diagnostics in an industrial machine is illustrated in Fig. 3.7. The bottom side nut of the machine is the most vibrating region because it is the most significant part of the fault diagnosis of the water pump.

Class Activation Map (CAM) is an explanation method suitable for convolutional neural networks with a global average pooling (Zhou et al. 2016). It highlights the discriminative region of the input image to identify the class of the image. It uses activation maps of the last convolutional layer to train linear classifiers for each class for final class estimation. The image regions that play an important role in prediction are identified by projecting the weights of the output layer onto the convolutional feature map. CAM requires retraining linear classifiers, one for each class, making this method time-consuming.

The authors of Selvaraju et al. (2016) have proposed Gradient-weighted CAM (Grad-CAM) to reduce the time complexity of CAM. Grad-CAM is the general form of CAM suitable for any convolutional neural network architecture. It uses the gradients of the target feature, flowing into the final convolutional layer to visualize the class activation maps. Moreover, it highlights the important regions in the input

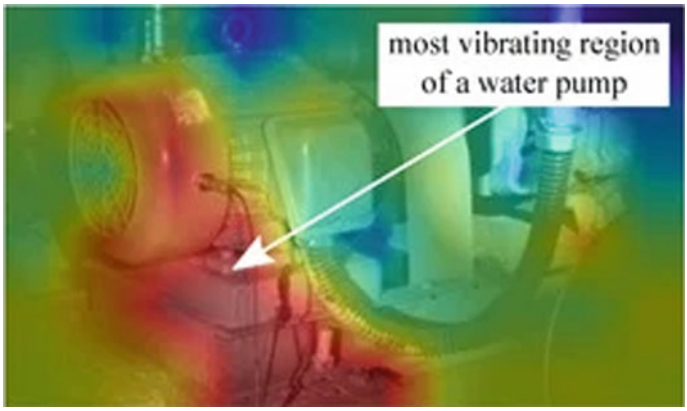


Fig. 3.7 Example of visual explanation (Sun et al. 2020)

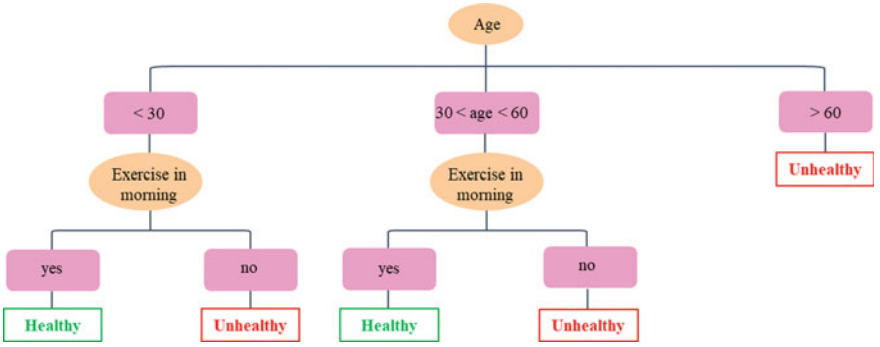
image for the selected feature. However, Grad-CAM fails in the localization of objects with multiple occurrences of the same class. The author of Selvaraju et al. (2016) also proposed a variation of Grad-CAM known as Guided Grad-CAM to obtain fine-grained pixel-scale representation. Guided Grad-CAM upsamples and fuses class-specific saliency map with the visualizations generated by guided backpropagation.

Grad-CAM++ (Chattopadhay et al. 2018) is an extension of the Grad-CAM that provides better visual explanations for the convolutional neural network. It enhances the object localization capability by extending multiple object instances in a single image. It utilizes second-order gradients of feature maps from the last convolutional layer. Additionally, it uses a specific class score as a weight to generate a visual explanation for the corresponding class label. The importance of each pixel is captured separately in the gradient feature map by assigning different weights to each pixel. Grad-CAM++ is more suitable for multi-label classification problems.

3.3.3 Rule-Based Explanation

The rule-based explanations are the simplest form of explanation. They describe the model inference mechanism using a set of IF–THEN rules or a tree (Islam et al. 2022). An example of a rule-based explanation along with a decision tree is illustrated in Fig. 3.8. In this example, two significant attributes, *age* and *exercise in the morning,* are considered to predict whether a person is healthy or unhealthy. Most rule-based explanations are ante-hoc methods that interpret models with a global scope. Ensemble learning and decision tree are popular examples of rule-based explanations. This type of explanation is commonly used to develop recommendation systems for naive users.

Adaptive Neuro Fuzzy Inference System (ANFIS) (Sagir and Sathasivam 2017; Keneni et al. 2019; Mehrdad Aghamohammadi and Madan 2019) is another popular



IF-THEN rules

Is a person healthy?

IF age < 30 AND Exercise in morning = yes, THEN Healthy.

IF age < 30 AND Exercise in morning = no, THEN Unhealthy.

IF 30 < age < 60 AND Exercise in morning = yes, THEN Healthy.

IF 30 < age < 60 AND Exercise in morning = no, THEN Unhealthy.

IF age > 60, THEN Unhealthy.

Fig. 3.8 Example of rule-based explanation

rule-based inference system that combines fuzzy inference rules with a neural network. Fuzzy Inference Systems (FIS) is a mapping method of a given input to output using the theory of fuzzy sets. There are two broad categories of FIS, namely the Mamdani fuzzy system and the Sugeno fuzzy system (Keneni et al. 2019). The general structure of the Mamdani fuzzy system is as follows:

IF p is A and q is B, THEN z is C;

where p and q are linguistic input variables and Z is the linguistic output.

Similarly, in the case of the Sugeno fuzzy system, the rules format is as follows:

IF p is A and q is B, THEN z is f(p,q);

where p and q are fuzzy sets on a universe of discourse and z is an output in the form of mathematical function f(p,q).

3.3.4 Textual Explanation

The textual explanations present the model prediction by learning text explanations as sets of words denoting the features that influence the model prediction (Bennetot

et al. 2019). They use Natural Language Processing (NLP) techniques to describe the model prediction in natural language. Moreover, they utilize numerous methods for generating symbols that represent the inner working of a model (Barredo Arrieta et al. 2020b). However, textual explanations are the least common among all forms of explanations due to their high computational requirement for NLP tasks. Generally, textual explanations are generated for individual prediction for precise and specific results. They are suitable for the general user. They are popular in applications like interactive question-answering systems (Mohseni et al. 2021). An example of textual explanation is given below based on Fig. 3.8.

Example of textual explanation: "*The person is classified as 'unhealthy' RATHER THAN 'healthy' because person age is more than 30 and no exercise in morning*".

The textual explanations are based on either factors or features that influence the model prediction or representative examples that support the prediction. If explanations are constructed for humans, they should be contrastive or counterfactual (Stepin et al. 2021; Zucco et al. 2018). Several researchers emphasize that good explanations are contrastive that explain the "Why", "Why not", and "What-if" of an AI-based system. The contrastive explanation is an effective method for mental model formation. Moreover, contrastive explanations improve the understanding without providing a full causal analysis (Kim et al. 2016). The counterfactual explanations are used to explain predictions of individual instances (Myers et al. 2020). A counterfactual explanation is a human-friendly explanation that describes a causal situation in the form: If an event "P" had not occurred, "Q" would not have occurred. Additionally, the counterfactual explanation identifies external factors that affect the model output.

3.4 Frameworks for Model Interpretability and Explanation

Researchers have designed several state-of-the-art frameworks to develop interpretable machine learning models. This section discusses the six promising XAI frameworks developed in python. These six frameworks are selected based on their explanation generation capability, application, and success on standard AI systems.

3.4.1 Explain like I'm 5

Explain Like I'm 5 (ELI5) is a python framework that helps to debug machine learning and deep learning models. ELI5 is one of the simple frameworks that finds the importance of each feature to the output for understanding the inner working of a model. However, its explanation is limited to parametric linear models and decision tree-based models. ELI5 provides two major functions: eli5.show_weights() function to inspect model parameters and eli5.show_prediction() function to inspect an individual prediction and determine why the model predicts this.

3.4.2 Skater

Skater is another popular open-source python framework designed to understand the inner workings of the black box model. The Skater framework evaluates and explains predictive models based on independent (input) and dependent (target) variables in a post-hoc manner (Linardatos et al. 2021). Moreover, it enables better model insight and debug options by keeping humans in the loop. The Skater framework supports a variety of models which can be explained either at the local or global level. In addition, it also supports object-oriented and functional programming paradigms to provide better scalability and parallelism.

3.4.3 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) (Palatnik de Sousa et al. 2969) is a surrogate-based explanation method that explains a model's prediction by fitting a local surrogate model whose predictions are easy to explain. LIME explains each prediction to understand how the black box model works in that local fidelity. It observes the effects of individual predictions by perturbing the original data. Although LIME is popular and simple, random perturbation of LIME results in unstable interpretation results. The authors of Zafar and Khan (2021) have proposed a deterministic version of LIME known as DLIME to deal with this limitation. Unlike LIME, DLIME uses hierarchical clustering to group the data and k-nearest neighbours to find the cluster where the given instance belongs.

3.4.4 Shapley Additive Explanations

Shapley Additive Explanations (SHAP) is used to explain an instance's prediction by computing each feature's contribution to the prediction. Shapely values are used to

identify the effect of individual features on the model outcome (Failed 2019). Shapley values provide explanations by assigning a value called weight to each feature for a particular prediction. SHAP can guarantee consistency and local accuracy because of its thorough approach to considering all possible predictions, such as using all possible combinations of inputs. SHAP is available in two variants (i) KernelSHAP (Lundberg and Lee 2017) and (ii) TreeSHAP (Linardatos et al. 2021). Kernel SHAP is a model agnostic method based on LIME concepts and Shapley values. The major drawback of Shapley values is their computational complexity. Tree SHAP computes exact SHAP values for decision trees based models. Asymmetric Shapley Values (ASV) is a SHAP variation that incorporates a causal graph of the cause-effect relationship between variables in the model explanation process. Unlike SHAP, where shapely values are symmetrical, ASV uses asymmetric shapely values. The model fairness analysis is a major application of ASV values because it can capture the indirect effects of the variable on a model.

3.4.5 Anchors

The anchors method explains each model prediction by finding IF–THEN rules called anchors (Ribeiro et al. 2018). An anchor explanation is a rule framed using input and the model prediction at a local level. Anchors are high precision and model-agnostic explanation methods that use reinforcement learning to construct rules without knowledge about the model. Moreover, they can explain nonlinear models because they work on feature predicates. The key limitation of Anchors is that they only support textual and tabular data. They can produce explanations in the form of tabular data and text depending on the application domain.

3.4.6 Deep Learning Important Features

Deep Learning Important FeaTures (DeepLIFT) is another popular explanation framework for the deep neural network. It calculates the importance score of each feature. It explains the model by computing the difference in model output from some reference output based on the difference of the input from some reference input. The reference input represents some default input (Shrikumar et al. 2017). Moreover, DeepLIFT provides different considerations for positive and negative contributions.

The aforementioned XAI frameworks are critically examined and their comparative analysis is presented in Table 3.1. All six frameworks support model agnostic post-hoc explanation at the local level. In addition to local explanation, SKATER and SHAP support the global explanation.

Table 3.1 Comparison of explainable framework

Explainable framework	Method			Form of explanation supported
	Local/global	Post-hoc/atte-hoc	Model agnostic/model specific	
ELI5	Local	Post-hoc	Model agnostic	Textual
SKATER	Local and global	Post-hoc	Model agnostic	Textual
LIME	Local	Post-hoc	Model agnostic	Textual and visual
SHAP	Local and global	Post-hoc	Model agnostic	Textual and visual
ANCHORS	Local	Post-hoc	Model agnostic	Textual and tabular
DeepLIFT	Local	Post-hoc	Model agnostic	Textual

3.5 Conclusion and Future Directions

While Artificial Intelligence has a long history, Explainable AI, also known as XAI, is a relatively new interdisciplinary research field. XAI research has been growing rapidly due to the increasing demand for diverse frameworks and methods to produce interpretable and understandable results from black box models. Many methods of Explainable AI have been proposed in the literature to understand the inner working of black box models and their predictions. This chapter provides a selective and summarized overview of various methods and forms of explanation for XAI. A taxonomy of methods suitable for XAI followed by a classification of explanation forms has also been proposed. Moreover, a comparative analysis of six popular XAI frameworks has been presented to help the research community select a suitable explainable framework.

XAI brings significant benefits to many application domains relying on AI-based systems. The potential application domains such as healthcare, finance, military, criminal justice and transportation require more attention to the human's role in existing explainability methods because the consequence of decisions can be dangerous. It has been observed that little attention has been given to combining different interpretability methods to achieve easy to understand and human-centric explanations. Hence, developing general and interactive explanations methods with emerging NLP techniques will be a new research direction in XAI.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of Conference on Human Factors in Computing Systems, pp. 1–18 (2018). https://doi.org/10.1145/3173574.3174156

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

Aghamohammadi, M., Madan, M., Hong, J.K., Watson, I.: Predicting heart attack through explainable artificial intelligence. In: International Conference on Computational Science—ICCS 2019, vol. 1, pp. 633–645 (2019). https://doi.org/10.1007/978-3-030-22741-8

Alicioglu, G., Sun, B.: A survey of visual analytics for Explainable Artificial Intelligence methods. Comput. Graph. **102**, 502–520 (2022). https://doi.org/10.1016/j.cag.2021.09.002

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **11**(5), 1–13 (2021). https://doi.org/10.1002/widm.1424

Barredo Arrieta, A. et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion (2020a). https://doi.org/10.1016/j.inffus.2019.12.012

Barredo Arrieta, A. et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020b). https://doi.org/10.1016/j.inffus.2019.12.012

Bennetot, A., Laurent, J.L., Chatila, R., Díaz-Rodríguez, N.: Towards explainable neural-symbolic visual reasoning, *arXiv Learn.* (2019)

Chakraborty S. et al.: Interpretability of deep learning models: a survey of results (2018). https://doi.org/10.1109/UIC-ATC.2017.8397411

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097

Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable Artificial Intelligence, *Wiley Interdiscip.* Rev. Data Min. Knowl. Discov. **11**(1), 1–21 (2021). https://doi.org/10.1002/widm.1391

Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. In: CEUR Workshop Proceedings, vol. 2071 (2018)

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning, *arXiv Prepr. arXiv1702.08608*, no. Ml, pp. 1–13 (2017). http://arxiv.org/abs/1702.08608

Dosilovic, F.K., Brcic, M., Hlupic, N.: Explainable artificial intelligence: a survey. In: Proceedings of 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, pp. 210–215 (2018). https://doi.org/10.23919/MIPRO.2018.8400040

Ehsan, U., Harrison, B., Chan, L., Riedl, M.O.: Rationalization: a neural machine translation approach to generating natural language explanations. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 81–87 (2018). https://doi.org/10.1145/3278721.3278736

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of 2018 IEEE 5th International Conference on Data Science Advanced Analytics DSAA 2018, pp. 80–89, (2019). https://doi.org/10.1109/DSAA.2018.00018.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–45 (2018). https://doi.org/10.1145/3236009

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(4), 1–13 (2019). https://doi.org/10.1002/widm.1312

Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? No. Ml, pp. 1–28 (2017). http://arxiv.org/abs/1712.09923

Ibrahim, M., Louie, M., Modarres, C., Paisley, J.: Global explanations of neural networks: mapping the landscape of predictions. *CoRR arXiv1902.02384*, pp. 1–10 (2019). http://arxiv.org/abs/1902.02384

Inam, R., Terra, A., Mujumdar, A., Fersman, E., Feljan, A.V.: Explainable AI—how humans can trust AI. Ericsson, no. April, pp. 1–22, 2021. https://www.ericsson.com/En/Reports-and-Papers/White-Papers/Explainable-Ai--How-Humans-Can-Trust-Ai

Islam, S.R., Eberle, W., Ghafoor, S.K., Ahmed, M.: Explainable artificial intelligence approaches: a survey. CoRR, pp. 1–14 (2021). http://arxiv.org/abs/2101.09429

Islam, M.R., Ahmed, M.U., Barua, S., Begum, S.: A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. **12**(3) (2022). https://doi.org/10.3390/app12031353

Keneni, B.M. et al.: Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access, vol. 7, no. c, pp. 17001–17016 (2019). https://doi.org/10.1109/ACCESS.2019.2893141

Kim, B., Khanna, R., Koyejo, O.: Examples are not enough, learn to criticize! Criticism for interpretability. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 2288–2296 (2016)

Kim, I., Rajaraman, S., Antani, S.: Visual interpretation of convolutional neural network predictions in classifying medical image modalities. Diagnostics (2019). https://doi.org/10.3390/diagnostics9020038

Kim, B. et al.: Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: 35th International Conference on Machine Learning, ICML 2018, vol. 6, pp. 4186–4195 (2018)

Krajna, A., Brcic, M.: Explainable artificial intelligence : an updated perspective explainable artificial intelligence : an updated perspective. (2022)

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 1–45 (2021). https://doi.org/10.3390/e23010018

Lundberg, S., Lee, S.-I.: A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), May 2017, pp. 1–10. http://arxiv.org/abs/1705.07874. Accessed 30 Aug 2019

Messalas, A., Kanellopoulos, Y., Makris, C.: Model-agnostic interpretability with Shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), July 2019, pp. 1–7. https://doi.org/10.1109/IISA.2019.8900669

Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. **11**(3–4), 1–45 (2021). https://doi.org/10.1145/3387166

Moradi, M., Samwald, M.: Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst. Appl. **165**, 113941 (2021). https://doi.org/10.1016/j.eswa.2020.113941

Myers, C.M., Freed, E., Pardo, L.F.L., Furqan, A., Risi, S., Zhu, J.: Revealing neural network bias to non-experts through interactive counterfactual examples (2020). http://arxiv.org/abs/2001.02271

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., Costa da Silva, E.: Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors **19**(2969), 1–18 (2019). https://doi.org/10.3390/s19132969

Rajaraman, S., Candemir, S., Kim, I., Thoma, G., Antani, S.: Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl. Sci. (2018). https://doi.org/10.3390/app8101715

Ras, G., Van Gerven, M., Haselager, P.: Explanation methods in deep learning: users, values, concerns and challenges, pp. 19–36 (2018). https://doi.org/10.1007/978-3-319-98131-4_2

Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: 32nd Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2018, vol. 32, no. 1, pp. 1527–1535 (2018)

Sagir, A.M., Sathasivam, S.: A novel adaptive neuro fuzzy inference system based classification model for heart disease prediction. Pertanika J. Sci. Technol. **25**(1), 43–56 (2017)

Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization, *CoRR*, vol. abs/1610.0 (2016). http://arxiv.org/abs/1610.02391

Shi Zhang, Q., Chun Zhu, S.: Visual interpretability for deep learning: a survey. Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018). https://doi.org/10.1631/FITEE.1700808

Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences, *CoRR*, vol. abs/1704.0 (2017). http://arxiv.org/abs/1704.02685

Stepin, I., Alonso, J.M., Catala, A., Pereira-Farina, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021). https://doi.org/10.1109/ACCESS.2021.3051315

Sun, K.H., Huh, H., Tama, B.A., Lee, S.Y., Jung, J.H., Lee, S.: Vision-based fault diagnostics using explainable deep learning with class activation maps. IEEE Access **8**, 129169–129179 (2020). https://doi.org/10.1109/ACCESS.2020.3009852

Yasaka, K., Abe, O.: Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med. **15**(11), 1–4 (2018). https://doi.org/10.1371/journal.pmed.1002707

Zafar, M.R., Khan, N.: Deterministic local interpretable model-agnostic explanations for stable explainability. Mach. Learn. Knowl. Extr. **3**(3), 525–541 (2021). https://doi.org/10.3390/make3030027

Zhang, Y., Weng, Y., Lund, J.: Applications of explainable Artificial Intelligence in diagnosis and surgery. Diagnostics **12**(2) (2022). https://doi.org/10.3390/diagnostics12020237

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

Zucco, C., Liang, H., Di Fatta, G., Cannataro, M.: Explainable sentiment analysis with applications in medicine. In: Proceedings of—2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, pp. 1740–1747 (2019). https://doi.org/10.1109/BIBM.2018.8621359

Chapter 4
Methods and Metrics for Explaining Artificial Intelligence Models: A Review

Puja Banerjee and Rajesh P. Barnwal

Abstract Deep learning (DL) solutions have been facing the long-standing problem of making the Explainable Artificial Intelligence (XAI) an integral part of the machine learning pipeline. In recent times, multiple deep learning approaches have been established for solving the enhanced complications aroused due to high predictive capacity. Though DL models demonstrate exceptionally high accuracy but the same comes with computationally complex and difficult to interpret black-box architectures. Several efforts are being made to develop the methods for making such high-precision black-box models explainable so that the trustworthiness and reliability of such models can be established. The chapter provides an overview of XAI, different methods of XAI, and metrics associated with those methods. Further, the chapter also discusses the motivational factors behind XAI, its applications, and its taxonomy. For clarity on the XAI implementation stage, Pre-model, In-model, and Post-model explainability are elaborated along with the model-agnostic and model-specific techniques. The chapter concludes with a brief discussion on a simple use-case of implementing the XAI method in a real-life problem followed by enumerating possible future research directions.

4.1 Introduction

Most effective Artificial Intelligence (AI) techniques are still black-box. The users and AI professionals are unable to interpret and understand the reason behind the decisions of those techniques. The absence of such a transparent system can result

P. Banerjee
Academy of Scientific and Innovative Research, Ghaziabad, India
e-mail: puja.cmeri20a@acsir.res.in

R. P. Barnwal (✉)
AI & IoT Lab, IT Group, CSIR-Central Mechanical Engineering Research Institute, Durgapur, India
e-mail: r_barnwal@cmeri.res.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_4

in severe consequences in the areas of health diagnosis, finance management, military, self-driving vehicles, etc. Thus, methods for explaining AI decision-making processes have experienced a huge increase of attention from the research community to its application domain. The methods of deep learning have advanced the state-of-the-art of Artificial Intelligence to the next level. However, even with such extraordinary improvements, the absence of explainations in the deep learning predictions and the lack of control over the internal functioning of model development act as major drawbacks. Thus efforts are focused on making deep learning models manageable and interpretable to the end-users. The International market of Explainable AI is sub-divided on the grounds of contributing better services and solutions. Depending on its domain of applications, the market of XAI is classified as drug discovery, fraud detection, advertising, recommendation engines, computer vision (e.g., in case of classifying images, visual question answering, image captioning), security, natural language processing (classification of text, sentiment analysis) as well as supply chain management, etc. Moreover, based on user industry, the XAI international market is classified into telecommunication, medicine, healthcare, retail, public sectors, logistics, entertainment, military, defense, etc. Based on region, the market of XAI is divided into North America, Europe, and the Asia Pacific. In the report published by Next Move Strategy Consulting (Next Move Strategy Consulting (NMSC) 2022), the international market of XAI is expected to generate an amount of $21.78 billion (USD) by 2030.

Deep Learning is a common and most contemporary subset of the wide Machine Learning which field of study utilizes Artificial Neural Networks(ANN) (Goodfellow et al. 2016) for prediction and inference. Though the advancement of ANN started between the 1940s and 1960s, researchers did not find sufficient computational power and data for testing the limits of the developed methods. Deep Learning uses deeper networks (more number of model layers) for prediction. The deep learning structure enabled non-linear modeling of the data for prediction, as opposed to the traditional paradigms that follow linear modeling for prediction. The term “deep”, in these methods is derived from the idea of using multiple layers in learning networks. The layers comprise of input and an output layer and multiple hidden layers between these two layers. On the other hand, shallow networks are networks that have up to two hidden layers. Deep Learning can achieve higher accuracy because of the multiple hidden layers which can better model higher levels of abstractions in the data. These models can also better leverage the data in big data sets for prediction. Since these models can efficiently leverage the information in big data sets (often containing millions of data points), the chance of an incorrect prediction due to the unavailability of a data point critical to the prediction gets significantly reduced. This mitigates the risk of inaccurate predictions and therefore leads to an overall improvement in the prediction accuracy.

4.1.1 *Bringing Explainability to AI Decision—Need for Explainable AI*

XAI is a field of research that makes AI-based prediction systems transparent and understandable to end-users. This term was first coined in the year 2004 (Adadi and Berrada 2018) for describing the capability of a system for explaining the behavior of entities managed by the AI models in simulating applications in games (Van Lent et al. 2004). Although the concept is still comparatively new, the difficulty of explaining the model has survived since the middle of the 1970s when researchers started studying explainations for the expert AI systems. Moreover, the advancement toward solving such problems has decelerated as AI reached a point of inflection with the extraordinary progress in Machine Learning techniques. Since then the center of research in the Artificial Intelligence domain has shifted towards model development and algorithm implementation. AI and Machine learning generally demonstrates practical success in many different application domains. Some of the popular applications sectors include autonomous vehicles and drones, speech recognition systems, face recognition, navigation system, social networking, etc. But the central problem of such models is that these models are not transparent and thus black-box. This means that even if the fundamental principles of the mathematical are well understood, they lack explicit interpretive knowledge representation. This results in increasing ethical, legal, and privacy issues. This in turn results in applying black-box approaches in personal, defense, and business operations that could be more difficult and untrustworthy due to their adverse implications. Thus it necessitates the need for explainable AI systems that can introduce more transparency and clarity. Based on the existing literature, the need for explaining the AI systems may stem from the reasons that the user should get a trustworthy, dependable, compliant, effective, fair, and robust decision as shown in Fig. 4.1.

Various factors for which explainability is required in AI models are as follows:

- **Explain to Justify**: Explaining an AI decision furnishes the information required to be justified, generally when the decision is unexpectedly made. This makes sure that an auditable and provable way is there for defending the algorithmic decisions for being ethical and fair, which would result in introducing trust in the AI decisions.
- **Explain to Control**: XAI is important not because of justifying any AI decision but also helps in preventing incorrect AI decisions. Thus, a better understanding of the behavior of AI systems provides significant visibility of unpredictable flaws and vulnerabilities. This helps in the rapid identification and correction of errors in critical situations, which in turn enables enhanced control over the system.
- **Explain to Improve**: Another cause for developing XAI models is to make continuous improvements. An AI model which can be understood and explained is the one to be easily improved. This is where the end-users get to know the reason why an AI system produces such specific outputs. This shows the possibility that how XAI could act as the base for ongoing improvement of man and machine interaction.

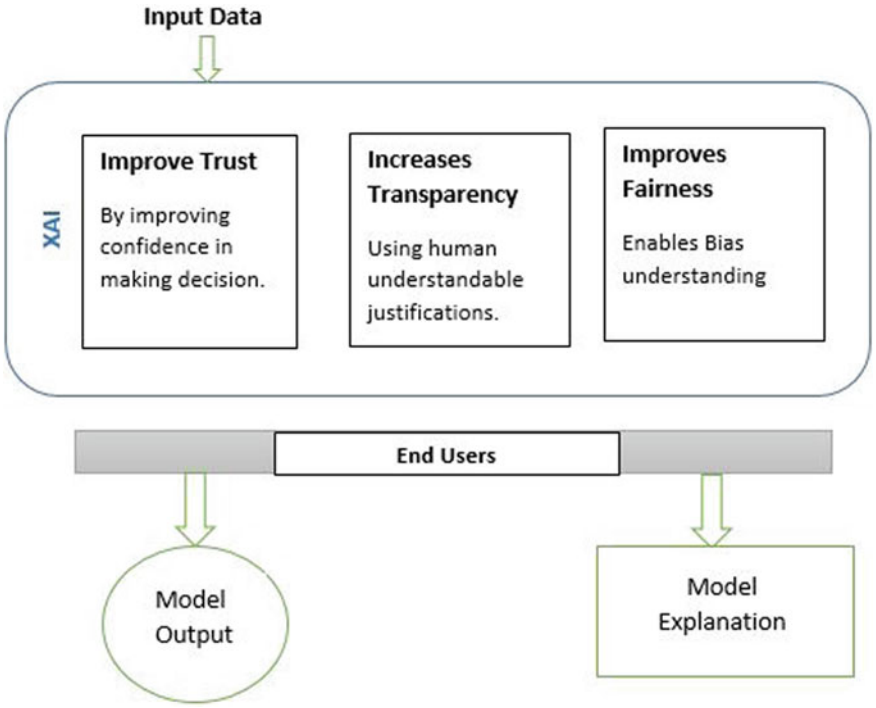


Fig. 4.1 Factors making XAI important

The key equivalent terminologies used in the literature for XAI and their interpretation are mentioned in Table 4.1.

Interpretability and explainability are approached via two categories, i.e, the transparency-based approach and the post-hoc-based approach. Transparency and estimated performance are contrary objectives and thus there should be a trade-off while developing an AI model. When a system is already robust and self-contained, introducing transparency is not necessary. But, when it is a part of another complicated system, then introducing transparency is good for debugging ability. Post-hoc interpretability takes out information from a trained model and does not specifically depend on the working of the model. An advantage of this type of approach is that it impacts the model's performance, which is treated as a black box.

4.2 Taxonomy of Explaining AI Decisions

Interests of researchers in the domain of Explainable AI are emerging. Previous works were concentrated on providing explainations of any decision undertaken based on knowledge-based expert AI systems. The main reason for these kinds of interest

Table 4.1 Equivalent key terminology used for XAI

Key terms	Description
Black-box AI	The AI model is a black box when it does not reveal anything about the internal design and structure (Suman et al. 2010) of the system. Because it is difficult for black box models to provide suitable explainations, the problem related to these systems is known as the *black-box problem*
Interpretable AI	Interpretable systems are those where the users can not only visualize the parameters necessary for any prediction but can also understand how the input variable is mathematically connected to the outputs. Researchers use the terms *interpretability* and *explainability* (Koh and Liang 2017) interchangeably. Others used terms such as *comprehensibility* or *understandability* (Bojarski et al. 2017) to refer to the same issue, whereas the term *interpretable AI* is more preferred in the industry
Responsible AI	This term of XAI takes societal, moral, and ethical values into consideration. The pillars of Responsible AI are Accountability, Transparency, and Responsibility (Dignum 2017)
Third-wave AI	Recently, the term third wave in AI has also surfaced, where the system constructs an explainatory model for classifying a real phenomenon and provides reason to their tasks and situations

in research related to XAI aroused because of current advancements in the area of Artificial Intelligence, it could be applied in a large range and into different domains, XAI is used in solving problems related to improper and unethical use, bias in any AI decisions, lack of transparency in the developed models. In addition to it, the latest laws imposed by the government illustrated that research on XAI should be done to a greater extent. XAI helps the AI systems in uncovering the hidden bias in any kind of decision caused due to the internal parameters of an opaque model.

A thorough review of different approaches was presented by Mueller et al. (2019) in the year 2019 which presented numerous types of explainations in AI systems and divided it into three generations i.e, the first generation -which includes efficient systems from the 70s. This system attempts in expressing the internal process of a system by putting knowledge directly from experts. The example includes changing the rules into natural expressions. The second-generation systems-these are intelligent systems, which are made to provide cognitive support to the human-machine

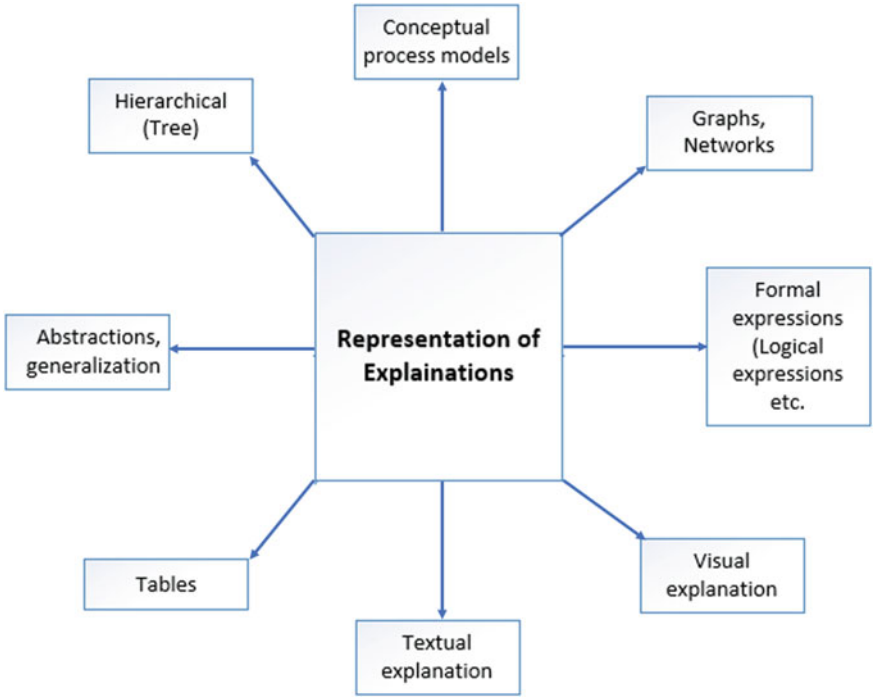


Fig. 4.2 Different representations for providing explainations for AI decisions

interaction and take the knowledge and reasoning capabilities of humans into consideration. For example, an interfacing system is being arranged in such a way that it complements the knowledge lacking by the users, and the third-generation systems, which utilizes techniques from the present era, started in the year 2015. Like first-gen systems, then disclose the internal working of the AI system. Although the systems are *black-box* in nature, it helps in clarifying their working. In addition to it, in recent days, researchers are using advanced computer technologies such as visualizing the data, and video animations, which have a huge potential in driving further research in XAI. Multiple ideas are being proposed for providing explainable decisions that aim in developing fair and trustable AI decisions.

Figure 4.2 illustrates different formats and representations of explaination. The concepts illustrated here show the *common ground* for research in the XAI domain. This is generated by properly reviewing and doing a thorough literature survey on AI.

For generating explaination, varieties of taxonomy are available, which include approaches that take notice of the type of explaination to be provided, a model which is needed to be explained, the scope for the explaination, or by combining all of these approaches. The taxonomy of Explainable Artificial Intelligence is shown in Fig. 4.3. Considering an example where XAI techniques have been applied the

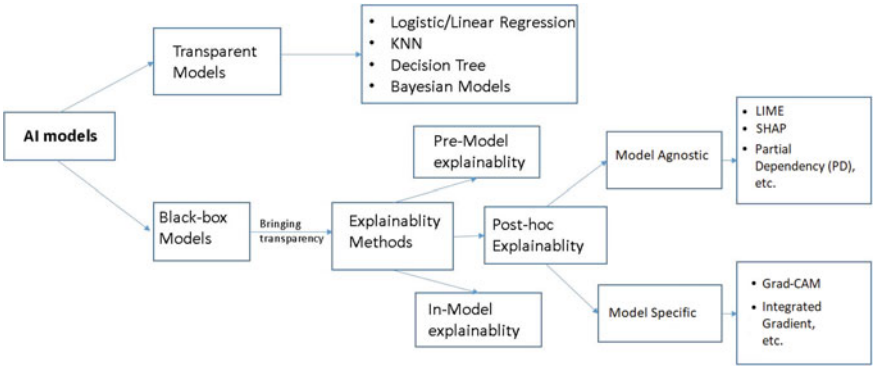


Fig. 4.3 Taxonomy of explainable artificial intelligence

methods for explaination are classified into an intrinsic and post-hoc method. In the intrinsic method, explainability is achieved by imposing constraints on the AI model. Post-hoc methods explain the model after its training. The post-hoc explainability method is further classified as Model-specific and Model-agnostic methods as well as local and global explainability methods. Moreover, Arya et al. (2019) showed a hierarchical relationship between different methods of explaination, the author also segregated explainations that are based on different used techniques and associated with different relations.

4.3 Methods of Explainable Artificial Intelligence

With improvements in hardware and a rise in data availability, the benefits of predictive performance increase using complex and opaque models. However, the primary issues that need to be addressed adequately are interpretability and explainability. With this kind of approach, we need to start our modeling with a trained black-box predictor and training data. While some methods deal with the issue of interpretability, others deal with explainability. The method is called model-agnostic when it operates on the inputs and outputs of the black-box model and is called model-specific when it uses the idiosyncrasies associated with some kind of representation.

These methods, in general, predict with an explaination, which is in the form of feature importance for a particular decision. Methods like layer-wise relevance propagation (Montavon et al. 2019) and sensitivity analysis (Christopher Frey and Patil 2002) have been presented to explain the predictions from DL models. Deep Taylor decomposition propagates explaination from Neural Network output to the input contribution. Model-agnostic methods are used to capture the degree of input influences on the system outputs. SHapley Additive exPlainations is a Local approximation method (Antwarg et al. 2019), which is used in explaining the prediction $f(y)$ for a single input y. SHAP is a framework that helps in estimating the number

of additive feature attributions which other multiple works follow in general. An explaination is generated by highlighting the relevant region of the input image.

The AI explaination is classified into three types:

- **Model-based explainations**: It represents those explainations that uses a model for explaining the original task (Krarup et al. 2019). In this type of explainations the task model explains itself or more interpretable is being generated to explain the task model.
- **Attribute-based explainations**: It generally measures the power of explaination of the input features and utilizes the ranks for explaining the task model (Jain et al. 2020). For instance, the feature importance explaination approach belongs to this class of explainations.
- **Example-based explainations**: This type of method selects instances from the testing and training data sets and can create new instances (van der Waa et al. 2021) for explaining the task model. For example, an instance is being selected which can be predicted well by the model as explainations, or by producing counterfactual examples for explainations.

Numerous explaination techniques have been described and evaluated which fall under following six broad approaches.

- **Feature relevance**: These approaches for explaining AI decisions focus on the inner functioning of the model and highlight features that best explain the output of the model.
- **Local explainations**: These approaches segment the solutions and provide explainations for smaller segments that are less complex. This tackles explainability by first dividing the solution and then generating explainations for less complex solution sub-spaces, which are relevant to the entire model. These types of explainations are given by those techniques which can only explain parts of the functionalities of the entire system, but not the complete whole.
- **Visualization**: These types of approaches allow end-users to visualize the behavior of the model, often by minimizing the complexity of the problems. These techniques of explainability aim to visualize the behavior of a model. Existing multiple visualization methods come with techniques that allow a simple human interpretable visualization approach. Visualizations can also be used with other kinds of methods for improving the understanding, and are considered the best way for explaining the complex interaction between the variables involved in the model to those users unfamiliar with AI modeling.
- **Explaination by example**: These approaches bring out specific representative data for explaining the overall behavior of the model.
- **Text explainations**: These types of approaches convert the explainations into natural language text. Text explaination generates symbols that represent the working of an AI model. These symbols can depict the algorithm rationale by using a semantic mapping technique that maps the model to the symbols.
- **Model simplification**: These approaches focus on building a new model that is less complex than a model that is to be explained.

4.3.1 Techniques of Explainable AI

- **Explainations by simplification**: These are the types of techniques that are used to explain an AI model. Because basic models are sometimes only representatives of particular parts in a model, this category includes local explainations of AI decisions. Nearly all the techniques, which are following this kind of path, working in simplifying the model, are based on some rule extraction methods. One of the familiar contributions to this kind of approach is the Local Interpretable Model Agnostic Explainations (LIME) (Ribeiro et al. 2016) method. This type of approach categorizes the explaination by simplifying it and provides local explainations. Along with LIME, there is another approach for the extraction of rules known as Genetic Rule Extraction. Even though it was not deliberated for extracting the rules from non-transparent models, the generic proposition of Genetic Rule Extraction is extended in accounting for explainability purposes. Explaining by Simplification is addressed from a separate perspective, where a technique for distilling and inspecting a black-box model is being presented. Within it, two major ideas are being illustrated: a technique for model distillation and comparison for auditing the risk of black-box scoring models; and a statistical test is being conducted for checking whether the auditing data is lacking the important features with which it was trained.
- **Feature relevance explainations**: These types of explaination methods are used for describing the working of an opaque model by ranking the feature and by calculating the influence, relevancy, and importance of each feature which results in predicting the output. A combination of propositions is being established inside this category, and they utilize different types of algorithmic approaches with a similar targeted goal. One of the important contributions to this is SHapley Additive exPlainations(SHAP) (Fidel et al. 2020). Researchers have introduced a method for calculating an importance score of the feature for every prediction with a set of necessary properties like consistency and local accuracy etc which is lacking by the antecedent. Other techniques for tackling the contribution of each feature for the prediction have an alliance with the game theory and local gradients.
- **Explaination by visualization**: This technique is used for achieving model agnostic explainations. Some of the works presented by Cortez et al. (2011) presented a detailed portfolio of the visualization method which helps in explaining a black box model made on some set of expanded techniques. Another set of visualization techniques presented by the same author (Cortez and Embrechts 2013) where three novel Stochastic-algebraic methods are Data-based Stochastic algebraic method, Monte-Carlo Stochastic algebraic method, and Cluster-based Stochastic algebraic method is being shown as one novel input measures. And finally, Goldstein et al. (2015) presented Individual Conditional Expectation plots as a mechanism for visualizing the estimating the model with the help of a supervised learning algorithm. Explaining the results visually is rare in the field of model-agnostic techniques. The design of these kinds of methods ensures that they can be easily applied to any kind of model without considering its inner structure, it creates

visualizations from the inputs and outputs of a non-transparent model, which is a difficult task. This is the reason nearly all the methods used for visualization fall into this type of categorical work together with the feature relevance technique, which provided the details that are finally presented to the end-user. Some of the important Explainability methods based on visualization techniques are as follows:

4.3.2 Stages of AI Explainability

The method of explainability could be classified in terms of the stages when this kind of method could be applied: before i.e., (pre-model), during the explaination i.e, (in-model), or after i.e., (post-model) when the AI model is built (Carvalho et al. 2019). Pre-modeling explainability methods are not dependent on the model, as because they are applied to the data itself. Different stages of Explainable AI are shown in Fig. 4.4. Pre-model explainability usually occurs before the model is selected, as because it is also important for exploring and having good knowledge about the data before the model is created. Pre-model explainability is, nearly related to data interpretability, which consists of explaining data analysis (Tukey et al. 1977) methods. Visualizing data is very critical in the case of pre-model explainability, which consists of a graphical representation of the data aiming to provide better knowledge . Model explainability deals with AI models which are inherently interpreting. Post-modeling explainability helps in improving interpretability after the model is built, thus it is referred to as post-hoc.

4.3.2.1 Pre-model Explainability Methods

Pre-model explainability is a collection of several strategies aimed at better comprehending the dataset used in model construction. Exploratory data analysis, dataset description standardization, explainable feature engineering, and dataset summarising methodologies are the four primary categories of pre-model explainability. These pre-model explainability methodologies are being discussed as follows:

- **Exploratory data analysis**: Exploratory data analysis aims to extract a summary of a dataset's primary properties. The mean, standard deviation, range, missing samples, and other statistical features of the dataset are frequently included in this summary. An exploratory data analysis task looks at the relative frequency of faulty and non-defective photographs which can show a problem with class imbalance, with considerably fewer defective photographs than non-defective ones. After identifying the problem in the training dataset, a variety of methods can be used to alleviate the problem and improve the classifier's performance. However, relying solely on statistical features is insufficient while analyzing data. Real-world

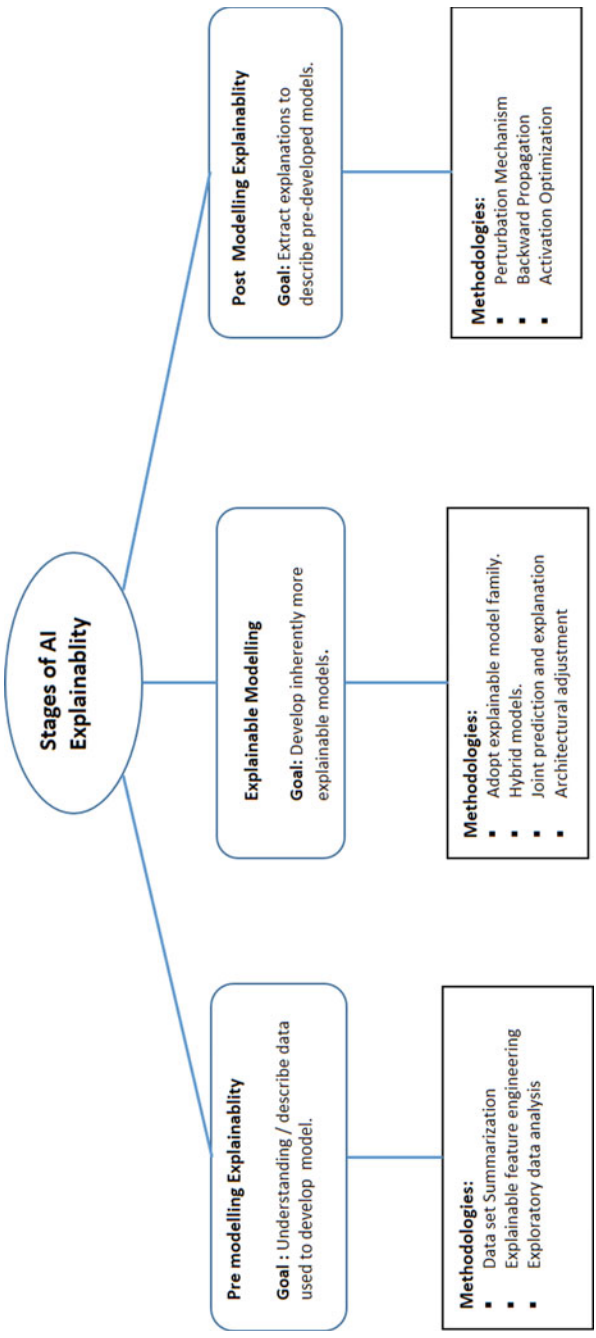


Fig. 4.4 Stages of explainable artificial intelligence (Khaleghi 2022)

datasets are generally complicated and multidimensional, i.e., they contain a significant number of features.

- **Dataset description standardization**: In most cases, datasets are released without adequate documentation. Standardization might help alleviate concerns like systematic bias in AI models and data exploitation by ensuring effective communication between dataset creators and users.
- **Explainable feature engineering**: Feature selection is a technique for limiting the input variable to a model by using only relevant data and removing noise. It is the technique of selecting suitable characteristics for a machine learning model automatically based on the type of challenge. Domain-specific and model-based feature engineering are the two basic ways to achieve explainable feature engineering. To extract and/or identify features, domain-specific techniques rely on domain expert knowledge and ideas obtained from exploratory data analysis. Model-based feature engineering, on the other hand, employs a variety of mathematical models to identify a dataset's underlying structure.

4.3.2.2 Explainable Modeling Methodology

Explainable modeling is also known as In-model Explainability. This methodology is being applied during the explaination of the model. Models which are transparent in nature provide some kind of interpretability themselves. These models are approached based on the domain where they become interpretable, like, algorithmic transparency and decomposability. A model becomes transparent if it can be easily understandable. Some of the transparent models are shown in this section.

- **Logistic and Linear Regression**: It is a type of classification model which is used for predicting a binary variable. However, linear regression would be analog when the dependent variable becomes continuous. Such kind of model would create dependencies between the predicted variables and predictors thus restricting the flexible fitting of the data (Rao 2003). The stiffness exhibited by the model leads to the model being maintained under the category of transparent methods. Although logistic and linear regression entirely meets the properties of transparent models, these methods might also need explainability techniques (including visualization techniques), especially when a model needs to be explained to non-expert audiences (Hellevik 2009). This kind of model are widely used for a long time, which has motivated researchers to create ways for explaining the model predictions to non-expert users.
- **Decision Trees**: This is another kind of model that could easily fulfill all the restrains of transparency. This is a hierarchical structure for making decisions applied to regression and classification problems. It is a simulatable model, whose properties can either make them algorithmically transparent or decomposable. Decision trees are always been categorized under the different types of transparent models (Schetinin et al. 2007). The applications of these models have been closely related to decision-making contexts. This provides an answer to why their complexity and understandability have been considered a crucial matter (Tan et al.

2020) always. A simulatable decision tree is a type of decision tree that can be managed by a human user which indicates that the model size is relatively small and the features and their corresponding meaning are easily understandable. Any increment in model size converts the model into a decomposable one because its size restricts its full simulation. Moreover, a further increase in this size by using the complicated feature is supposed to inject more algorithmic transparency into the model by making it lose the previous characteristics.

- **K-Nearest Neighbor**: The predictions made by the k-nearest neighbor are explained by the data points of the k-neighbor which are neighborhood points whose features were averaged for making the prediction (Peterson 2009). Visualizing the individual cluster which contains alike instances provides an interpretation of why an instance belongs to a particular group. Our analysis suggests that because of the lack of explainability method this is actual and direct at the same time (i.e., it does not produce any illusion of explainability by model approximation) and utilizes the potential of the explainability method in different applications. Some of the recent works have introduced external knowledge and imparted that into the model for improving the interpretation. These XAI methods can fill the gap by introducing the domain knowledge in the model in a model agnostic and transparent procedure. As we mentioned before, it should be kept in mind that the k-nearest neighbor's(KNN) class for providing transparency is dependent on the features, the numbers of neighbors available, and the distance function that measures the similarity in between instances of the data. A high K value hinders full simulation of the performance of the AI model by human users. Similarly, using complex features and distance functions can restrict model decomposability, thus limiting the interpretability only to the transparency of its algorithmic functioning.
- **Bayesian Model**: A Bayesian model is a probabilistic graphical model where its links represented condition-based dependencies in between the set of variables. Considering an example, the Bayesian network can show the relationships between the diseases and their symptoms. If the symptoms are being given, the network could be used for computing probabilities of the existence of different kinds of diseases. Similar to the General Algebraic Modeling System, these kinds of models also provide a clean representation of the relationship between the features and targets. Again, the Bayesian models lack behind concerning the working of transparent models.

4.3.2.3 Post-model Explainability Methodology for AI Models

There arise some cases where the models are not able to meet any of the criteria which makes them transparent in nature, thus a different method needs to be conceived and applied to the model for explaining its decision. Post-model explainability is also known as post-hoc model explainability. The purpose of the post-hoc modeling explainability method is to communicate information that is understandable by the human user and it shows how a model, which is already developed, predicts the output based on provided input. Different kinds of approaches for Post-hoc explaination are shown in Table 4.2.

Table 4.2 Popular approaches for Post-hoc explaination

Types of explainations	Methods	Properties
Local explaination	Rule-based	Here the common sense knowledge and the domain-specific expertise is being represented in the form of plausible rules
	Feature importance	The explainer provides each feature with an important value which shows how much the particular feature is important for the prediction
	Saliency map	It is used in image and video processing and it shows which part is important for the AI decision
	Counterfactual	The explaination could provide a link between that what could have happened when the input of the model is being changed in a particular way
Global explaination	Representation based	It determines the model's reliance on concepts that are semantically logical for humans
	Collection of local explaination	It picks a subset of local explainations to constitute the global explaination after providing a local explaination for every instance of data
	Model distillation	It helps in learning feature shapes which provide the relation between the input feature and the prediction of the model

4.3.3 Types of Post-model Explaination Methods

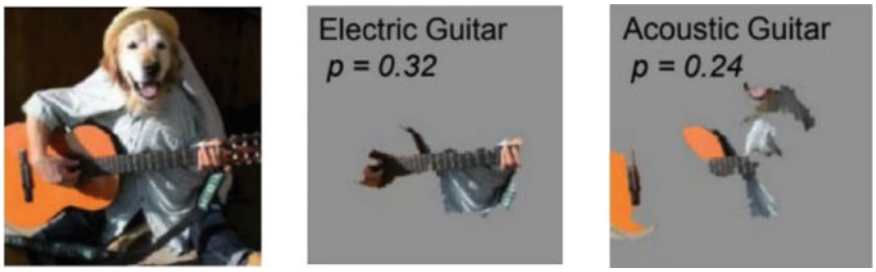
The identified trends surrounding the post-hoc explainability method for different kinds of AI models are as follows.

4.3.3.1 Model-Agnostic Techniques

These type of techniques are designed for plugging it into any kind of model intending to extract some of the information from the prediction process. Some of the techniques

for simplification are being used for generating proxies that mimic their precursor with the motivation of finding out something tractable and less complex. During other times, the main focus is to extract knowledge directly from the model and by directly visualizing it for the betterment of interpretation of their behavior. By following this taxonomy, model-agnostic techniques depend on feature relevance explaination, simplification of the model, and different visualization techniques:

- **SHAP**: SHAP (SHapley Additive exPlainations) is a game theory-based model, which is used for explaining the predictions of any ML model. Using the classic Shapley values from game theory and their related extensions, this method connects optimal credit allocation with locally generated explainations. DeepSHAP (Lundberg and Lee 2017) can be described as a high-speed approximation algorithm for SHAP values in deep learning models that are developed based upon a connection with DeepLIFT. SHAP assigns an equal importance value to every feature for a specific model prediction. The novel components of this method comprise: (1) the identification of an entirely new class of additive feature importance measures, and (2) theoretical results which show that there exists a distinctive solution within this kind of class having a set of desirable properties.
- **LIME**: The LIME framework (Ribeiro et al. 2016) proposed a much more simple approach to all of the previously discussed methods. In essence, LIME perturbs the input data and analyzes the change in the model decisions. For the image classification task, LIME divides the input image into interpretable components after generating a collection of perturbed instances. Next, every perturbed instance is run through the model to get a probability score. Following this step, a simple locally weighted linear model learns on this dataset. Figure 4.5 illustrated the explaination produced by using the LIME method for providing explaination of *Electric guitar* and *Acoustic guitar*. Finally, LIME calculates the super-pixels with the highest positive weights for explaination.
- **Layer-wise relevance propagation**: It is used for visualizing the decision of a convolutional neural network. This method provides a heatmap in the input space which indicates the importance of every pixel that contributes to the final classification. In contrast to the susceptibility maps produced by guided back-propagation,



(Original) (Explaining *Electric guitar*) (Explaining *Acoustic guitar*)

Fig. 4.5 Explaination by using LIME method on images (Ribeiro et al. 2016)

this method can directly highlight the positive contribution concerning network classification of the input space. The layer-wise propagation used by this method is subjected to conservation properties, where the information obtained by a neuron should be re-distributed to the lower layers in a uniform amount. This kind of behavior is also shared by other works on explainations (Landecker et al. 2013; Schütt et al. 2017). Assuming l and m be neurons present in two successive layers, propagating relevance scores R^{lm} in a layer into the neurons of the lower layer is attained by applying :

$$R^{lm} = \sum_{m} \frac{z^{lm}}{\sum_{l}^{z^{jm}}} \tag{4.1}$$

z^{lm} represents the extent to which a neuron l contributed to making the neuron k relevant. The denominator helps in enforcing the conservation property (Bach et al. 2015). The procedure for propagation gets terminated once the input features are covered.

The Post-hoc local explainations and feature relevance methods have gradually become the most adopted method for explaining deep neural networks.

- **DeepLIFT**: DeepLIFT (Deep Learning Important FeaTures) is a method that is used for decomposing the network predictions on a particular input by the back-propagation of the contributions of all neurons to every input feature (Li et al. 2021). In essence, DeepLIFT compares the activation of every neuron with its *reference activation* and allocates the contribution scores based on the differences. By having an option to separately consider the positive and negative contributions, DeepLIFT can reveal the dependencies missed by other approaches. The scores are efficiently computed in a single backward pass.

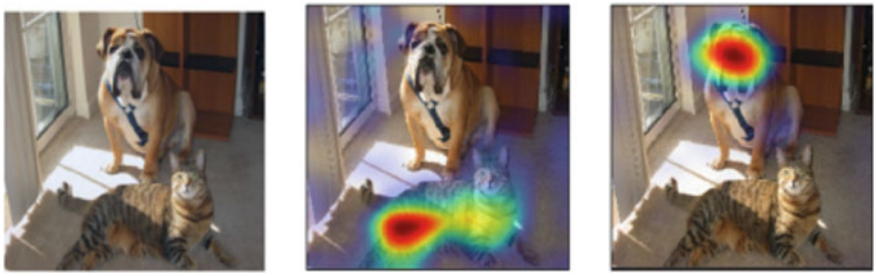
Some of the Explainability methods for visualization of the AI prediction are shown in Table 4.3.

4.3.3.2 Model-Specific Techniques

Model-specific techniques are based on the details of the applied machine learning or deep learning model's distinctive structures. These techniques are employed specifically for a model architecture, such as a convolutional network model CNN. These methods take advantage of the internals of machine learning models such as neural networks and apply a reverse engineering approach to explain how the relevant DL or ML algorithm is making the decision. The benefits of employing model-specific models are that they allow us to gain a better understanding of the choice by revealing the model's internal workings, as well as allowing us to create a more tailored explainable model. On the other hand, such models necessitate going through the entire structure of the model, putting the model's performance at risk because the ML or DL model will be recreated. The structure of the algorithms is usually exam-

Table 4.3 Visualization methods for explainable Artificial Intelligence

Methods	Description	Uses	Limitations
Deep SHAP (Lundberg and Lee 2017)	It is used for computing the SHAP values based on game theory. The algorithm is fast and is connected with DeepLIFT using multiple background samples	It helps in determining attributions for non-neural models like Decision trees, Support Vector Machines(SVM), etc	Mathematically articulated problems arise when Shapley values are interpreted as feature importance measures
Deep LIFT (Gilpin et al. 2018)	Here a reference input is used for computing the reference value of the hidden units. It avoids placing potentially misleading importance on bias terms. It consists of two variants—Rescale rule and RevealCancel. RevealCancel treats positive and negative contributions in a neuron	Rescale is sometimes related to e-LRP but it can not be applied to models that involve multiplicative rules. RevealCancel manages such situations by using RevealCancel for convolutional and Rescale for fully connected layers which helps in reducing the noise	It can not be applied to models that involves multiplicative rules
Integrated Gradients (Qi et al. 2019)	It aims to explain the relationship between a model's predictions in terms of its features. It works by computing the average gradient because the input differs from the baseline to the value of the actual input, unlike the Gradient input which uses a single derivative at the input.	It has many use cases which include understanding the feature importance and debugging the model performance, it is also used for augmenting the accuracy metrics.	The heatmap generated by the integrated gradients is diffused
Saliency Maps (Simonyan et al. 2013)	These compute the absolute value of the partial derivative of the output neuron w.r.t the input features for detecting the most influential feature. It shows the importance of a pixel to the human visualization system	This method captures the instinct such that the information of a location is proportional to the activation level. It processes the images for differentiating the visual features present in images.	It is noisy because deep neural networks do not filter out irrelevant features during forward propagation
Layer wise relevance propagation (LRP) (Montavon et al. 2019)	The purpose of LRP is to provide an explaination of a neural network's output with context to the input's domain It distributes the prediction score layer with the layer having a backward pass on the network by using specific rules like e-rule when the numerical stability is ensured	It brings explainability to the AI decisions that help to scale up the potential of complex deep neural networks. LRP works by backward propagation of the network, using a set of designed rules for propagation. This method does not interact with training the network, so it could be easily applied to pre-trained classifiers	This method is limited to the CNN models with ReLU activation
Grad-CAM (Selvaraju et al. 2017)	A gradient-based class activation map is produced by using the gradients of the targeted concept as it goes to the last convolution layer	This method applies to CNN which includes the fully connected layers, structured output, and reinforcement learning	Lack of explaining decisions produced by deep networks in the domains like natural language processing, and reinforcement learning



(Original) (Explaining *Cat*) (Explaining *Dog*)

Fig. 4.6 Grad-CAM showing visualization heat map for differentiating the images as *Cat* and *Dog* (Selvaraju et al. 2017)

ined using model-specific approaches. Some of the model specific techniques are illustrated as follows.

- **Grad-CAM**: Grad-CAM is the acronym for Gradient weighted Class Activation Mapping. Several visualization approaches and gradient-based methodologies (Zeiler and Fergus 2014) have been used for providing explainations in deep learning solutions. Gradient weighted Class Activation Mapping applies the gradient of the targeted conception into the final convolutional layer of the model for building a localization map that highlights the main region of the input image. Grad-CAM observes both the forward and backward passes and gives better visualizations by representing a local influence of the input image in predicting the output classes. We only consider backward passes in the de-convolution methods. This algorithm is a class discriminated localization technique that brings out higher resolution visual explaination from the CNNs (Convolutional Neural Networks). A Grad-CAM based visual explaination of the AI model identified as *Cat* and *Dog* using the heatmap is shown in Fig. 4.6.
  For producing the localization map for a specific image class i, Grad-CAM finds out the gradient score g^i before the softmax layer with reference to the feature map f^k :

$$g^i(f^k) = \frac{dy^i}{df^k} \tag{4.2}$$

  here, k is denoted as an index of the channel and thus the averaged gradient score is denoted as m^c where:

$$m^c = \frac{1}{UV} \sum_i \sum_j g^i(f^k) \tag{4.3}$$

  here U and V represent the length and the width of an image provided as input. Last feature map of the class is represented as N^k. Where $N^k \subset P^{U*V}$, and, P is real number. Thus from Equation 2, The Grad-CAM can be represented as Equation

3:

$$Grad - CAM = relu(\sum_{k=1}^{\infty} m^c)N^k \tag{4.4}$$

The weight of the feature maps is being used and the weighted sum is being calculated which generates the final heat map.

- **Integrated Gradient**: Integrated Gradient (IG)) is a method for computing the gradient of the predictions of a model to its input features and does not require us to modify the original deep neural network (Saibi et al. 2006). We can apply IG to any model that is differentiable like images, texts, or any type of structured data. Methods such as Integrated Gradients are model-specific and require knowledge about the internal model for computing the gradients of the model layers.
- **Saliency maps**: The objective of computing saliency maps is to find out the image regions that are different (or conspicuous) from their neighbors based on the image features (Thompson and Bichot 2005). Given an image, we first compute the basic image features such as color, orientation, intensity, etc. These processed images are then used to construct Gaussian pyramids that are in turn used to create feature maps at different scales. Next, the saliency map is created by taking the average of all the feature maps (Zeiler and Fergus 2014). This technique measures information at the end of every network scale that is subsequently combined to form a single saliency map.

4.4 Metrics for Explainable Artificial Intelligence

Different evaluation metrics should accompany the measures adopted in XAI research (Hoffman et al. 2018). The evaluation metrics are expected to achieve the identified properties of explainability as objectives. The quantitative metrics for both model-based and example-based explainations are largely used to assess the ease with which they can be understood, whereas the quantitative metrics for attribution-based explainations are mostly used to assess the fidelity with which they can be explained. We have identified the different properties of explainability by studying the corresponding definitions. These identified properties of explainability are used as objectives that we expect the evaluation metrics to achieve. Standardization organizations such as ISO have also raised the need for evaluation metrics for measuring the reliability of AI systems (Russakovsky et al. 2015). Their standardization document has outlined different challenges related to the implementation and the use of AI systems. Over-reliance and under-reliance on the AI system are the primary concerns raised by this work. Over-reliance occurs when a user is more reliant on the automation aspects without considering the associated limitations. This can have adverse outcomes (Yang et al. 2019). Under-reliance occurs when the user frequently disagrees with the correct AI system decisions.

4.4.1 Evaluation Metrics for Explaining AI Decisions

As discussed in Sect. 4.3, the methods of explaination are broadly classified into three types namely Model-based Explainations, Attribute-based Explainations, and Example-based Explainations. In this section, the quantitative metrics are being discussed for measuring the qualities of these explaination methods. Apart from the evaluation of the AI explaination methods, proper selection of the evaluation metrics plays a major role in evaluating the system accurately. Different types of metrics that evaluate the extent of the explainability by different methods are as follows:

- **Subjective metrics**: It is designed for questioning the users based on tasks and the explainations provided, these questions are asked when the task gets executed or afterward for obtaining the subjective response from the user on the explainations. Some of the examples of these types of metrics are the confidence, and trust of users that have an enormous grasp over the focal points for evaluating the explainable system. Hoffman et al. (2018) proposed a metric that is used for the subjective evaluation of an AI system. It considers factors like user trust, understanding, and satisfaction. Zhou et al. (2016) have looked over the factors like the uncertainty that affect the trust of users in informed machine Learning decision makings. They established that explaination generated because of influence in the training data points remarkably affects the user's trust in case of informed decision making.
- **Objective metrics**: It mentions the objective information of a task to a user before or after the task is being performed. Such kinds of examples are human metrics, which include behavior and physiological measures of humans when informed decision making takes place, another such metric is task-related metrics, which include time length for completing the task and performance of the task. Schmidt and Biessmann (2019) showed that fast and accurate decisions mean instinctively understanding the explainations provided. It resulted in deriving a trust metric that is based on the explainability metrics.
- **Computational metrics**: These metrics are known as mathematical indicators for determining the quality of explainations generated by an XAI system. The measurement of these kinds of explainations is generally being carried out by using necessarily developed equations. Thus, these metrics may be used without any kind of human intervention as guidelines for preparing the explaination techniques.
- **Cognitive metrics**: The explainations provided to the end-users are being measured by using cognitive metrics. The assessment of human subjects is a blunt indication of explainations, as we know the initial goal of XAI is to convey the reasons behind machine judgments to people.

Accuracy is one of the most commonly used metric (Rosenfeld 2021), it is very easy for understanding although noticing only these metrics would give an incomplete suggestion regarding the performance of a model. Multiple established metrics are there which would provide a thorough insight into the performance of the model. The metrics used for quantifying the explainations are generally very specific to the different types of machine learning problems and models. Some of the widely used

Table 4.4 Metrics for quantitatively explaining AI decisions

Types of explaination	Metrics	Explaination properties
Model-based explainations	Model size (Guidotti et al. 2018)	Simplicity
	Interaction strength (Markus et al. 2021)	Simplicity
	Level of agreement (Lakkaraju et al. 2017)	Clarity
Attribution-based explainations	Effective complexity (Nguyen and Martínez 2020)	Broadness and simplicity
	Recall of important features (Ribeiro et al. 2016)	Soundness
	Selectivity and continuity (Montavon et al. 2018)	Soundness and clarity
	Mutual information (Nguyen and Martínez 2020)	Broadness and soundness
	Sensitivity (Montavon et al. 2018)	Soundness
Example-based explainations	Diversity (Nguyen and Martínez 2020)	Simplicity
	Non-representatives (Nguyen and Martínez 2020)	Simplicity and completeness

metrics are Loss, Confusion Matrix, Accuracy, Mean Absolute Error, Root Mean Square Error, Accuracy, etc.

Table 4.4 shows the metrics for quantitatively explaining the AI decisions for explaining the classification.

4.5 Use-Case: Explaining Deep Learning Models Using Grad-CAM

As a use-case of XAI, we implemented some of the visual methods of XAI for explaining the learned models for classifications of fresh tea leaves (Banerjee et al. 2022). We have applied a well-known technique for introducing visual explainations of predictions from Neural Network-based models, which makes them more interpretable. We have used transfer learning for classifying different types of tea leaves. The aim here is to reveal the underlying features that are responsible to explain the relationship between the predictions of a tea-leaf classification model by using popular visualization techniques. In this work, the Grad-CAM method has been used that use the gradients of the targeted concept. We performed our experiment on our image dataset created by imaging and labeling the freshly harvested tea leaves

from *Banuri experimental tea farm*, Palampur, India. Our data consists of 965 tea leaf images where the combination of 1-Leaf 1-bud consists of 261 data samples, 2-Leaf 1-bud has 174 data samples, 3-Leaf 1-bud have 279 data samples, 4-Leaf 1-bud has 199 data samples, and 5-Leaf 1-bud has 52 data samples. Because of the class imbalance in our dataset, we have implemented data augmentation, which reduced the class imbalance. We used the final layer of convolution for producing a localization map that highlights the most important regions of the image while determining the class of the tea-leaf. While performing our experiments, we found that the Grad-CAM method takes out features from the pre-trained VGG16 model. We implemented transfer learning by using well-known pre-trained backbone models like VGG16 and InceptionV3. For our tea leaf dataset, we discovered that the performance evaluation metrics of our built model on 1L1B(1-Leaf 1-bud) have overall best performance, with precision, recall, and F1-scores of 0.90, 0.90, 0.90. Our 2L1B data (2-Leaf 1-bud) has precision, recall, and F1scores of 0.73, 0.62, and 0.67, respectively. Whereas 3L1B (3-Leaf 1-bud) has a precision of 0.80, recall of 0.82, and F1-score of 1. Moreover, 4L1B (4-Leaf 1-bud) has a comparatively low precision of 0.69, but has a good recall of 0.92, and has an F1score of 0.79. 5L1B (5-Leaf 1-bud) has a precision of 1 and has a comparatively poor recall of 0.18, and an F1score of 0.30 which shows the model has certain limitations in the identification of the actual distinguishing features responsible for classifications for tea leaves. Our trained classification model, which employs VGG16, predicts an average of 0.83, 0.69, 0.74 precision, recall, and F1-score, with an accuracy of 80%. And by using InceptionV3 our model could able to have achieved an accuracy of 76.41% respectively. We obtained explainations based on the classification in our dataset of the tea leaf images by providing the pre-trained models as input to the Grad-CAM pipeline for producing class-specific heat maps. We used Grad-CAM as an explaination method for explaining our prediction of the model. The model generates explainations for different leaf image categories. Grad-CAM exclusively highlights the important portions of the image which are of utmost importance for predicting the class of the tea leaf image. Figure 4.7 shows the generated heatmaps for explaining the most important image regions which are influential in predicting the result using pre-trained model VGG16 and InceptionV3 models.

The heatmap produced by Grad-CAM for the pre-trained VGG16 model focuses particularly on the Leaf and Bud portion of the input image, whereas the heatmap generated by Grad-CAM with pre-trained InceptionV3 model has an inappropriate focus on the stem region of the leaf image. Grad-CAM uses the final feature map of the model for creating heatmaps that highlight the image pixels responsible for the prediction of the image class.

4.6 Challenges and Future Directions

Explainable AI is an active area of research for handling the black-box nature of deep learning models. The most recent literature survey, reveals that though there

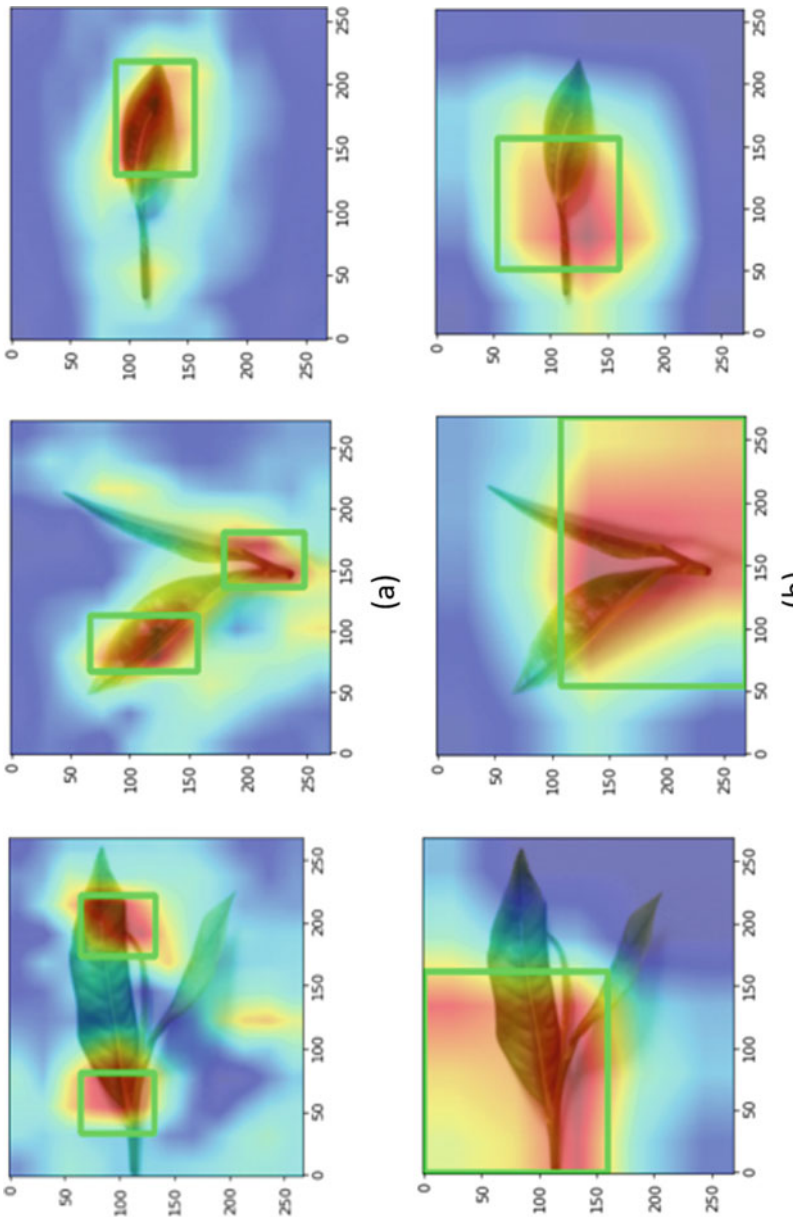


Fig. 4.7 Explaining the prediction by using Grad-CAM and pre-trained **a** VGG16, and **b** InceptionV3 models (Banerjee et al. 2022)

are several works published in this domain of research there are still many important problems to be resolved by the research communities. Though there are a lot of AI black-box methods are there which need explainations but due to high-dimensional issues, the image classification and object detection problem in computer vision is challenging. The black-box models that yield highly accurate results normally need to be explained using visual explaination for better perception of the underlying working of the model behind the same. However, the main challenge for visual explainations of a classifier's output arises from the complex nature of the AI model and underlying data. Since the visual explaination method relies on the algorithm for the generation of human perceptible feedback which becomes subjective in nature. Moreover, it is very challenging to properly explain complex classification models accurately. The classifiers which provide good performances are gradually becoming more complex due to the involvement of the numerous parameters and the operations performed by them which in turn makes them complex and tough to explain. Because of the different architectures of the AI methods, the problem to design an effective framework for explaining the underlying decision process increases. Other challenges include the use of multiple types of data for training the models. For explaining an AI model, the most common strategy is to trace back to the input data. Different types of data require different types of explainations. Though the AI model based on image data is relatively easier for generating visual interpretations of the decision process, the same using the textual, speech, or nominal data is difficult to get explained similarly. Even in image data-based AI models, there are no objective metrics available that measures the quality of explainations. This paves the way for further research in this area. In this section, we endeavored to identify a few main research questions in the field of XAI. These questions include some necessary aspects like how do we evaluate the Extent of Explaination for an AI model? While referring to the AI perspective, metrics are defined as any quantification of the extent of explaination which helps in evaluating its quality and suitability.

Evaluation metrics explain the model's performance. It is used in measuring the quality of the explaination. Though there are multiple metrics used for tasks of classification, ranking, clustering, regression, modeling topics, etc. there are very limited metrics are there for measuring the extent of explaination in different data types and AI tasks. This in turn forms another research question whether a method of explaination can be devised that is invariably applicable to any data type or AI model?

There are different methods of explainable AI that target explaining the decision-making of the AI system and thus can identify the problems associated with the underlying model. But it needs to be investigated how explainable AI models can help in improving the prediction accuracy or inhibiting the failure points in real-life problems. XAI consists of a set of frameworks and tools which helps in interpreting and understanding the predicted output of the AI models. It is also imperative to study the applicability of the explaination method on a local instance of data or globally on an entire dataset. There are techniques where an AI model can be explained locally and globally based on the given input data but rarely any specific XAI method is available which can effectively evaluate the quality of explainations both in the local

and global dataset. Also, there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these kinds of methods for increasing trust in them. Moreover, these kinds of metrics could be used for standardizing the XAI solutions.

4.7 Conclusion

Explainable AI (XAI) is an emerging technological paradigm of which most enterprises are conscious. The methods and processes of XAI provide several advantages. Explainability in pre-modeling is a feasible but under-focused approach for avoiding transparency problems. Pre-modeling explainability methods mainly focus on the explainability of data rather than the model itself. Whereas Post-modeling/Post-hoc explainability is a collection of different types of methods with a common goal of gaining a better understanding of the working of the trained model. Based on Post-hoc explaination the methods are classified into model-agnostic and model-specific techniques. Moreover, the metrics used for explaining the AI decisions are quantitatively evaluated as Subjective metrics, Objective metrics, Computational metrics, and Cognitive metrics for evaluating the AI system accurately. Although there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these methods for enhancing trust in them. For increasing transparency in the developed model, it is necessary to produce an intuitive explaination. In this chapter, we have presented a comprehensive overview of the methods and metrics for explaining decisions made by AI models. We also covered the taxonomy of XAI in ample detail and discussed different strategies used for providing explainations behind the working of the data-based learned models. We have presented a selected overview of works to assist researchers and practitioners in understanding insights, accessible resources, and unresolved difficulties in using XAI methodologies. A use-case of implementing a popular XAI visualization method is also been demonstrated in this chapter. Though the research work in the area of XAI is in full swing but still has many gray areas to be addressed by the global AI communities. The research directions section in this chapter is an endeavor to summarize the identified research gaps and unanswered research questions for prospective XAI researchers.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)

Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:1909.03012 (2019)

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130, 140 (2015)

Banerjee, P., Banerjee, S., Barnwal, R.P.: Explaining deep-learning models using gradient-based localization for reliable tea-leaves classifications. In: 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1–6. IEEE (2022)

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U.: Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 (2017)

Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019)

Christopher Frey, H., Patil, S.R.: Identification and review of sensitivity analysis methods. Risk Anal. **22**(3), 553–578 (2002)

Cortez, P., Embrechts, M.J.: Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 341–348. IEEE (2011)

Cortez, P., Embrechts, M.J.: Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. **225**, 1–17 (2013)

Dignum, V.: Responsible artificial intelligence: designing AI for human values (2017)

Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89. IEEE (2018)

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Statist. **24**(1), 44–65 (2015)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press (2016)

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 1–42 (2018)

Hellevik, O.: Linear versus logistic regression when the dependent variable is a dichotomy. Qual. Quant. **43**(1), 59–74 (2009)

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

Jain, A., Ravula, M., Ghosh, J.: Biased models have biased explanations. arXiv preprint arXiv:2012.10986 (2020)

Khaleghi, B.: The how of explainable AI: explainable modelling. https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed

Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning, pp. 1885–1894. PMLR (2017)

Krarup, B., Cashmore, M., Magazzeni, D., Miller, T.: Model-based contrastive explanations for explainable planning (2019)

Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)

Landecker, W., Thomure, M.D., Bettencourt, L.M., Mitchell, M., Kenyon, G.T., Brumby, S.P.: Interpreting individual classifications of hierarchical networks. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 32–38. IEEE (2013)

Li, J., Zhang, C., Zhou, J.T., Fu, H., Xia, S., Hu, Q.: Deep-lift: deep label-specific feature learning for image annotation. IEEE Trans, Cybern (2021)
Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inf. **113**, 103,655 (2021)
Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, pp. 193–209 (2019)
Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Process. **73**, 1–15 (2018)
Next Move Strategy Consulting (NMSC): explainable AI market size , share, forecast, industry analysis report | 2021 - 2030. https://www.nextmsc.com/report/explainable-ai-market
Nguyen, A.P., Martínez, M.R.: On quantitative aspects of model interpretability. arXiv preprint arXiv:2007.07584 (2020)
Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
Qi, Z., Khorram, S., Li, F.: Visualizing deep networks by optimizing with integrated gradients. In: CVPR Workshops, vol. 2 (2019)
Rao, S.J.: Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis (2003)
Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)
Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, pp. 45–50 (2021)
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
Saibi, H., Nishijima, J., Ehara, S., Aboud, E.: Integrated gradient interpretation techniques for 2D and 3D gravity data interpretation. Earth Planets Space **58**(7), 815–821 (2006)
Samek, W., Müller, K.R.: Towards explainable artificial intelligence. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 5–22. Springer (2019)
Schetinin, V., Fieldsend, J.E., Partridge, D., Coats, T.J., Krzanowski, W.J., Everson, R.M., Bailey, T.C., Hernandez, A.: Confident interpretation of Bayesian decision tree ensembles for clinical applications. IEEE Trans. Inf. Technol. Biomed. **11**(3), 312–319 (2007)
Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:1901.08558 (2019)
Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nat. Commun. **8**(1), 1–8 (2017)
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
Suman, R.R., Mall, R., Sukumaran, S., Satpathy, M.: Extracting state models for black-box software components. J. Object Technol. **9**(3), 79–103 (2010)

Tan, S., Soloviev, M., Hooker, G., Wells, M.T.: Tree space prototypes: another look at making tree ensembles interpretable. In: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, pp. 23–34 (2020)

Thompson, K.G., Bichot, N.P.: A visual salience map in the primate frontal eye field. Prog. Brain Res. **147**, 249–262 (2005)

Tukey, J.W., et al.: Exploratory Data Analysis, vol. 2. Reading, MA (1977)

Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the National Conference on Artificial Intelligence, pp. 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2004)

van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103,404 (2021)

Yang, F., Du, M., Hu, X.: Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831 (2019)

Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer (2014)

Zhou, J., Arshad, S.Z., Yu, K., Chen, F.: Correlation for user confidence in predictive decision making. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction, pp. 252–256 (2016)

Chapter 5
Evaluation Measures and Applications for Explainable AI

Mayank Chopra and Ajay Kumar

Abstract Machine learning advances, particularly deep learning, have enabled us to design models that excel at increasingly complicated tasks. Because of the growing size and complexity of these models, it's becoming more difficult to grasp how they arrive at their forecasts and when they go incorrect or even worse. Now, think of a situation in which we humans could open these black-box learning models and translate the content into a human-understandable format. This is known as Explainable Artificial Intelligence and there has been a lot of research in this field over the last few years mainly focusing on how to explain different types of models. The advancement of this research, raised a very important query: "Why does a model need to be explained?" So, the most accurate answer to this question is "TRUST". TRUST that the models are making the correct decisions over the correct assumptions. TRUST that we can tell what happened when a model fails. TRUST that we can do on a model implemented on a large scale that the predictions are made in line with expectations. It's hard to trust a system that's not transparent about its internal processes. This paper discusses the evaluation measures and application areas of XAI. Some XAI-related concepts were also mentioned.

Keywords Explainable artificial intelligence · Interpretable machine learning · Machine learning · XAI · Deep learning

5.1 Introduction

It is crucial for an artificial intelligence (AI) program to be able to give accessible justifications that properly justify its conclusions in circumstances when people are expected to make key decisions powered by AI. An adequate justification can

M. Chopra (✉) · A. Kumar
Department of Computer Science and Informatics, Central University of Himachal Pradesh (H.P.), Dharamshala, India
e-mail: mayankchopra.it@gmail.com

A. Kumar
e-mail: ajaykr.bhu@hpcu.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_5

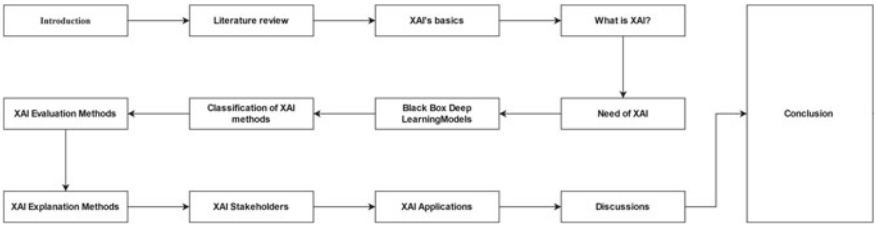


Fig. 5.1 Work Flow of the paper

enhance the system's trust, enabling better human-AI collaboration (Villata et al. 2013). Explanations can assist people in determining how much, they must believe the explanation source. In both industry and academia, AI algorithms are gaining a lot of traction (Attaran and Deb 2018). Since Machine Learning (ML)-based procedures are swamped with thousands of hardly analyzable variables to be improved during the training phase, a variety of these procedures are frequently referred to as "black-box" algorithms (Stepin et al. 2021). Because of this, the algorithm's outcome is difficult to explain. The failure to comprehend these automated judgments diminishes users' trust in such systems, reducing their usability (Ribeiro et al. 2016). Furthermore, many current explainable AI systems deliver summaries of automatically generated forecasts instead of complete explanations (Rudin 2019). As a consequence, the necessity to justify automated judgments with a good description of why the algorithm makes a specific decision has fueled rapid growth in the explainable AI research field (Anjomshoae et al. 2019).

In this paper, we have briefly reviewed the Explainable Artificial Intelligence Literature followed by the XAI's Evaluation Methods which is followed by the XAI's Explanation Methods. After that, we revisited the stakeholders and the applications. Some proposed frameworks in the field were also taken into consideration and in the end, we have concluded our study. Figure 5.1 shows the workflow of our paper.

5.2 Literature Review

The ideas of understandability and comprehensibility are difficult to describe precisely; however, numerous attempts have been made, with the most famous works comprising (Lipton 2018; Doshi-Velez and Kim 1702). Gilpin et al.'s (Gilpin et al. 2018) work is the other effort to identify the key concepts of understandability in machine learning. The researchers created a taxonomy for defining accuracy techniques for neural nets into three classifications while focusing primarily on deep learning. Between 2004 and 2018, Adadi and Berrada did a thorough analysis by collecting and processing 381 separate academic papers. They categorized all research on explainable AI into four categories, outlining the necessity for additional approval in the XAI domain including more human–machine interaction. After

noticing the community's habit of researching explainability only in the aspects of modeling, they called for incorporating it into other aspects of machine learning (Kumar and Chatterjee 2016). Finally, they suggested a potential area of study including the fusion of existing explainability approaches (Adadi and Berrada 2018). Murdoch et al. (Murdoch et al. 2019) published a fact sheet following the discovery of the shortage of regularity and a way to measure the effectiveness of classification strategies. They devised an interpretability paradigm to overcome the previously indicated gap. Classification power, describing accurateness, and appropriateness are three types of metrics presented by the Relevant, Descriptive, and Predictive framework for evaluating classification methodologies (Kumar et al. 2021). According to recent research, a new type of arrangement was provided that first differentiated truthful and subsequent methods, and then generated sub-classes (Arrieta et al. 2020). A separate ontology was created specifically for deep learning classification techniques due to the large number of them (Gautam and Chatterjee 2021).

5.3 Basics Related to XAI

5.3.1 Understanding

Understanding is linked to the human ability to spot connections and the perspective of a dilemma, and it is a prerequisite for explanations. Mechanistic understanding ("How does anything operate?") and functional understanding ("What is its objective?") are two types of understanding.

5.3.2 Explicability

Explicability refers to the ability to examine the characteristics of an AI system.

5.3.3 Explainability

Explainability goes beyond explicability by aiming to make the perspective of an AI system's logic, model, or findings for a judgement outcome available, so that beings can comprehend it.

5.3.4 Transparency

An AI system is transparent if its algorithmic activity in terms of decision outcomes or procedures can be comprehended by a human.

5.3.5 Explaining

Using explicability or explainability to enable a human to recognize a model and its intent is referred to as explaining.

5.3.6 Interpretability

Interpretability implies how an AI system's judgement could be clarified globally or locally and that the system's intent could be comprehended by a human.

5.3.7 Correctability

Correctability denotes the capacity of an AI system to be focused by a human actor to guarantee the right choices.

5.3.8 Interactivity

Interactivity occurs when it is feasible to progressively investigate and accommodate the inner structure of a model (correctability). This differs from local and global understandability, which relates to the presentation of outcomes and routes.

5.3.9 Comprehensibility

Comprehensibility, like interpretability, is based on local and global interpretations as well as functional knowledge. Furthermore, understandable AI satisfies interactivity. Both interpretable demonstration and invasion are viewed as essential elements for in-depth comprehension and thus as prerequisites to comprehensibility.

5.3.9.1 Human-Artificial Intelligence System

A human-AI system includes both computational elements and a human user who must come together to accomplish a purpose.

5.4 What is Explainable AI?

Explainability is a concept that stands at the crossroads of numerous fields of active AI research, with an emphasis on the following domains.

5.4.1 Fairness

Can we verify that choices taken by an AI system were done consistently?

5.4.2 Causality

Can one learn a system from facts that not only makes the right predictions but also offers some understanding of the core events?

5.4.3 Safety

Can we have confidence in the reliability of our AI system without recognizing how it makes its presumptions?

5.4.4 Bias

How can we be confident that the AI system hasn't picked up a distorted view of the world due to flaws in the training data or objective function?

5.4.5 Transparency

Everyone has a right to be informed about changes that influence us in ways, formats, and languages that we comprehend.

An XAI, also known as a "Transparent AI" or "Interpretable AI", is an AI whose activities are simple for humans to comprehend and evaluate. A civic privilege to explain can be implemented using XAI.

5.5 Need for Transparency and Trust in AI

The black box AI systems have found their way into many of today's modern implementations. Transparency and explainability are not critical requirements for machine learning models used as long as the overall efficiency of these systems is adequate. Even if these systems fail, the implications are unexceptional. As a result, the necessities for trust and openness in these types of AI systems are relatively low. The scenario is different in safety–critical applications. In this case, the opaqueness of ML techniques may be a restricting or indeed rejecting component. Particularly when a single misjudgment can endanger human life and health or lead to significant revenue damages, depending on an information system with unintelligible logic will not be an alternative. This lack of transparency is among the causes why the application of machine learning to areas such as healthcare is extremely careful than its application in the consumer, electronic commerce, or media industries.

5.6 The Black Box Deep Learning Models

The method of developing interpretations for AI system behavior will vary based on the type of ML techniques used: techniques that produce implicitly decipherable models vs deep learning algorithms that are intricate information and understanding methods and produce models that are implicitly indecipherable to actual users.

ML techniques such as Bayesian classifiers, decision trees, spare linear models, and additive models produce decipherable models in the sense that model components can indeed be instantly examined to comprehend the model's inferences. These technique makes use of relatively small internals, and also provide visibility and traceability in their decision-making. As long as the model is precise for the classification process, these strategies offer awareness of the AI system's decision-making.

Deep learning algorithms, one on either side, are a class of machine learning technique that sacrifices clarity and interpretability for the predictability. These techniques are now used to create applications such as consumer behavioral forecasting associated with high inputs, voice recognition, natural language processing, and computer vision.

The lack of transparency and understandability in the Deep Learning Algorithms makes them a black box. The black box model is a model which performs its predictions on its own without explaining anything for humans to understand.

The Black Box Problem occurs when artificially intelligent processor architectures are vague.

This figurative language is based on the notion that the function of a system may be explained by "gazing within." Although, modern computing systems are composed of well-known hardware components that present no physical barriers to peering inside. However, they may be seen as mysterious in the sense that it is tough to comprehend how such devices are designed.

With time prediction techniques have made considerable latest improvements in tackling the transfer among analysis and prediction affiliated with deep learning models—these techniques estimate deep-learning black-box models with simplified decipherable models that could be examined to clarify the black-box models.

These techniques are known as XAI because they transform black-box models into crystal models. They are gaining popularity because they allow Ai systems to undertake both predictive performance and interpretability goals.

5.7 Classification of XAI Methods

In the literature, several categorizations have been suggested to categorize various explainability techniques. In general, categorization methods are not exquisite; they can vary greatly based on the methodological features and can be categorized into several overlapping or non-overlapping classes at the same time. Various classification methods and taxonomies are briefly mentioned here, and a flow diagram for them is exhibited in Fig. 5.2 Flow diagram of the classification of XAI Methods.

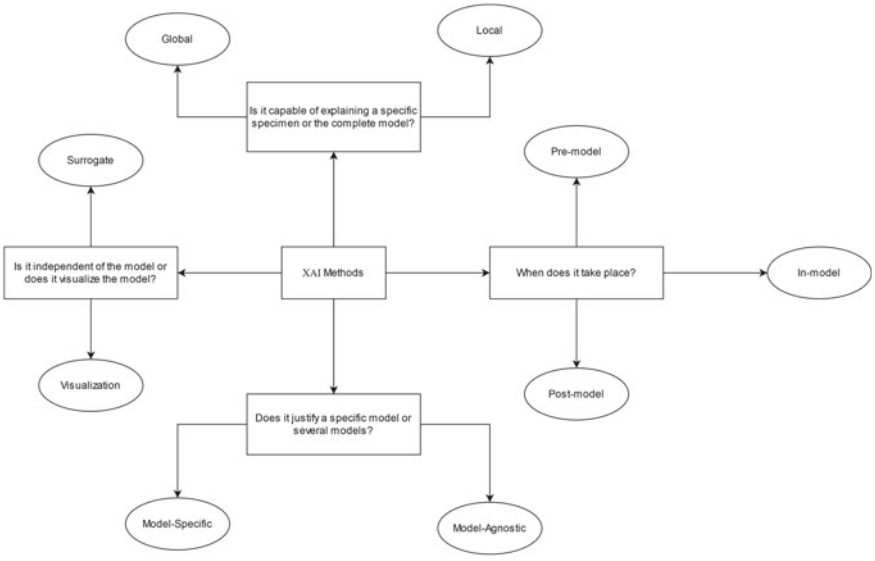


Fig. 5.2 Flow diagram of the classification of XAI Methods

5.7.1 Global Methods Versus Local Methods

Local explainable techniques are only useful for single model output. This can be accomplished by developing strategies for explaining a specific estimation or conclusion. Global methods, on the other hand, focus on the interior of a prototype by leveraging the knowledge base about the prototype, the mentoring, and the related data. It attempts to clarify the model's conduct in particular.

5.7.2 Surrogate Methods Versus Visualization Methods

Surrogate methods are ensembles of various models that are used to examine other black-box models. The black box models can be effectively comprehended by analyzing the surrogate model's judgments and correlating the black-box model's and surrogate model's judgments. Surrogate techniques are illustrated by the decision tree. The visualization techniques are not just a separate model, but they help clarify some aspects of the models through the image process, such as activation maps.

5.7.3 Model Specific Versus Model Agnostic

The metrics of the specific models are used to determine model-specific analysis techniques. The GNNExplainer (graph neural network explainer) is a type of model-particular explainability in which the complex nature of data depiction necessitates the use of the GNN specifically. Model Agnostic techniques are usually used in afterward assessment and are not restricted to a specific model architecture. These techniques lack explicit direct exposure to underlying model weights and architectural metrics.

5.7.4 Pre-Model Versus In-Model Versus Post-Model

Pre-model methods are self-contained and do not require a specific model architecture to be used. These methods include principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). In-model methods are interpretability techniques that are incorporated into the approach itself. Some methodologies are applied after creating a model and thus are referred to as post-model methods. These systems can build valuable observations about what a model gained throughout mentoring.

5.8 XAI's Evaluation Methods

Another significant aspect of the design process for XAI systems is the evaluation methodologies. Diverse metrics are prescribed to test the validity of the explanation for the desired purpose since explanations are created to meet a variety of interpretability aims.

5.8.1 Mental Model

A mental model is a depiction of how one perceives a system. according to cognitive psychology theories. Human–Computer Interaction (HCI) researchers examine users' conceptual models to see how much they know about intelligent systems in a variety of uses. One study looked at how individuals comprehend a smart grid system (Costanza et al. 2014), while another looked at how individuals comprehend and react to unpredictability in machine learning bus advent time predictions (Kay et al. 2016). Inquiring directly about the intelligent system's decision determining procedure is a good technique to examine users' understanding of intelligent systems. "Interviews, think-aloud, and self-explanations" provide useful data about an individual's thinking patterns and mental models when analyzed (Kim and Seo 1709).

5.8.2 Explanation Usefulness and Satisfaction

When evaluating explanations in expert systems, end-user convenience and the effectiveness of equipment justifications are also major determinants (Bilgic and Mooney 2005). To assess explanatory value for users, researchers use a variety of quantitative and qualitative criteria for comprehension, applicability, and adequacy of information (Miller 2019). Even though there are implicit methods of evaluating user acceptance (Hoffman 2018a, b), the evaluative assessment of contentment in descriptions, such as surveys and meetings, is a huge chunk of the literature.

5.8.3 User Trust and Reliance

User trust in expert machines is a cognitive and emotional factor that impacts whether users think a system is good or bad (Hoffman et al. 2013; Madsen and Gregor 2000). Swift trust (Meyerson et al. 1996), default trust (Merritt et al. 2013), and suspicious trust (Bobko et al. 2014) have all been used to describe the early individual faith and the trust-building with time.

5.8.4 Human-AI Task Performance

One of XAI's main goals is to assist end-users in becoming more prosperous in jobs using machine learning algorithms (Höök 2000). As a result, human-AI task productivity is a metric that applies to all user kinds. Users can adapt the intelligent system that caters to their requirements with the help of explanations. By delivering model interpretations, visual analytics tools also assist domain specialists in performing their responsibilities more effectively. Domain specialists can detect models and modify hyper-parameters to their particular data by evaluating model design, features, and machine output ambiguity. The necessity for model interpretation in a text (Hu et al. 2014; Liu et al. 2015; Wise et al. 1995) and multimedia (Bryan and Mysore 2013; Choo et al. 2010) assessment tasks has been investigated in visual analytics research.

5.8.5 Computational Measures

In the discipline of machine learning, computational measurements are commonly used to assess the correctness and completeness of classification strategies in terms of elucidating whatever the model has discovered. Instead of human subject investigations, computer tools should be used to determine the sincerity of explanations to the black-box model. The integrity of an improvised strategy in creating genuine justifications for model predictions is referred to as ad-hoc explainer integrity. As a result, a set of computational criteria for assessing the validity of produced interpretations, uniformity of interpretive outcomes, and authenticity of ad-hoc classification procedures to the initial black-box concept (Robnik-Šikonja and Bohanec 2018) have been developed.

5.9 XAI's Explanation Methods

In this section, we have listed some of the open-source methods used for the explanation.

5.9.1 Lime

An algorithm for faithfully explaining the outcomes of any encoder by estimating them effectively with an explainable fashion (Ribeiro et al. 2016).

5.9.2 Sp-Lime

An optimization strategy that identifies a group of relevant samples with justifications to handle the "trusting the model" problem (Ribeiro et al. 2016).

5.9.3 DeepLIFT

DeepLIFT is a deep learning recurrent estimation interpretive approach (Shrikumar et al. 1605). The ES attributes for a linear variant of the deep network are called DeepLIFT parameters. The use of DeepLIFT as an efficient framework for sampling-free estimation of ES attributes is motivated by this link. ES values can also be utilised to validate certain linearization choices made by DeepLIFT (Lundberg and Lee 1611).

5.9.4 Layer-Wise Relevance Propagation

Another method to approach the estimations of compositional networks (Bach et al. 2015). It is similarly an estimate of ES values, with the key contrast with DeepLIFT being the standard input used to estimate the impact of absent information (Lundberg and Lee 1611).

Some of the open-source methods used for the explanation are discussed above. Several typologies for classifying techniques for understandability have been suggested. Methods are classified as follows: characteristic value evaluation, reasoning from examples, and latent space traversal (Hase and Bansal 2020).

5.9.5 Characteristic Value Evaluation

Evaluation of characteristics value focuses on providing details on why the model uses particular aspects. The gradient-focused methods initially presented for vision by Simonyan et al. (1312), which could be transformed to be used with text data (Li et al. 1506), are the most prevalent of all these techniques. A variety of alternative methods for evaluating characteristic values all over data realms were suggested. LIME and Anchor are domain-agnostic approaches. Simple models, such as scattered linear models and instruction lists, are used in these techniques to estimate compound model actions relatively circling inputs. They demonstrate the evaluated results of explicitly interpretable characteristics on model results. What is "local" to input is described domain-specifically using an agitation allocation circling on that input?

5.9.6 Reasoning from Examples

Prototype models classify occurrences based on similarities to previously classified instances. Two papers on computer vision prototype models presented neural models that gain knowledge from prototypes correlating to image parts (Hase et al. 2019). These samples are being used to construct classifier features that are interpretable firsthand.

5.9.7 Latent Space Traversal

These methodologies navigate a model's dormant area to demonstrate how the model behaves when its input varies. Reaching the decision threshold in a classification setting could expose conditions required for a model's estimation of the native input. There are various techniques for developing a vision prototype (Joshi et al. 1806).

5.10 Explainable AI Stakeholders

A total of four stakeholders were considered in Explainable AI namely: Developers, Theorists, Ethicists, and Users (Preece et al. 1810).

5.10.1 Developers

Persons that are interested in developing AI applications. Many participants of this class provide their duties in industry, huge organisations, small and medium enterprises, and government, however, some are academicians or scholars who develop systems that can be used for a multitude of purposes, such as to help them with their work. Both the terms 'explainability' and 'interpretability' are used in this domain. Their key motivation for achieving explainability/interpretability is to improve the efficacy of their programs by offering aid to the testing process, troubleshooting, and review (Preece et al. 1810).

5.10.2 Theorists

Persons are enthusiastic about gaining knowledge and improving AI theory, notably in the domain of deep neural networks. Individuals are usually employed at scientific or corporate research facilities. Most are engaged practitioners; however, theorists

vary from developers in that their prime task is to improve the recent advancements in AI instead of providing real-world applications. The term 'interpretability' is more frequently used among theorists than 'explainability' (Preece et al. 1810).

5.10.3 Ethicists

Decision-makers, reporters, and reviewers who are worried about AI systems' justice, responsibility, and clarity. This group includes numerous computer engineers and technologists, but it also includes public scientists, attorneys, media people, bankers, and legislators. Members of this group also use the terms 'intelligibility' and 'legibility' in addition to 'explainability' and 'interpretability' (Preece et al. 1810).

5.10.4 Users

Individuals who employ artificial intelligence (AI) systems. The extensive mass of persons who participate in the growing compositions on AI explainability/interpretability belong to the first three circles, but this fourth one does not. Users require justifications to allow them to understand whether or not to respond in reaction to the outputs of the system, and/or to defend their decisions. This circle encompasses both real-life applications by users and engaged in procedures that are influenced by AI (Preece et al. 1810).

5.11 Applications

In this section, we have discussed some of the real-world implementations of Explainable Artificial Intelligence.

5.11.1 XAI for Training and Tutoring

After-action review (AAR) tools are common in combat training assistance, allowing trainees to reflect on their exercises and, presumably, learn from them. Mission statistics, a list of completed and failed objectives, and, in certain cases, a mission replay feature are all common aspects of these tools. It has been challenging for users of such AAR tools to grasp how their directions transfer into the behaviours of computer-controlled things with such constrained AAR tools due to increasingly complicated artificial intelligence (AI) in such training aids. Student users additionally have the drawback of having fragmented and sometimes incorrect domain knowledge: they

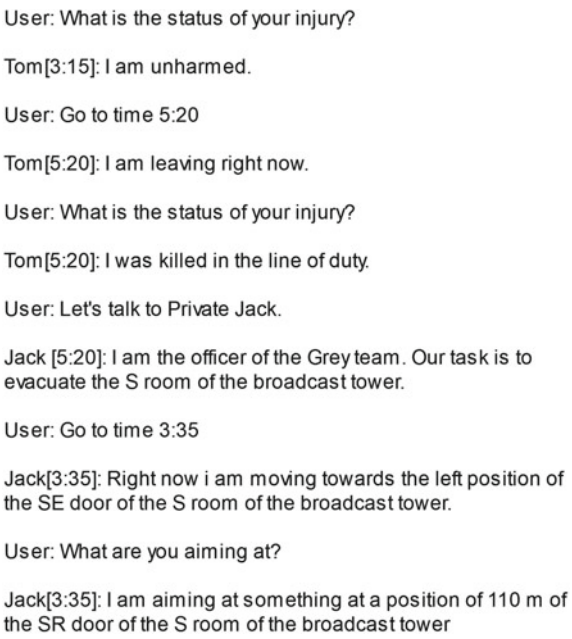


Fig. 5.3 XAI dialogue demo (Lane et al. 2005)

must not only learn new strategic knowledge and abilities but also comprehend emergent behaviors and their causes. They present a mechanism that allows entities in a strategic game to respond to questions regarding the actions. They explain continuing efforts to incorporate a smart instructor into the XAI framework and show how XAI may be used to deliver more meaningful after-action reviews (Lane et al. 2005). Figure 5.3 shows the XAI dialogue demo.

5.11.2 XAI for 6G

Explainable Artificial Intelligence (XAI) for 6G was described in detail, including public and legal incentives, criteria of understandability, efficiency versus understanding of research commutation, strategies to ameliorate explainability, and a framework for incorporating XAI into future wireless systems (Guo 2020). Figure 5.4 Illustrate XAI connectivity with UAV surveillance.

5.11.3 XAI for Network Intrusion Detection

Deep neural networks were utilized to detect network intrusions, and an explainable AI framework was suggested to give transparency to the machine learning pipeline

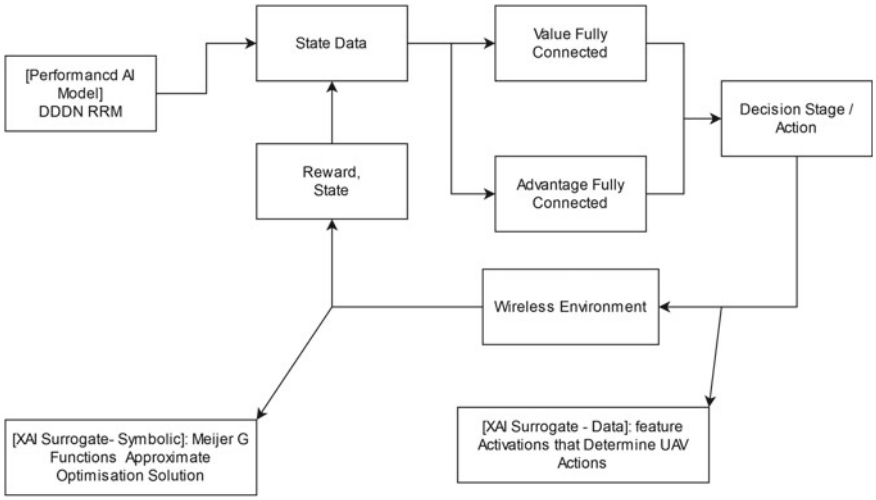


Fig. 5.4 Illustration of XAI connectivity with UAV surveillance (Guo 2020)

at every level. This is accomplished through the use of Explainable AI algorithms, which aim to make machine learning models not much like a black box by offering clarifications for why a prediction is made. Interpretations provide us with quantifiable data on which attribute impacts the likelihood of a cyber-attack or to what scale (Mane and Rao 2103).

5.11.4 XAI Planning as a Service

Explainable Planning can be implemented as a service, i.e., as a shell over a prevailing regulatory structure that allows the use of the existing planner to help answer incompatible questions. A prototype framework was introduced to help with this, as well as some instances of how a planner might be used to solve different types of incompatible issues. The key benefits and drawbacks of such an approach were then reviewed, followed by a questionnaire for Explainable AI Planning as a service that identified various possible research directions (Cashmore et al. 1908). Figure 5.5 illustrate the architecture of Explainable Planning as a service.

5.11.5 XAI for Prediction of Non-Communicable Diseases

Suggested a deep neural network framework based on Deep Shapley Additive Explanations (DeepSHAP) and supplied with a methodology for feature extraction for NCD prediction and explanation in the inhabitants. The DeepSHAP approach has

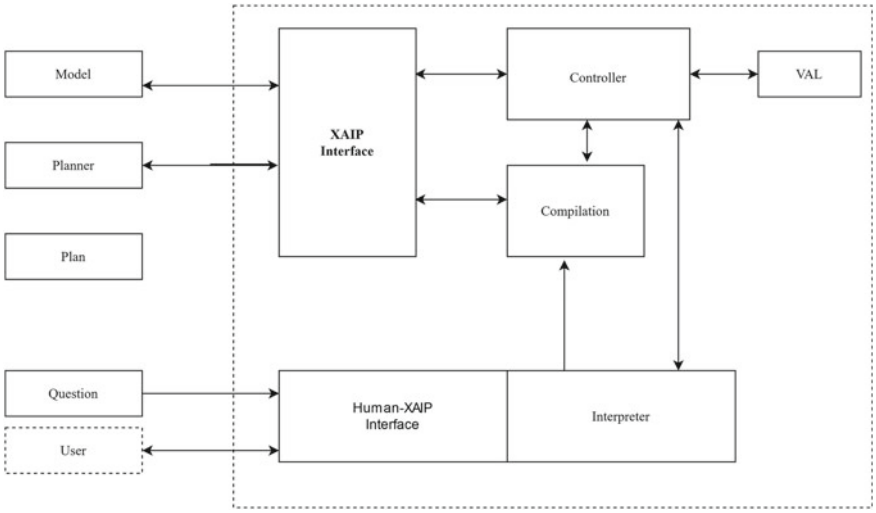


Fig. 5.5 Architecture for explainable planning as a service (Cashmore et al. 1908)

three constituents: first, the indicative characteristic is selected utilizing a flexible net-based immersed feature extraction approach; second, a deep neural network classifier is regulated with hyper-parameters and used to validate the algorithm with the chosen attributes batch; and third, the DeepSHAP approach provides two types of prototype elaboration. The developed scheme surpasses several existing models. Furthermore, the suggested model can aid the medical diagnosis of NCDs by offering a broad perception of changes in disease prospects at both the regional and international levels. The test findings show that key criteria that should have functioned to develop a confidence AI framework to differentiate between sufferers with COVID-19 signs and other sufferers may be interpreted using LIME (Davagdorj et al. 2021). Figure 5.6 shows the detection and understanding of noncommunicable illnesses using exploratory.

5.11.6 XAI for Scanning Patients for COVID-19 Signs

The objective of this project is to create a deep learning-based concept that can accurately recognize COVID-19 sufferers on a CT scan and a chest X-ray image collection. Eight alternative deep learning algorithms were updated and evaluated on two datasets: one with four hundred CT scan images and the other with four hundred chest X-ray images in this study. With a Ninety Five percent confidence interval, NasNetMobile surpassed those other models entitle of performance on CT scan (81.5–95.2%) and chest X-ray (95.4–100%) image datasets. In addition, the model's interpretability is simplified using Local Interpretable Model-agnostic Explanations

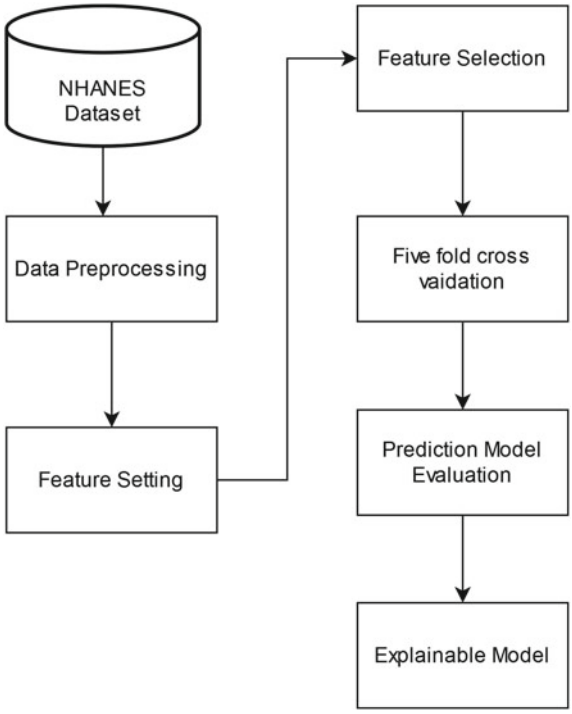


Fig. 5.6 Detection and understanding of noncommunicable illnesses using exploratory (Joshi et al. 1806)

(LIME) (Ahsan et al. 2007). Figure 5.7 displays the complete process flow of the study.

Many frameworks established in the subject of Explainable Artificial Intelligence are discussed in this section.

A design and assessment framework is suggested for end-to-end explainable artificial intelligence systems composition, as demonstrated through a prototype and suggestions, that integrates design goals with evaluation techniques (Mohseni et al.

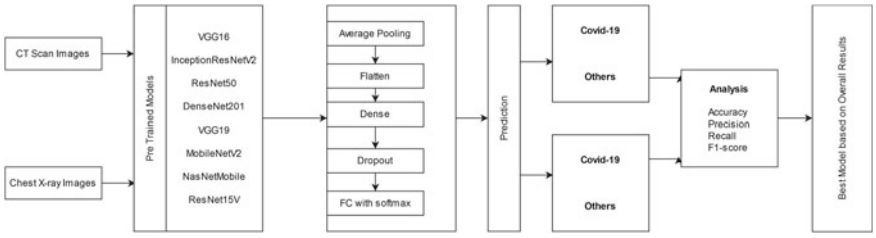


Fig. 5.7 Complete study process flow (Preece et al. 1810)

2021). A new functional explanation framework was developed, which differed from the majority of XAI's previous explanation frameworks. A functional explanation is evaluated using three criteria: correlation, completeness, and complexity. The correlation ensures that the explainer's information can bridge the gap between the explainer and the explainee. The level of correlation is measured using Pearl and Mackenzie's ladder of causation. Explainer information should be complete to describe the underlying systems. Six requirements were provided for a detailed description.

The explanation should be as straightforward as possible for the explainee, albeit this criterion is less crucial than correlation and completeness. The current state of XAI approaches was then assessed using these three variables. According to the data, the rule-extraction approach could provide the greatest standard of explanation on the hypothetical level in terms of correlation (Cui et al. 2019).

An explainable AI framework for imaging air filters was described, which can be expanded to other imaging applications including damage, flaws, or abnormalities. A feedforward neural network was used to build a quantifiable length pseudometric, which is then used as an exclusionary predictor for spatial bootstrap unorganized picture data. After that, this classifier is utilized as a forecaster in a Bayesian inference/regression model to forecast air filter erode degree with a 95% confidence level. The AI model's explainability was attained by using a "twin" model having understandable insights components and correlating them with the AI model's components. This explainable AI model could be used in a range of applications that use artificially produced structured and non—structured imagery in predictive nurturing and health management (Krishnamurthy et al. 2020).

A normative paradigm for Explainable Artificial Intelligence was designed to solve the Black Box Problem. This normative paradigm not only demonstrates the utility of the analytic approaches being produced in Explainable AI but rather they possess some restrictions. Diagnostic classification and feature-detector-identification approach, like input heatmapping, function best when the system's variables can be evaluated logically. Although semantic applicability is typically desirable, it is not always necessary, according to the framework study (Zednik 2021).

A novel theoretical framework to unite exploration and methods evolved in the domain of explainable AI (XAI) was proposed to solve the increasing diversity and absence of cooperation on what defines a comprehensible or explainable model was introduced to address the growing diversity and deficiency of agreement on what establishes an explainable or interpretable model. Two key concepts, "explanation" and "interpretation," underpin this paradigm. These concepts are further framed within a generic workflow that restricts other crucial semantics such as input/output realms and creates a distinction between low-level mathematical concepts and the high-level, human-understandable arena of non-functional constraints. Furthermore, the framework was used to demonstrate how it might aid in the evaluation of existing XAI approaches, demonstrating the amount by which each addresses distinct components of the explainability process (Palacio et al. 2021).

5.12 Possible Research Ideology and Discussions

In this section, we look at new work opportunities for XAI and spot possible scientific pathways that can be adopted to tackle them efficiently shortly.

An examination of the XAI literature reveals that, given the relatively latest and multidisciplinary nature of XAI, an organized approach to core ideas has still not been properly developed. The latest research, on the other hand, is supplying more detailed concepts that are causing a significant influence and forming the domain.

The field's infancy is indeed noticeable in the strategies to develop and construct XAI frameworks and outcomes that have been motivated by devs. Using XAI to reduce prejudices, guarantee social responsibility, and equity requires much data preprocessing and modeling techniques as it does demonstrate that it is properly considered. As a developer resource, XAI can be extremely useful in ensuring that a model relies on causative factors rather than commonalities, that data is equitably spread, and that the functionalities used are appropriate. Possessing the subject matter expertise, that a practitioner may have had in his domain is never sensible, but it is supposed to create sound models. Again, more cross-disciplinary research in XAI is required to make sure that it can collude across competent realms and acquire competence from all relevant areas. When it comes to the use of XAI in Artificial Based architectures, the main discussions show that organizations must take precautions when selecting an XAI approach for addressing a particular requirement. To produce XAI concepts that fulfill regulators and enforce adherence, particularly in healthcare and finance, showed a strong tier of involvement, as this is a considerable challenge in incorporating ML in these immensely governed aspects. The writings strongly imply that XAI will reduce this restriction; even so, no considerable evidential research has been conducted to demonstrate how XAI may fulfill regulators' demand to maintain adherence or undertake evaluations/audits.

How to validate interpretations and associated paradigms is again an evolving trend in XAI. On the one side, investigators should guarantee that sentient explanation is accepted, while on the other side, frameworks must display details as it is, without obscuring the metrics and instead of constructing convincing outcomes.

Conceptual debates about sentient interpretations are presently influenced by sociological viewpoints, with little regard for what tech is proficient in providing. This strengthens the social-technical duality, resulting in a highly fragmented XAI domain centered on either how we as beings analyze interpretations or how one can technologically retrieve features from complicated models. While the present state is warranted by the domain's infancy, more research is necessary to clarify the various XAI stakeholders' necessities and how they can be met with much more aimed interpretations than the two presiding clusters of devs or users.

Whilst the research discussion is far from conclusive, it has identified two directions for future XAI studies:

- The need to contribute to the key parties and their various explanatory requirements. This research direction emphasizes the significance of examining the micropolitics of XAI in groups and its impact on tasks.

- The demand for a systematic perspective in researching XAI, keeping in mind both the sociotechnical elements of XAI, as well as the procedure and result factors of XAI, as well as the credible and narrative components of XAI. This research path underlines the value of conceptualizing and scientifically understanding the multidimensionality of XAI a novel type of socio-technical system, as well as its significance for AI techniques in society and business.

5.13 Conclusion

In the last decagon, as principled considerations, laws, and the need to govern these models have grown, the desire to unlock the renowned 'AI black box' has gained a lot of traction. On the XAI, we conducted a rigorous literature review. We begin by determining the evaluation procedures. Second, the techniques of explanation were discussed, followed by the stakeholders, applications, and frameworks. From the viewpoint of the stakeholders, it is considered that this field necessitates more investigation. It is also brought to our knowledge that black box approaches are unsuitable for some applications, such as in the health world, where a system's incorrect calls might be extremely dangerous. Explainability was also emphasized as a requirement for resolving legal issues that have arisen as a result of the expanding use of AI systems. We anticipate that it will enhance community awareness of the primary categories among scholars (methods, applications, frameworks, etc.).

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Ahsan, M.M., Gupta, K.D., Islam, M.M., Sen, S., Rahman, M., Hossain, M.S., et al.: Study of different deep learning approach with explainable AI for screening patients with COVID-19 symptoms: using CT scan and chest X-ray image dataset (2020). arXiv preprint arXiv:2007.12525

Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (2019)

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. fusion **58**, 82–115 (2020)

Attaran, M., Deb, P.: Machine learning: the new 'big thing' for competitive advantage. Int. J. Knowl. Eng. Data Min. **5**, 277–305 (2018)

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**, e0130140 (2015)

Bilgic, M., Mooney, R.J.: Explaining recommendations: satisfaction versus promotion. In: Beyond Personalization Workshop, IUI (2005)

Bobko, P., Barelka, A.J., Hirshfield, L.M.: The construct of state-level suspicion: a model and research agenda for automated and information technology (IT) contexts. Hum. Factors **56**, 489–508 (2014)
Bryan, N., Mysore, G.: An efficient posterior regularized latent variable model for interactive sound source separation. In: International Conference on Machine Learning (2013)
Cashmore, M., Collins, A., Krarup, B., Krivic, S., Magazzeni, D., Smith, D.: Towards explainable AI planning as a service (2019). arXiv preprint arXiv:1908.05059
Choo, J., Lee, H., Kihm, J., Park, H.: iVisClassifier: an interactive visual analytics system for classification based on supervised dimension reduction. In: 2010 IEEE Symposium on Visual Analytics Science and Technology (2010)
Costanza, E., Fischer, J.E., Colley, J.A., Rodden, T., Ramchurn, S.D., Jennings, N.R.: Doing the laundry with agents: a field trial of a future smart energy system in the home. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2014)
Cui, X., Lee, J.M., Hsieh, J.: An integrative 3C evaluation framework for explainable artificial intelligence (2019)
Davagdorj, K., Bae, J.-W., Pham, V.-H., Theera-Umpon, N., Ryu, K.H.: Explainable artificial intelligence based framework for non-communicable diseases prediction. IEEE Access **9**, 123672–123688 (2021)
Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). arXiv preprint arXiv:1702.08608
Gautam, A., Chatterjee, I.: An overview of big data applications in healthcare: opportunities and challenges. In: Knowledge Modelling and Big Data Analytics in Healthcare, pp. 21–36 (2021)
Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (2018)
Guo, W.: Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun. Mag. **58**, 39–45 (2020)
Hase, P., Bansal, M.: Evaluating explainable AI: which algorithmic explanations help users predict model behavior? (2020). arXiv preprint arXiv:2005.01831
Hase, P., Chen, C., Li, O. Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2019)
Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A.: Trust in automation. IEEE Intell. Syst. **28**, 84–88 (2013)
Hoffman, R.R.: Theory → concepts → measures but policies → metrics. In: Macrocognition Metrics and Scenarios, pp. 3–10. CRC Press (2018a)
Hoffman, R.R.: Theory concepts measures but policies metrics. In: Macrocognition Metrics and Scenarios, pp. 3–10. CRC Press (2018b)
Höök, K.: Steps to take before intelligent user interfaces become real. Interact. Comput. **12**, 409–426 (2000)
Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Mach. Learn. **95**, 423–469 (2014)
Joshi, S., Koyejo, O., Kim, B., Ghosh, J.: xGEMs: generating examplars to explain black-box models (2018). arXiv preprint arXiv:1806.08867
Kay, M., Kola, T., Hullman, J.R., Munson, S.A.: When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016)
Kim, J., Seo, J.: Human understandable explanation extraction for black-box classification models based on matrix factorization (2017). arXiv preprint arXiv:1709.06201
Krishnamurthy, V., Nezafati, K., Stayton, E., Singh, V.: Explainable AI framework for imaging-based predictive maintenance for automotive applications and beyond. Data-Enabled Discov. Appl. **4**, 1–15 (2020)
Kumar, A., Chatterjee, I.: Data mining: an experimental approach with WEKA on UCI Dataset. Int. J. Comput. Appl. **138** (2016)

Kumar, D., Mehta, M.A., Chatterjee, I.: Empirical analysis of deep convolutional generative adversarial network for ultrasound image synthesis. Open Biomed. Eng. J. **15** (2021)
Lane, H.C., Core, M.G., Van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring (2005)
Li, J., Chen, X., Hovy, E., Jurafsky, D.: Visualizing and understanding neural models in NLP (2015). arXiv preprint arXiv:1506.01066
Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue **16**, 31–57 (2018)
Liu, M., Liu, S., Zhu, X., Liao, Q., Wei, F., Pan, S.: An uncertainty-aware approach for exploratory microblog retrieval. IEEE Trans. Vis. Comput. Graph. **22**, 250–259 (2015)
Lundberg, S., Lee, S.-I.: An unexpected unity among methods for interpreting model predictions (2016). arXiv preprint arXiv:1611.07478
Madsen, M. Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems (2000)
Mane, S., Rao, D.: Explaining network intrusion detection system using explainable AI framework (2021). arXiv preprint arXiv:2103.07110
Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I trust it, but I don't know why: effects of implicit attitudes toward automation on trust in an automated system. Hum. Factors **55**, 520–534 (2013)
Meyerson, D., Weick, K.E., Kramer, R.M., et al.: Swift trust and temporary groups. In Trust in Organizations: Frontiers of Theory and Research, vol. 166, p. 195 (1996)
Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TIIS) **11**, 1–45 (2021)
Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. **116**, 22071–22080 (2019)
Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., Dengel, A.: Xai handbook: towards a unified framework for explainable AI. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S.: Stakeholders in explainable AI (2018). arXiv preprint arXiv:1810.00184
Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
Robnik-Šikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. In: Human and Machine Learning, pp. 159–175. Springer (2018)
Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: learning important features through propagating activation differences (2016). arXiv preprint arXiv:1605.01713
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps (2013). arXiv preprint arXiv:1312.6034
Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)
Villata, S., Boella, G., Gabbay, D.M., Van Der Torre, L.: A socio-cognitive model of trust using argumentation theory. Int. J. Approx. Reason. **54**, 541–559 (2013)
Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: Proceedings of Visualization 1995 Conference (1995)
Zednik, C.: Solving the black box problem: a normative framework for explainable artificial intelligence. Philos. Technol. **34**, 265–288 (2021)

Chapter 6
Explainable AI and Its Applications in Healthcare

Arjun Sarkar

6.1 Introduction

Due to the lack of high-end graphics or tensor processing units, previously, deep neural networks could not be implemented as state-of-the-art Artificial Intelligence (AI) algorithms. Rather, linear models were preferred, and they were easy to understand and interpret. Things started changing with the advent of more advanced processing units, in the last decade, when the algorithms took on real-world problems. The models began getting bigger and better. While this highly improved the model performances, this also led to a problem of model interpretability. With access to large datasets, such as ImageNet (Krizhevsky et al. 2017), and these larger non-linear models with millions of parameters, AI soon started taking on human performance, on certain tasks. In 2015, a deep learning model called ResNet (He et al. 2016) surpassed human accuracy at the ImageNet challenge. Soon, AI was implemented in real-world tasks, and companies, industries, and research facilities started adopting AI into their workflows.

AI has now become a part of our day-to-day lives. AI algorithms not only help with day to day tasks such as finding outlines in today's phone cameras or recommending movies on Netflix, but also take on more challenging tasks such as beating human beings at strategy games (Lee et al. 2016; Garisto 2019) and surpassing human beings in complex visual recognition tasks (Bartolo et al. 2020; Thompson and Baker 2021). The rise of deep learning algorithms (Lecun et al. 2015) and computational power over the years has led to this extreme advancement in AI.

Healthcare costs are on the rise all around the globe. AI in medicine and healthcare can reduce this cost and, at the same time, improve healthcare (Higgins and Madai

A. Sarkar (✉)
Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Jena, Germany
e-mail: arjun.sarkar786@gmail.com; arjun.sarkar@leibniz-hki.de

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_6

2020). Even though we read online how an AI beats doctors at predicting a particular disease every other day, the reality of acceptance of AI is still low in healthcare. While there are many reasons for this, one of the most prevalent ones is the explainability of the AI model. The 'black-box' nature of deep learning models is yet to be fully understood, and this causes a lack of trust and transparency. One error by an AI algorithm can be fatal for a patient in a hospital. Thus, the healthcare sector is cautious about implementing AI without completely understanding these algorithms. Most AI software implemented in hospitals today only helps diagnose and aid the doctor in making decisions. The accepted AI software goes through many regulations before being implemented in a hospital environment.

To build trust and reliability on these 'black-box' models, a new research field has emerged in recent years—eXplainable Artificial Intelligence (XAI). This field focuses on interpreting AI models and aims to provide an understandable way to explain AI predictions. At the rise of the deep learning era, most of the research was focused on improving model performance without caring much about explainability. But, that trend is now changing, with many researchers and companies looking to provide high AI accuracies and increased interpretability of the AI models.

A question may arise here: Why can an AI with high accuracy be trusted blindly? Initially, since the AI gives high accuracy, it may seem that the model can be implemented in a real-world situation. But many studies have shown that AI does not always learn the things that humans want it to learn (Lapuschkin et al. 2019). In the PASCAL VOC challenge (Everingham et al. 2010), it was often noticed that the AI was not precisely detecting the object of interest but making its classification based on context (Lapuschkin et al. 2016). For example, a classifier was often noticed to predict images of horses based on the watermark on the pictures and not on the actual horses. Similarly, the algorithm predicted an image as a train, not based on the train itself but the railway tracks (Lapuschkin et al. 2016). So, even though the model gave good accuracy, the correct predictions were often based on some artifacts. Usually, people can't comb through thousands of images on these big data challenges and figure out artifacts. So, these errors mostly go unnoticed. But these overfitting errors occur more often than expected. While this model trained on the PASCAL VOC dataset may perform overwhelmingly well on the test dataset, as the test data also belongs from the same distribution of images as the training data, the same model may fail miserably when tested on real-world data. This is just one of the many examples which fosters the need for the explainability of AI algorithms.

Sometimes, explainability is not about the end results but some intermediate learning. Deep learning algorithms have the power to find interesting patterns from images or text, which may be unknown to a human expert. When Deepmind's AlphaGo AI defeated the rank one human, Lee Sedol, at the game of Go, it played certain moves that other Go experts termed as 'not-human' (Thompson and Baker 2021). This just meant that a human being would not make that move, or that move was previously unknown to humans. Similarly, these algorithms can find patterns in medical images or correlate specific genes with certain diseases previously unknown to health experts. In the scientific and healthcare field, this can prove to be revolutionary. Often scientists and doctors focus more on certain patterns and features than

the final prediction, as those intermediate patterns can lead to new scientific discoveries. Due to the ability of deep learning algorithms to find patterns, these models have had massive success in the field of medicine (Ching et al. 2018; Piccialli et al. 2021), drug discovery (Chen et al. 2018; Gawehn et al. 2016), protein studies (Wang et al. 2017; Xu 2019), neuroscience (Marblestone et al. 2016; Richards et al. 2019), and radiology (Miotto et al. 2017; Kermany et al. 2018; Kuenzi et al. 2020).

The first part of the chapter looks at explainability from different aspects—the multidisciplinary nature of explainable AI in technological, legal, medical, and ethical aspects. Secondly, several explainability algorithms developed over the years which had significant impact on healthcare are explained in the next section. Finally, applications of these algorithms in real world medical tasks are showcased including the use of XAI in the recent COVID-19 pandemic.

6.2 The Multidisciplinary Nature of Explainable AI in Healthcare

The explainability of AI in the healthcare domain is not always a technological issue. It can often be due to combined medical, legal, or ethical issues (Amann et al. 2020).

6.2.1 Technological Outlook

The main issue of XAI is a technological problem: trying to explain an AI algorithm in a human-understandable form. The AI algorithm itself can achieve this explainability, or different models or methods can be used to describe a trained model (Rudin 2019). While the former can be achieved easily for linear models, the latter is necessary for the larger and more complex deep learning models.

Since the inception of XAI, various methods have been developed to try and explain these deep learning models. The explanation of linear models is always very accurate. But these models have severe performance issues compared to the more complex AI models (Esteva et al. 2019). So, there is a tradeoff between the complexity of a model and its explainability. Not only does model understanding help in understanding the final decision, but it also aids developers in tuning the parameters of the model and increasing performance. The problems of overfitting can be reduced or removed altogether. Researchers at Mount Sinai hospital trained a deep learning model to classify safe and high-risk patients based on X-ray images (Zech et al. 2018). The model produced high accuracies on the test set. But when the same model was tested on hospitals other than Mount Sinai, the model performance decreased. When XAI techniques were applied to the model, it was noticed that the model learned from the metadata of the X-ray machine at Mount Sinai hospital rather than on the actual X-ray images. The model was thus able to distinguish the

pictures easily from that particular X-ray machine but failed on images of other X-ray machines, as the metadata no longer matched. XAI techniques help identify and correct these problems before the model is deployed in a real-world scenario. This makes the model more robust, reduces integration costs, and saves time.

While certain XAI techniques help developers improve model parameters, other techniques help healthcare professionals without in-depth knowledge in programming understand predictions. Pointing out the position of infection in medical images has immense benefits for doctors and helps provide a second opinion when they are in doubt. AI also can find rare diseases that are not often known to even seasoned experts (Schaefer et al. 2020). During the recent COVID-19 pandemic, while many algorithms were developed to classify whether patients were infected or not, very few were deployed on the field due to a lack of proper explainability (Fuhrman et al. 2022).

6.2.2 Legal Outlook

The explainability of AI in healthcare is a legal need in nearly every country. In different sectors, the legal requirement for XAI is dissimilar. XAI is not a must in logistics, and a few errors are admissible. But in public administration or banking, XAI can play a vital role. A person whose loan has been rejected due to an AI model has the right to know the reason behind the rejection. In no other sector is XAI as mandatory as in the healthcare field (Schönberger 2019). This does not come as a surprise, as in healthcare, even one error has the potential of harming human life.

AI in healthcare is used for many applications, such as disease classification and diagnosis (Qiu et al. 2020), anomaly detection, patient positioning, image segmentation (Aslam et al. 2015), image super resolution (Chaudhari et al. 2018), and image registration (Ma et al. 2017; Wu et al. 2013). AI is meant to improve clinical applications and aid doctors, improve the standard of medical development and save patient lives. But to train a robust model, often sensitive patient data is required. These privacy issues must meet all legal requirements, from image acquisition to final prediction. Similarly, in recent years, anti-discrimination and explainability of Ai models have gained momentum (Deeks 2019).

Hospitals don't use the AI algorithm as a computer program, but the algorithm is wrapped in the form of a software with a user-friendly graphical user interface (GUI). It is a requirement by most regulatory bodies in the USA or the European Union to provide a level of transparency of the AI's output (Smith et al. 2020). Though these regulations are rather vague now, with no solid rules for explainability, these rules will supposedly get stricter as more emphasis is made on XAI.

One more budding question is about the awareness of these AI predictions and the disclosure to patients. That is, how much of the decision would be made by the AI and how much by the doctor, and finally, how the final prediction would be disclosed to the patient (Cohen 2020). One fear is that the legal system is not fast enough to keep with the rising pace of AI development. In healthcare, AI-based decisions need

strict laws such that they do not hamper innovation but also protect patients' rights and privacy. When these laws are clearly defined and AI researchers can overcome the problems with XAI, AI will be fast adopted in all healthcare sectors.

6.2.3 Medical Outlook

The medical outlook aims to bring semblance between the need for laboratory-based testing or replacing it entirely with AI-based algorithms. Laboratory testing and medical imaging are the methods that have been used since time immemorial for the proper diagnosis of a disease. These are methods that are understandable by medical experts. In laboratory testing and imaging, doctors access the results and the images and find meaningful patterns that point to certain complications or diseases. On the other hand, when trained on the images and the corresponding infection types, a deep learning model may predict the condition correctly but does not indicate the patterns it uses to come to that prediction. XAI can be helpful here in showing these patterns and can be much faster than even the most trained experts.

Even though an AI algorithm can provide good overall accuracy and low error rates (Weng et al. 2017), the algorithm cannot be perfect because of data inconsistency due to noise and imaging errors. The trained experts need to look at the false positives and negatives, and they cannot always be heavily reliant on AI. AI bias is another such complication that cannot be removed entirely. For example, suppose the training data is sampled from a large population of people from Europe in skin cancer prediction. In that case, the same algorithm will fail if deployed in Africa due to the differences in skin tone and color (Wen et al. 2022).

XAI can be crucial in deciding the amount of disagreement between a medical expert and an AI. The results of XAI in the medical field are often visual representations or textual explanations. These explanations can be beneficial to the medical experts in making a final decision on the diagnosis. Without XAI, the clinician has to choose blindly whether to trust the AI or not, but with XAI, the person can understand why the AI makes that particular decision. If an algorithm keeps performing poorly, the results can be reported to the developers, and the developers can understand the reason for the poor performance using explainability. In case the algorithm works well, the clinicians can trust it better when they understand the reason for its good performance (Cutillo et al. 2020).

6.2.4 Ethical Outlook

As more and more healthcare institutes adopt AI into their framework, certain ethical aspects need to be examined. One of these issues is protecting the autonomy of the patient and informing the patient about the use of the deep learning models for their diagnosis. If a patient is not informed whether a doctor or an AI algorithm predicted

a particular disease, this can hamper the patient's trust towards the doctor. A more critical situation would be if the prediction is a false positive or false negative, and the patient is mistreated. The patient can challenge the institution, and the healthcare facility will not be able to provide a concrete reason for such a prediction. This is one more reason for introducing a proper XAI before using the AI algorithm blindly. A solution to such a scenario can be first asking the patient for their permission to use AI for the diagnosis and later explaining the results of the AI to them.

One more ethical conundrum can arise with the rise of AI in healthcare. If AI systems start taking over more healthcare positions in the future, it can limit clinicians' decision-making rather than enhance them. This situation should be avoided, as the best outcome for a patient is to be not entirely dependent on an AI, but a decision of AI carefully analyzed by a doctor.

6.2.5 Patient Outlook

The patient outlook is a perspective that focuses on patients and considers them as an active part of the healthcare decision process (Baker 2001). This refers to the treatment process in which each patient is provided with a treatment best suited for that individual. The idea is to provide patient-centered medicine. But deep learning models predicting risk may not be able to provide such treatment. As doctors do not understand the inner workings of the models as well, neither can they inform the patient about the reason for the predicted risk. XAI can prove to be helpful and continue maintaining the patient-centered approach.

Similarly, wearable devices and smartwatches can now predict certain risk factors in patients (Bhattacharya and Lane 2016; Mauldin et al. 2018). Previously these devices provided similar treatment and health plans to all users. But recently, most companies have been trying to give each user a different health plan based on their activity, heart rate, and sleeping patterns (Coutts et al. 2020; Nweke et al. 2018). Risk assessment explained in the form of text or visual data builds trust and increases transparency and continues building towards a patient-centered innovation in healthcare.

6.3 Different XAI Techniques Used in Healthcare

There are various explainability methods for AI in medicine, and there are multiple ways to classify these methods. Some such taxonomies are explained below.

6.3.1 Methods to Explain Deep Learning Models

Since deep learning models are all black-box models and cannot be easily explained, most modern research is focused on trying to explain these models. Saliency maps is one such technique very commonly used to interpret convolutional neural networks (Itti et al. 1998). These saliency maps are pixels of the image that the convolutional neural network considers essential to the final prediction. Saliency is represented on the image as a visual heatmap or topography.

There are multiple gradient-based techniques used for the explainability of deep learning models (Simonyan et al. 2014). The gradient-based approach shows how much a change in the input would affect the output. Saliency maps are also based on this gradient technique. The Krizhevsky network (Krizhevsky et al. 2017) beat the previous methods and was considered one of the best gradient-based explainability methods. Another algorithm that improves gradient explainability is the DeepLIFT algorithm (Shrikumar et al. 2017). The DeepLIFT algorithm enhances the previous methods by multiplying the input signal to the gradient. The model's superiority was evident when tested on genomic data and natural images. The algorithm assigns a weighted score to the activation of all neurons in the model and shows some crucial connections or features that the previous models failed to identify. DeconvNets or Deconvolution (Zeiler and Fergus 2014) is another method to understand convolutional neural networks (CNN). Unlike regular convolutional layers that extract features from image pixels, deconvolution does the opposite—mapping features to pixel values. Deconvolution is generally used to understand what a convolutional neural network learns in every convolutional layer.

Class Activation Maps (CAM) (Zhou et al. 2016) and its more advanced counterparts Grad-CAM (Selvaraju et al. 2020) and Grad-CAM++ (Chattopadhay et al. 2018) are some of the most famous interpretability methods used to explain the results of convolutional neural networks. CAM helps to identify important locations of the image trained by a model to predict the class of the image. Activations from the final layer of the convolutional model are concatenated to create a feature vector. The weighted sum of this vector is fed into a SoftMax layer to calculate the final result. The result is displayed as a heatmap. While CAM gave good results, it could not be applied to any convolutional model. To overcome this problem, Grad-CAM was developed. Grad-CAM (Selvaraju et al. 2020) can generate the localization maps for any convolutional neural network, regardless of its shape or structure. Grad-CAM++ (Chattopadhay et al. 2018) further improves Grad-CAM by better visualization of the output maps and better object localization for multi-label classification.

The RISE algorithm (Petsiuk et al. 2019) slightly differs from the CAM algorithms. This algorithm considers each pixel of an image to generate a saliency map. Random masks are multiplied elementwise with the images and fed into the network. The model generates a probability score and a saliency map of the input image, which is obtained by combining the masks.

While many algorithms are trying to explain the results of the convolutional neural network models, some studies suggest that none of these techniques are correct in

interpreting the networks (Kindermans et al. 2018). The authors tested some of these explainability methods on simple linear models, but the methods could not correctly interpret the linear models. Hence, the authors argue that if these techniques cannot even explain simple linear models, is their explanation of large complex non-linear models, correct? They further proposed two additional models, PatternNet and PatternAttribution, which work well on linear models and more complex models.

In Natural Language Processing (NLP), a different method is used for explainability (Lei et al. 2016). Small pieces of the input text are added to the model as input, and the model aims to generate the entire text from these small text fragments. Finally, the generated text provides some context and justification for the generated text in terms of the input text.

LIME (Ribeiro et al. 2016) or Local interpretable model-agnostic explanations is an XAI method that can interpret any black-box model. It is also one of the most famous and commonly used interpretability methods for tabular data, text, and images. LIME can interpret individual predictions of a model. It tweaks the feature values of a single data sample and creates an impact of the tweak on the output. Even though LIME can be a simple and powerful interpretable model, it has certain drawbacks. Some studies have shown that choosing poor parameters can cause the model to give different results and miss many essential features completely. This can be a severe problem when the model is deployed in the field. The DLIME algorithm was proposed to overcome this problem. Random sampling used in LIME is replaced in DLIME by choosing clusters of similar data and selecting the most relevant cluster by running k-nearest neighbors (KNN). The authors of DLIME also proved the superiority of DLIME over LIME by testing it on three separate medical datasets.

Shapely Additive explanations (SHAP) (Lundberg and Lee 2017) is another often used interpretability technique. SHAP is a model inspired by game theory. It computes the importance of each feature for all predictions. The SHAP values are a combination of three important properties, namely, accuracy, missingness, and consistency. The authors demonstrate how SHAP is more intuitive and more human interpretable than other XAI methods. Various other model agnostic models such as Anchors (Ribeiro et al. 2018), DeepSHAP (Chen et al. 2021), Protodash (Kim et al. 2016), Permutation Importance (PIMP) (Altmann et al. 2010), and Contrastive Explanation Methods (CEM) (Dhurandhar et al. 2018) are often used for explainability as well.

In deep learning, attention is a trendy concept (Vaswani et al. 2017). The concept of attention was inspired by how humans pay attention to different parts of images or other data sources to analyze them. The MDNet network was created (Xia et al. 2020) to directly map medical imaging and corresponding diagnostic reports. It contained an image model as well as a language model. Attention mechanisms were used to visualize the detection process. This attention mechanism allowed the language model to discover the predominant and distinguishing features used to map the images and diagnostic reports. This was the first study to use the attention mechanism to gain insight from the medical image data. SAUNet, an interpretable U-Net version (Ronneberger et al. 2015), was created (Sun et al. 2020). It also added a secondary

shape stream to capture important shapes-based information in addition to the regular texture features. An attention module was used in the U-Net decoder. SmoothGrad (Hooker et al. 2019) was used to create spatially and shape attention mappings to visualize the high activation area of the images.

These are some commonly used XAI methods for deep learning models in healthcare. All these models have some significance, but there is no one idea to explain all kinds of text and image data or on all sorts of models. SHAP and its advancements are comprehensive and understandable XAI methods of all the methods. Grad-CAM is commonly used for convolutional model interpretation, even for industrial AI software deployed in hospitals.

6.3.2 Explainability by Using White-Box Models

White-box models are transparent models and are easily understandable or interpretable. This category contains mainly linear models, decision trees, and some complex models that are easy to interpret. Some of these complex models used in medical imaging and healthcare are listed here.

Microsoft came up with an interpretable model for predicting pneumonia risk, which also had great accuracy (Caruana et al. 2015). The authors discussed that while interpretable linear models and decision trees could not give good results, neural nets gave much better results, but at the cost of explainability. High-performance generalized additive models with pairwise interactions (GA2Ms) were proposed and tested on two real medical data case studies. The authors also mentioned that the model could be scaled to work on thousands of patient data without losing accuracy and still being highly interpretable.

Another technique that utilizes Boolean rules to create predictive models was proposed—Boolean Rule Column Generation (Dash et al. 2018). This technique uses easy-to-understand Boolean rules with some clauses and conditions. Humans easily understand these clauses and conditions. GLM or Generalized Linear Rule Models (Wei et al. 2019) use an ensemble of rule-based features. GLMs are easy to interpret and complex simultaneously, as the rules can capture non-linear dependencies. TED or Teaching Explanations for Decisions (Hind et al. 2019) is a framework that tries to produce explanations like a human expert rather than explaining the inner workings of an AI model.

Not much research has been done in the complex white-box model development domain. No white-box model can produce the same high accuracy as deep learning models. The white-box models are also very domain-specific, unlike various computer vision and natural language processing neural networks used on various real-world tasks.

6.3.3 *Explainability Methods to Increase Fairness in Machine Learning Models*

AI models are not just theoretical analysis techniques anymore, but with every passing day, more and more models are adopted in real-world applications. Any discrimination or inequality in these models can potentially impact human lives. The fairness of these AI models is another part of XAI that tackles ethical and social aspects. Usually, bias in the models is checked by implementing the model in a different setting, such as a different demographic, and evaluated. Many techniques developed in recent years focus on tackling the bias and discrimination in these models.

The method of disparate impact testing (Feldman et al. 2015) is a model-evaluation tool that can assess the fairness and accuracy of a model but does not provide any details or insight into the causes of bias. It uses simple experiments to highlight differences between model predictions and errors for different demographic groups. It can also detect biases in terms of ethnicity, gender, marital status, or demographics. Another data preprocessing technique was suggested to remove bias from machine-learning models (Calmon et al. 2017). The authors developed a convex optimization to learn a data representation that meets the three stated goals: controlling discrimination, limiting distortion in individual instances, and preserving utility. Adversarial debiasing (Zhang et al. 2018) is an approach to tackling biases regarding demographic segments in machine-learning systems. It involves selecting a feature about the element of interest and then simultaneously training both the primary and adversarial models. The main model is trained to predict the label. The adversarial model, based on the primary model's prediction for each instance, attempts to predict the segment. The goal is to maximize the main machine learning system's accuracy in correctly predicting the label while minimizing the adversarial ability. Adversarial biasing can be used for both classification and regression tasks.

Many methods to make classifiers aware of discriminatory biases need data modifications or algorithm tweaks (Kamiran et al. 2012). They are also not flexible regarding multiple sensitive features handling and control over performance versus discrimination tradeoff. Two new methods, Reject Option-based Classification and Discrimination-Aware Ensemble were developed to solve these problems.

Counterfactual fairness (Kusner et al. 2017) captures the intuition that a decision that affects an individual is fair if it affects the same person in both the real and counterfactual worlds. The individual would then be part of a different demographic. It was also argued that causality in fairness must be addressed. Consequently, a framework was developed to model fairness using tools of causal inference. The authors state that any measure of causality in fairness should not be based on counterfactuals. It is also essential to ensure that counterfactual causal guarantees can be used. Based on the concept of counterfactual fairness, the proposed framework allows users to create models that can take sensitive attributes that could reflect social biases towards people and compensate accordingly. A recent study (Kearns et al. 2018) found that most machine-learning fairness notions and definitions only focus on predefined social segments. It was also pointed out that while such simple

constraints can force classifiers at the segment level to attain fairness, they could lead to discrimination against sub-segments that contain specific combinations of sensitive feature values. The authors suggested that fairness be defined across an infinite or exponential number of sub-segments. These were then determined using the space of sensitive feature values. An algorithm was developed to produce the fairest distribution of sub-segments over classifiers.

One study (Elisa Celis et al. 2019) pointed out that while recent research has attempted to attain fairness regarding a particular metric, specific metrics have been overlooked. Furthermore, some proposed algorithms lack solid theoretical support. The authors developed a meta-classifier that could handle multiple fairness constraints concerning multiple non-disjoint sensitive elements. Another work pointed out that many existing notions about fairness regarding treatment and impact are too restrictive and strict. This can lead to poor model performance. The authors suggested notions of fairness that were based on the collective preferences of different demographic groups to address this issue. Their concept of fairness, more specifically, tries to define which outcome or treatment the various demographic groups would prefer if given a choice.

Fairness is still a new area of machine learning interpretability. However, the incredible progress made over the past few years has been remarkable. Many methods can ensure fair resource allocation and protect the most vulnerable demographics. Several techniques can be used to manipulate data before training models, make algorithmic changes during training, and adjust post-hoc. However, these methods tend to focus too heavily on group fairness. They often overlook individual-level factors at both the local and global levels, leading to the mistreatment of individuals. A small portion of scientific literature deals with fairness in images or text. This gap is still a significant one that needs to be explored in the future.

6.3.4 Explainability Methods to Analyze Sensitivity of a Model

Interpretability methods are used to evaluate and challenge machine learning models to ensure they are reliable and trustworthy. These methods use some form of sensitivity analysis. Models are evaluated for their stability and their ability to predict the impact of subtle but intentional changes in inputs. Sensitivity analysis may interpret changes in output across a range of examples or just one.

The sensitivity index is a traditional method of sensitivity analysis that represents each input variable using a numerical value. First-order indices measure the contribution of one input variable to the output variation. Second, third, and higher-order indexes measure the interaction contribution between two, three, or more input variables to that output variance. The total-effect indices combine the contributions of higher-order and first-order interactions with the output variance.

Sobol (2001) proposed an output variance sensitivity analysis based on ANOVA decomposition. He suggested using Monte-Carlo methods to approximate sensitivity indices higher and first order. Fourier Amplitude Sensitivity Test (FAST) (Cukier et al. 1973) is a method to improve the approximation of Sobol's indexes. The Fourier transformation converts a multi-dimensional integral to a one-dimensional integrated. These algorithms were further enhanced to an RBD-FAST (random balance designs-FAST) algorithm (Plischke 2010), which improved computational efficiency. Morris's method (Morris 1991) of global sensitivity analysis, also known as the one-step-at-a-time (OAT) method, is another option. Although the Morris method is complete, it can be very computationally expensive. Fractional factorial designs (Saltelli et al. 2008) needed to be developed and used in practice to perform sensitivity analysis more efficiently.

6.4 Application of XAI in Healthcare

There are two main types of explanations for deep neural networks in medical images: those that use standard attribution-based approaches and those that use novel, often domain-specific or architecture-specific methods. Many attribution methods can be used to assign an attribution value, contribution, or relevance to each network input feature. An attribution method determines the importance of an input element to the target neuron, which is often the output neuron for a classification problem. Heatmaps show the arrangement of all input features according to the shape of the input samples. Non-attribution is a methodology that is validated on an issue rather than using separate analyses using pre-existing methods. These included concept vectors, attention maps, return of a similar image, and text justifications.

Some applications of XAI in healthcare are explained in this section. While each healthcare domain has hundreds of studies where XAI has been used, only a few examples from each domain are listed.

6.4.1 Medical Diagnostics

One study (Kavya et al. 2021) proposed a computer-aided framework for allergy diagnosis. They evaluated several ML algorithms and then chose the most effective one using k-fold cross-validation. They developed a rule-based approach to the XAI method by creating a random forest. If-Then rules and explanations representing each path within a tree are extracted using medical data. The computer-aided framework was also deployed on the mobile app by the authors, which can be used to assist junior clinicians in verifying the diagnostic predictions. Another study (Dindorf et al. 2021) suggests an explanation-independent classifier for spinal positions. SVM and radiofrequency were used as ML classifiers. Then, they applied LIME to predict the classification. The authors of another study (El-Sappagh et al. 2021) suggested an

RF model to diagnose and detect Alzheimer's progression. The authors also used SHAP to identify the essential features of the classifier. Next, they used a fuzzy rule-based method. SHAP could provide a local explanation for specific patient diagnosis/progression prediction explanations about feature impacts. The fuzzy rule-based system could also generate natural language forms that can aid patients and doctors in understanding the AI model. One paper suggested an XAI framework to assist doctors in diagnosing hepatitis patients (Peng et al. 2021). The authors compared intrinsic XAI methods such as logistic regression, decision trees, and kNN to the more complex models SVM, XGBoost, and RF. The authors also used the post-hoc methods SHAP and LIME and partial dependence plots (PDP). For chronic wound classification, a CNN model was proposed (Sarp et al. 2021). For XAI, the authors used LIME, which aided clinicians in better diagnosis.

6.4.2 Medical Imaging

Due to their simplicity, attribution-based methods were used in most medical imaging literature. Researchers can efficiently train a neural network architecture that is suitable without making it difficult to explain. They also have access to an attribution model. A pre-existing deep learning model or a custom model can obtain the best results on a given task. The existing model implementation is more straightforward and can leverage transfer learning techniques. In comparison, custom models can concentrate on specific data and avoid overfitting with fewer parameters. Both are useful for medical imaging datasets.

Analyzing the post-model data using attributions can show if the model is learning the right features or if it's learning the wrong features. This allows researchers to adjust the hyperparameters and architecture of the model to get better results with test data and potentially in a real-world setting.

6.4.2.1 Brain Imaging

Different methods were analyzed in a study to compare their robustness in CNN's Alzheimer's classification using brain MRI. The methods that were compared were LRP (Bach et al. 2015) and Guided backpropagation (GBP). The L2 norm was calculated between the average attribution maps for multiple runs to check the repeatability of heatmaps of identically trained models. Because occlusion covers more area, it was an order of magnitude lower than the baseline occlusion. LRP performed better than all other methods, indicating a fully attribution-based method. LRP also had the highest similarity in the sum, density, and gain (sum/density), for the top 10 regions across all attributions. Another study (Pereira et al. 2018) used GradCAM and GBP to examine the clinical coherence between the features learned from a CNN for automatic grading brain tumors using MRI. Both methods activated the tumor and surrounding ventricles, which could indicate malignancy. They were both correctly

graded in cases. This focus on non-tumor areas and spurious patterns in GBP maps can lead to errors that indicate unreliability.

6.4.2.2 Breast Imaging

SmoothGrad and IG were used to visualize features in a CNN for classifying estrogen receptor status using breast MRI (Papanastasopoulos et al. 2020). The model learned relevant features from both dynamic and spatial domains, with each contribution. Visualizations showed that the model had learned some irrelevant features from pre-processing artifacts. These observations led us to make changes in our pre-processing and training methods. A previous study to classify breast mass from mammograms (Hassan et al. 2020) used two different CNNs, AlexNet (Szegedy et al. 2015) or GoogleNet (Krizhevsky et al. 2017)—and used saliency maps for visualizing image features. Both CNNs were able to detect the contours of the mass, which is the essential clinical criteria. They also showed sensitivity to context. In another study (Amoroso et al. 2021), the authors also presented an XAI framework to help breast cancer patients. The framework was used to identify a patient's most important clinical feature.

6.4.2.3 Skin Imaging

GradCAM and KernelSHAP were used to compare the features of a set of 30 CNN models trained for melanoma detection (Young et al. 2019). GradCAM and Kernel SHAP were used to compare the features of a suite of 30 CNN models trained for melanoma detection. It was found that even high-accuracy models would sometimes focus on features that were not relevant to the diagnosis. The attribution maps of both methods showed differences in the models' explanations. This demonstrated that different neural network architectures learn various features. A further study (Molle et al. 2018) showed how CNN features were used to classify skin lesions. By scaling the feature maps of activations to the input size, the features for the two last layers were visualized. The layers looked for indicators such as lesion borders, color irregularities, and risk factors such as lighter skin or pink textures. However, some spurious features such as hair and artifacts had no significance.

6.4.2.4 X-ray Imaging

Some studies have used attribution-based diagnostic methods in addition to the more popular imaging modalities. These include both image inputs and non-image inputs. One study used CNNs to perform uncertainty and interpretability analyses on colorectal polyps (Wickstrøm et al. 2020). This is a precursor to rectal cancers. CNN used GBP to create heatmaps. They were found to use the shape and edge information to make predictions. The uncertainty analysis also revealed higher levels of

uncertainty in samples that were misclassified. The authors (Lundberg et al. 2018) presented a model that uses SHAP attributions for hypoxemia. This study was done to analyze preoperative and in-surgery factors. The resulting attributions were consistent with known factors such as BMI, physical state (ASA), tidal volumes, inspired oxygen, and others.

Attribution-based methods were the first method of visualizing neural networks. They have evolved from simple gradient-based class activation maps to more advanced techniques such as Deep SHAP. These visualizations show that models are learning relevant features in most cases. Any spurious features were flagged and corrected by the readers. The identification of relevant features can be improved by smaller models and custom variants to the attribution methods.

6.4.2.5 CT Imaging

DeepDreams inspired attribution method (Mordvintsev et al. 2015) was presented in (Couteaux et al. 2019) to explain the segmentation and classification of tumors from liver CT images. Based on the DeepDreams concepts, this innovative method can be applied to black-box neural networks. The algorithm performed a sensitivity assessment of the features and maximized the activation of target neurons by performing gradient ascent. Comparing networks trained on synthetic and real tumors showed that the former was more sensitive than the latter to clinically relevant features. At the same time, the latter was also more focused on other features. The network was sensitive to both intensity and sphericity with domain knowledge.

6.4.2.6 Retina Imaging

As a diabetic retinopathy (DR) tool, grading by ophthalmologists, a system that produced IG heatmaps and model predictions was investigated (Sayres et al. 2019). The assistance provided by the system was shown to improve the accuracy of the grading over that of an expert without any help or the model predictions. Although the initial grading process was slower, users soon found that it improved their grading experience. This is especially true when heatmaps and predictions are used. Patients without DR saw a decrease in accuracy when model assistance was used. Expressive gradients (EG) were proposed as an extension to IG for weakly supervised segmentation (Yang et al. 2019). Compact CNNs performed better than larger ones, and EG highlighted regions of interest more effectively than traditional IG or GBP methods. EG extends IG by enriching input-level attribution maps with high-level attribution maps. A comparative analysis of various explainability models, including DeepLIFT, DeepSHAP, IG, etc., was performed for a model for detection of choroidal neovascularization (CNV), and diabetic macular edema (DME), and drusens from optical coherence tomography (OCT) scans (Singh et al. 2020).

6.4.3 Surgery

In one study (Kletz et al. 2019), the authors presented a CNN-based medical app to learn the representations of instruments in laparoscopy. They validated their model using different datasets. To help explain how the model classified an instrument, they also provided activation maps from different CNN layers. XAI-CBIR (Chittajallu et al. 2019) was proposed to explain surgical training. XAI-CBIR provides an example post hoc explanation of XAI methods. It extracts representative examples to offer explanations. It uses a self-supervised deep learning model to extract semantic descriptions from MIS video frames. It also used a saliency map to explain visually why it believes the image retrieved is similar to the query. Minimally invasive surgery (MIS) videos can be retrieved using the XAI CBIR system.

6.4.4 Detection of COVID-19

Understanding the COVID-19 data associated with COVID-19 is necessary to fully understand the clinical applications of explainable AI for COVID-19 assessment (Fuhrman et al. 2022). While reverse transcription-polymerase chain reaction (RT-PCR) tests are the most common tool for COVID-19 detection, radiography, and CT scans can supplement RT-PCR testing to improve detection accuracy and throughput. For COVID-19 diagnosis, only X-ray or CT finding may not be sufficient. Therefore, differential diagnosis is difficult because of the subtle differences in COVID-19 and non-COVID-19 pneumonia (Cleverley et al. 2020). For improved differential diagnosis and detection accuracy, explainable AI can be helpful (Dong et al. 2021; Salehi et al. 2020). A standard language for describing COVID-19 can also be used. These datasets are publicly available and can meet data requirements. This includes large-scale projects such as the NIH-funded Medical Imaging and Data Resource Center.

This case study is mainly focused on radiography and chest CT. However, other modalities such as PET/CT, lung ultrasound, and MRI may also play a part in COVID-19 patient care. The development of AI systems to assess COVID-19 is similar to other disease evaluations in many ways. The most common use of explainable AI in COVID-19 assessment is to ensure the model accurately focuses on regions of concern in the input image that indicate disease presence. This is usually done through heatmap visualization. Some studies have had mixed success (Mei et al. 2020; Xiong et al. 2020; Wehbe et al. 2021). One study (Wehbe et al. 2021) shows heatmaps for both negative and positive COVID-19 cases. They note that the negative examples have a low influence on the lungs. Another study (Xiong et al. 2020) shows heatmaps that accurately highlight COVID-19 in lung segmentation and identify regions without significant COVID-19 content to aid the classification decision. This finding is limited in understanding the model's performance and should be investigated further before clinical implementation. COVID-19 can be easily confused

with other diseases such as viral pneumonia. One study differentiates between these, and the results of this study are precise (Jin et al. 2020). The authors divided CT scans into four types of pneumonia and COVID-19. They also identified phenotypic mistakes that were common for humans and AI readers. Grad-CAM and Guided GradCAM were used to visualize the most critical image regions. The authors also provided segmentation for diseased areas. Like previous works, Grad-CAM indicates that the model identifies high-value regions within and outside of the lungs. However, Guided GradCAM does not capture all of the diseased lung tissue. They also use t-SNE to visualize feature embeddings from the various disease classes and identify image features that may be problematic in the classification decision. Another study (Zhang et al. 2020) presents another unique use of explainable AI in the COVID-19 assessment. In this case, the authors use clinical metadata and quantitative lesion features to create classifiers that can predict patient prognosis. They use Shapley numbers to assess how each feature impacts the risk classifier. This includes whether it increases or decreases a prediction output. They also evaluate the effectiveness of different drug administrations and the patient's response to treatment. This type of analysis is beneficial for understanding images indicative of high risk. It is helpful when combined with clinical metadata.

A method called GSInquire was used in a recent study to detect COVID-19 using chest X-ray images (Wang et al. 2020). It produced heatmaps that were used to verify the features of the COVID-net model. GSInquire was created to be an attribution method that performed better than other methods such as SHAP or Expected gradients. It uses the new metrics impact score and coverage. The impact score was the percentage of features that strongly impacted the model decision or confidence. Impact coverage was determined in relation to the inclusion of factors that could be adversely affected. While these studies use python programming to create the deep learning models, the Cognex VisionPro Deep Learning Software classified Covid-19 X-ray images using their deep learning-based graphic user interface (GUI) (Sarkar et al. 2021). The software has built-in Grad-CAM for interpretability, highlighting the regions of interest. A trained medical expert can then look at the Grad-CAM and judge the efficacy of the software.

6.5 Conclusion

Despite its rapid growth, explainable AI is still not a mature field. It often suffers from a lack of formality and poorly defined definitions. Although many machine learning interpretability methods and studies have been developed in academia and other institutions, they do not often form an integral part of machine-learning workflows or pipelines.

This chapter examines the role of explicable AI in clinical decision-support systems from technological, legal, and ethical perspectives.

There are many applications of XAI within the healthcare industry. The concept of explainability has many implications for all stakeholders. Developers, doctors,

and legislators face challenges regarding medical AI. Combining multiple modalities, such as medical images and patient records, is possible to make decisions and attribute model decisions to each one. This could simulate a physician's workflow where images and patient parameters are used to make a diagnosis. This can improve accuracy and provide more detailed explanations. Despite making impressive strides in explaining the diagnosis, there are still many steps to meet regulators and end-users needs.

There is still much to be done in machine learning interpretability methods. Many studies have been conducted over the years and demonstrated many opportunities for improvement. They also highlighted the potential benefits and enhancements these methods bring to existing machine-learning workflows. However, they also revealed their weaknesses and performance limitations. However, we believe that explainable AI still has many unexplored areas and great potential to be explored in the future.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26** (2010). https://doi.org/10.1093/bioinformatics/btq134

Amann, J., Blasimme, A., Vayena, E., et al.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **2** (2020). https://doi.org/10.1186/s12911-020-01332-6

Amoroso, N., Pomarico, D., Fanizzi, A., et al.: A roadmap towards breast cancer therapies supported by explainable artificial intelligence. Appl. Sci. (Switzerland) **11** (2021). https://doi.org/10.3390/app11114881

Aslam, A., Khan, E., Beg, M.M.S.: Improved edge detection algorithm for brain tumor segmentation. Procedia Comput. Sci. (2015)

Bach, S., Binder, A., Montavon, G., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE (2015). https://doi.org/10.1371/journal.pone.0130140

Baker, A.: Book: crossing the quality chasm: a new health system for the 21st century. BMJ **323** (2001). https://doi.org/10.1136/bmj.323.7322.1192

Bartolo, M., Roberts, A., Welbl, J., et al.: Beat the AI: investigating adversarial human annotation for reading comprehension. Trans. Assoc. Comput. Linguist. **8** (2020). https://doi.org/10.1162/tacl_a_00338

Bhattacharya, S., Lane, N.D.: From smart to deep: Robust activity recognition on smartwatches using deep learning. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016 (2016)

Calmon, F.P., Wei, D., Vinzamuri, B., et al.: Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems (2017)

Caruana, R., Lou, Y., Gehrke, J., et al.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings—2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 (2018)

Chaudhari, A.S., Fang, Z., Kogan, F., et al.: Super-resolution musculoskeletal MRI using deep learning. Magn. Reson. Med. (2018). https://doi.org/10.1002/mrm.27178

Chen, H., Engkvist, O., Wang, Y., et al.: The rise of deep learning in drug discovery. Drug Discov. Today **23** (2018)
Chen, H., Lundberg, S., Lee, S.I.: Explaining models by propagating shapley values of local components. In: Studies in Computational Intelligence (2021)
Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface **15** (2018). https://doi.org/10.1098/rsif.2017.0387
Chittajallu, D.R., Dong, B., Tunison, P., et al.: XAI-CBIR: explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In: Proceedings—International Symposium on Biomedical Imaging (2019)
Cleverley, J., Piper, J., Jones, M.M.: The role of chest radiography in confirming covid-19 pneumonia. BMJ **370** (2020)
Cohen, I.G.: Informed consent and medical artificial intelligence: what to tell the patient? SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3529576
Couteaux, V., Nempont, O., Pizaine, G., Bloch, I.: Towards interpretability of segmentation networks by analyzing deepDreams. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2019)
Coutts, L.V., Plans, D., Brown, A.W., Collomosse, J.: Deep learning with wearable based heart rate variability for prediction of mental and general health. J. Biomed. Inform. **112** (2020). https://doi.org/10.1016/j.jbi.2020.103610
Cukier, R.I., Fortuin, C.M., Shuler, K.E., et al.: Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J. Chem. Phys. **59** (1973). https://doi.org/10.1063/1.1680571
Cutillo, C.M., Sharma, K.R., Foschini, L., et al.: Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit. Med. **3** (2020)
Dash, S., Günlük, O., Wei, D.: Boolean decision rules via column generation. In: Advances in Neural Information Processing Systems (2018)
Deeks, A.: The judicial demand for explainable artificial intelligence. C. Law Rev. **119** (2019)
Dhurandhar, A., Chen, P.Y., Luss, R., et al.: Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Advances in Neural Information Processing Systems (2018)
Dindorf, C., Konradi, J., Wolf, C., et al.: Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (Xai). Sensors **21** (2021). https://doi.org/10.3390/s21186323
Dong, D., Tang, Z., Wang, S., et al.: The role of imaging in the detection and management of COVID-19: a review. IEEE Rev. Biomed. Eng. **14** (2021). https://doi.org/10.1109/RBME.2020.2990959
Elisa Celis, L., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: a meta-algorithm with provable guarantees. In: FAT* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019)
El-Sappagh, S., Alonso, J.M., Islam, S.M.R., et al.: A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci. Rep. **11** (2021). https://doi.org/10.1038/s41598-021-82098-3
Esteva, A., Robicquet, A., Ramsundar, B., et al.: A guide to deep learning in healthcare. Nat. Med. **25** (2019)
Everingham et al. 2010Everingham, M., van Gool, L., Williams, C.K.I., et al.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88** (2010). https://doi.org/10.1007/s11263-009-0275-4
Feldman, M., Friedler, S.A., Moeller, J., et al.: Certifying and removing disparate impact. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)
Fuhrman, J.D., Gorre, N., Hu, Q., et al.: A review of explainable and interpretable AI with applications in COVID-19 imaging. Med. Phys. **49** (2022)

Garisto, D.: Google AI beats top human players at strategy game StarCraft II. Nature (2019). https://doi.org/10.1038/d41586-019-03298-6
Gawehn, E., Hiss, J.A., Schneider, G.: Deep learning in drug discovery. Mol. Inform. **35** (2016)
Hassan, S.A., Sayed, M.S., Abdalla, M.I., Rashwan, M.A.: Breast cancer masses classification using deep convolutional neural networks and transfer learning. Multimed. Tools Appl. **79** (2020). https://doi.org/10.1007/s11042-020-09518-w
He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
Higgins, D., Madai, V.I.: From bit to bedside: a practical framework for artificial intelligence product development in healthcare. Adv. Intell. Syst. **2** (2020). https://doi.org/10.1002/aisy.202000052
Hind, M., Wei, D., Campbell, M., et al.: TED: teaching AI to explain its decisions. In: AIES 2019—Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019)
Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems (2019)
Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998). https://doi.org/10.1109/34.730558
Jin, C., Chen, W., Cao, Y., et al.: Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat. Commun. **11** (2020). https://doi.org/10.1038/s41467-020-18685-1
Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Proceedings—IEEE International Conference on Data Mining, ICDM (2012)
Kavya, R., Christopher, J., Panda, S., Lazarus, Y.B.: Machine learning and XAI approaches for allergy diagnosis. Biomed. Signal Process. Control **69** (2021). https://doi.org/10.1016/j.bspc.2021.102681
Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: 35th International Conference on Machine Learning, ICML 2018 (2018)
Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172** (2018). https://doi.org/10.1016/j.cell.2018.02.010
Kim, B., Khanna, R., Koyejo, O.: Examples are not enough, learn to criticize! Criticism for interpretability. In: Advances in Neural Information Processing Systems (2016)
Kindermans, P.J., Schütt, K.T., Alber, M., et al.: Learning how to explain neural networks: PatternNet and PatternAttribution. In: 6th International Conference on Learning Representations, ICLR 2018—Conference Track Proceedings (2018)
Kletz, S., Schoeffmann, K., Husslein, H.: Learning the representation of instrument images in laparoscopy videos. Healthc. Technol. Lett. (2019)
Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM (2017). https://doi.org/10.1145/3065386
Kuenzi, B.M., Park, J., Fong, S.H., et al.: Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell **38** (2020). https://doi.org/10.1016/j.ccell.2020.09.014
Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems (2017)
Lapuschkin, S., Binder, A., Montavon, G., et al.: Analyzing classifiers: fisher vectors and deep neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
Lapuschkin, S., Wäldchen, S., Binder, A., et al.: Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. **10** (2019). https://doi.org/10.1038/s41467-019-08987-4
Lecun, Y., Bengio, Y., Hinton. G.: Deep learning. Nature (2015)
Lee, C.S., Wang, M.H., Yen, S.J., et al.: Human versus computer go: review and prospect [Discussion Forum]. IEEE Comput. Intell. Mag. **11** (2016). https://doi.org/10.1109/MCI.2016.2572559
Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. In: EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Proceedings (2016)
Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (2017)

Lundberg, S.M., Nair, B., Vavilala, M.S., et al.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. **2** (2018). https://doi.org/10.1038/s41551-018-0304-0

Ma, K., Wang, J., Singh, V., et al.: Multimodal image registration with deep context reinforcement learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017)

Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. **10** (2016). https://doi.org/10.3389/fncom.2016.00094

Mauldin, T.R., Canby, M.E., Metsis, V., et al.: Smartfall: a smartwatch-based fall detection system using deep learning. Sensors (Switzerland) **18** (2018). https://doi.org/10.3390/s18103363

Mei, X., Lee, H.C., Diao, K.Y., et al.: Artificial intelligence—enabled rapid diagnosis of patients with COVID-19. Nat. Med. **26** (2020). https://doi.org/10.1038/s41591-020-0931-3

Miotto, R., Wang, F., Wang, S., et al.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. **19** (2017). https://doi.org/10.1093/bib/bbx044

Mordvintsev, A., Tyka, M., Olah, C.: Inceptionism: going deeper into neural networks, google research blog. In: Google Research Blog (2015)

Nweke, H.F., The, Y.W., Al-garadi, M.A., Alo, U.R.: Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst. Appl. **105** (2018)

Papanastasopoulos, Z., Samala, R.K., Chan, H.-P., et al.: Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI (2020)

Peng, J., Zou, K., Zhou, M., et al.: An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. J. Med. Syst. **45** (2021). https://doi.org/10.1007/s10916-021-01736-5

Pereira, S., Meier, R., Alves, V., et al.: Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018)

Petsiuk, V., Das, A., Saenko, K.: RisE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference 2018, BMVC 2018 (2019)

Piccialli, F., di Somma, V., Giampaolo, F., et al.: A survey on deep learning in medicine: why, how and when? Inf. Fusion **66** (2021). https://doi.org/10.1016/j.inffus.2020.09.006

Plischke, E.: An effective algorithm for computing global sensitivity indices (EASI). Reliab. Eng. Syst. Saf. **95** (2010). https://doi.org/10.1016/j.ress.2009.11.005

Qiu, S., Joshi, P.S., Miller, M.I., et al.: Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain **143** (2020). https://doi.org/10.1093/brain/awaa137

Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)

Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 (2018)

Richards, B.A., Lillicrap, T.P., Beaudoin, P., et al.: A deep learning framework for neuroscience. Nat. Neurosci. **22** (2019)

Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2015)

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1** (2019)

Salehi, S., Abedi, A., Balakrishnan, S., Gholamrezanezhad, A.: Coronavirus disease 2019 (COVID-19) imaging reporting and data system (COVID-RADS) and common lexicon: a proposal based on the imaging data of 37 studies. Eur. Radiol. **30** (2020). https://doi.org/10.1007/s00330-020-06863-0

Saltelli, A., Ratto, M., Andres, T., et al.: Global sensitivity analysis: the primer (2008)
Sarkar, A., Vandenhirtz, J., Nagy, J., et al.: Identification of images of COVID-19 from chest X-rays using deep learning: comparing COGNEX VisionPro deep learning 1.0^{TM} software with open source convolutional neural networks. SN Comput. Sci. **2** (2021). https://doi.org/10.1007/s42979-021-00496-w
Sarp, S., Kuzlu, M., Wilson, E., et al.: The enlightening role of explainable artificial intelligence in chronic wound classification. Electronics (Switzerland) **10** (2021). https://doi.org/10.3390/electronics10121406
Sayres, R., Taly, A., Rahimy, E., et al.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology **126** (2019). https://doi.org/10.1016/j.ophtha.2018.11.016
Schaefer, J., Lehne, M., Schepers, J., et al.: The use of machine learning in rare diseases: a scoping review. Orphanet J. Rare Dis. **15** (2020)
Schönberger, D.: Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. Int. J. Law Inf. Technol. **27** (2019). https://doi.org/10.1093/ijlit/eaz004
Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128** (2020). https://doi.org/10.1007/s11263-019-01228-7
Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning, ICML 2017 (2017)
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014—Workshop Track Proceedings (2014)
Singh, A., Mohammed, A.R., Zelek, J., Lakshminarayanan, V.: Interpretation of deep learning using attributions: application to ophthalmic diagnosis (2020)
Smith, J.A., Abhari, R.E., Hussain, Z., et al.: Industry ties and evidence in public comments on the FDA framework for modifications to artificial intelligence/machine learning-based medical devices: a cross sectional study. BMJ Open **10** (2020). https://doi.org/10.1136/bmjopen-2020-039969
Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55** (2001). https://doi.org/10.1016/S0378-4754(00)00270-6
Sun, J., Darbehani, F., Zaidi, M., Wang, B.: SAUNet: shape attentive U-net for interpretable medical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2020)
Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
Thompson, B., Baker, N.: Google AI beats humans at designing computer chips. Nature (2021). https://doi.org/10.1038/d41586-021-01558-y
van Molle, P., de Strooper, M., Verbelen, T., et al.: Visualizing convolutional neural networks to improve decision support for skin lesion classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018)
Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
Wang, L., Lin, Z.Q., Wong, A.: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. (2020). https://doi.org/10.1038/s41598-020-76550-z
Wang, S., Li, Z., Yu, Y., Xu, J.: Folding membrane proteins by deep transfer learning. Cell Syst. **5** (2017). https://doi.org/10.1016/j.cels.2017.09.001
Wehbe, R.M., Sheng, J., Dutta, S., et al.: DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set. Radiology **299** (2021). https://doi.org/10.1148/RADIOL.2020203511

Wei, D., Dash, S., Gao, T., Günlük, O.: Generalized linear rule models. In: 36th International Conference on Machine Learning, ICML 2019 (2019)
Wen, D., Khan, S.M., Xu, A.J., et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit. Health **4** (2022)
Weng, S.F., Reps, J., Kai, J., et al.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE **12** (2017). https://doi.org/10.1371/journal.pone.0174944
Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. **60** (2020). https://doi.org/10.1016/j.media.2019.101619
Wu, G., Kim, M., Wang, Q., et al.: Unsupervised deep feature learning for deformable registration of MR brain images. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2013)
Xia, H., Sun, W., Song, S., Mou, X.: Md-net: multi-scale dilated convolution network for CT images segmentation. Neural Process. Lett. **51** (2020). https://doi.org/10.1007/s11063-020-10230-x
Xiong, Z., Wang, R., Bai, H.X., et al.: Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology **296** (2020). https://doi.org/10.1148/radiol.2020201491
Xu, J.: Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. U. S. A. **116** (2019). https://doi.org/10.1073/pnas.1821309116
Young, K., Booth, G., Simpson, B., et al.: Deep neural network or dermatologist? In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2019)
Zafar, M.B., Valera, I., Rodriguez, M.G., et al.: From parity to preference-based notions of fairness in classification. In: Advances in Neural Information Processing Systems (2017)
Zech, J.R., Badgeley, M.A., Liu, M., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. **15** (2018). https://doi.org/10.1371/journal.pmed.1002683
Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2014)
Zhang, K., Liu, X., Shen, J., et al.: Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell (2020). https://doi.org/10.1016/j.cell.2020.04.045
Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)
Zhou, B., Khosla, A., Lapedriza, A., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)

Chapter 7
Explainable AI Driven Applications for Patient Care and Treatment

Mukta Sharma, Amit Kumar Goel, and Priyank Singhal

Abstract The continuous development of technology has saved countless lives and improved the quality of living. Artificial Intelligence is reshaping the healthcare industry from hospital care to clinical research, drug development, to insurance, and has been able to reduce costs and improve patient outcomes. Most AI system works as a black box with little or no explanation which results in a lack of trust and accountability among patients and doctors. This chapter is written with the intent to share with the audience how exquisitely the health care sector has integrated with the technology. The chapter initiates with a brief description of the use of Artificial intelligence and technology in the health domain, and how computers are helping not only doctors, but patients, health care departments, and Insurance companies. This chapter later focuses on various AI-driven Applications which are used for patient care and treatment. This chapter shed light on the purpose and benefits of XAI along with a few real examples.

Keywords Electronic health record · Artificial intelligence · Machine learning · Deep learning · Explainable artificial intelligence · Clinical decision support system

7.1 General

Society is getting enormous benefits from the advancement and innovation in technology. Technology has become an indispensable aspect of our lives. The continuous

M. Sharma (✉) · A. K. Goel
Delhi, India
e-mail: hod.bca@tips.edu.in

A. K. Goel
e-mail: amit.goel@galgotiasuniversity.edu.in

P. Singhal
Moradabad, India
e-mail: priyank.computers@tmu.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_7

development in technology has saved countless lives and also have improved the quality of living. Technology is proving to be very relevant in the Health care sector, starting with EHR where the records of the patient are maintained electronically, a system to compile patient's medical history, which includes patients' past and present details. These days the data is collected through various devices like smartwatches, bands, patient's email, apps, etc. to provide a better monitoring system that will help in analyzing the health information.

Any IT gadgets or programming designed to enhance emergency clinic and authoritative efficiency, provide new bits of understanding about medications and therapies or improve the general nature of treatment is referred to as medical services innovation. Artificial Intelligence is being used extensively in various domains, like the reviews about the most-watched movie or series on Netflix, the most purchased product on Amazon, traffic congestion on road can be predicted by Google Maps, etc. are a few instances of the use of the AI.

AI technologies are reshaping the healthcare industry from hospital care to clinical research, drug development, to insurance, and have been able to reduce costs and improve patient outcomes. With the use of Artificial Intelligence in the subsequent years, it has been observed that there is a paradigm shift from what technology can do, to how technology should be used responsibly to improve health care services and patients' health. As most the AI system works as a black box; with very little or no explanation it results in a lack of trust and accountability amongst patients and doctors. The results generated by the AI tools could not be cross verified, and in case if they are generating a wrong decision; could lead to disaster specifically when it comes to the health care sector, which involves human life and one wrong decision can ruin a life. Therefore, it is essential and very crucial to use XAI, as it provides an explanation in natural language for a better understanding and rational decision making.

The present medical care industry is a $2 trillion behemoth (https://builtin.com/healthcare-technology). The apps and other health care devices with the use of AI have improved and helped in diagnosis, as it is more comfortable scanning and observing the pattern related to health like heart rate, blood pressure, footsteps took, calories burn/taken, to even monitor the sleep quality, etc. Artificial Intelligence is supporting emergency clinics by taking the decisions for better diagnosis based on analyzing and predicting the data.

With the enhancement in Artificial Intelligence, the medical domain has created robotic surgeries, where actually the physician is not even in the operation theater with the patient. The patient might be at the clinic or hospital in his home town eliminating any stress and hassle of traveling. Robotic surgeries also allow a minimally-invasive procedure that reduces the scars, and pain; which helps the patient cure fast (Bouronikous 2013). AI systems, such as deep learning or machine learning as the name suggests the machine is being trained by taking inputs and later producing outputs with no decipherable explanation or context.

Erik Birkeneder, a medical device, and digital health expert, in an interview with Forbes, "We can't be sure an AI system will discover those outliers or otherwise appropriately diagnose patients if it isn't properly trained with the relevant

data and we don't understand how it makes its decisions" (https://www.capestart.com/resources/blog/how-explainable-ai-for-health-care-helps-build-user-trust/#:~:text=When%20and%20in%20what%20context,undetectable%20by%20the%20human%20eye).

Many of the AI algorithms are really insightful an algorithm to estimate the brain age is based on more than 5000 brain scans using a deep learning algorithm and is good in predicting the age and identifying if someone is getting cognitive decline or dementia and also has the capability to trace back the neural network and the changes in the brain due to the age or any other reason. Similarly, the genetic algorithm works very fine, so the algorithms refer to the image visualization and take the decisions based on the visuals, the results are often correct.

Explainable AI is the need of the hour, though XAI has been in the existence for approximately 40 years. It is gaining good popularity now as people are using Artificial intelligence extensively in almost all domains; especially in medicine where we are dealing with human lives we need to be assured, need to trust the solution/output given by AI system. The human need to understand why this decision has been taken or why this diagnosis has been proposed by the AI system. Designers and developers need the ability to explain to improve system robustness and allow diagnostics to avoid prejudice, injustice, and discrimination, as well as to raise user trust in why and how decisions are made. It is essential to give people a feeling that they can trust the software, the output should be interpretive (predicted or inferred), especially in cases where the images are synthesized. In short, XAI is required where the user needs an explanation to make a decision.

7.2 Benefits of Technology and AI in Healthcare Sector

From the invention of X-ray equipment to advances in surgical techniques, technology has improved our health and prolonged our lives (Hosny et al. 2018), Scherman (2019). Continued developments, and research in innovations that cure illnesses especially using Artificial Intelligence, training the devices to not only collect the data through sensors, and actuators but also to analyze the data using numerous algorithms which can predict with accuracy and precision (Mojsilovic 2019). Technology is helping us start by maintaining and keeping the patient's records handy through EHR (Electronic Health Record) instead of conventional paper-based manual methods. Also, the technology has made it possible to connect through Telemedicine, use remote monitoring health, and also share our health information through wearable and sensors technology with our doctors (Gulavani and Kulkarni 2010). Sequencing the human genome has been one of the greatest advancements in medical technology (McDonough 2021).

The innovation is certainly helping and permitting us to analyze a huge amount of data. The way Amazon has imagined Alexa, which is a virtual assistant based on an AI framework. Alexa helps in responding/answering conversational questions immediately, even in a noisy environment. Another AI application can help examine

the information identified with individuals' wellbeing, permitting us to analyze the patterns they could turn into the key to well-being screening, early analysis, and treatment plan for a patient. The data accumulated by technology and sensors can have various advantages (Bouronikous 2013; Laal 2012; Luci 2015), for example:

- **Reduced healthcare cost and enhanced speed**—With the assistance of innovation, the data provided by the sensors can help in observing the well-being of the patients living at remote places. Real-time cautions can likewise assist a patient with counselling a specialist before health deteriorates. This remote monitoring of heath also eliminates hospital room expenses and staff costs. Splendid advancement can streamline claims planning, and cut costs by a colossal edge (Patel 2022).
- **Reducing healthcare waste**—According to the World Health Organization, healthcare waste accounts for about a quarter of all waste produced. An approximate 16 billion injections are issued per year around the world, however, the disposal of all needles and syringes is not done properly. Measures to guarantee the effective and ecologically sustainable disposal of medical wastes might assist to minimize harmful health and environmental implications. Providers and health insurers should follow these three ways to maximize healthcare costs during a period when 25% of spending is deemed unsustainable, (Thimbleby 2013), WHO/Unicef (2018).
- **Virtual Reality is helping in fighting Depression and Mental Health**—Approximately more than 800,000 peoples commit suicide every year; generally, because of emotional well-being issues. Psychological maladjustment can't be recognized by taking blood tests or breaking down information in an electronic clinical record. The data can suggest subtle traces of problems provided that they are well analyzed. Clinical psychologists and doctors identify the behavior and perform a cognitive test on the patients to know the situation. Especially the patients who have been born with any kind of prior traumas, doctors gradually trained the patient's brains to talk and build up tolerance through exposure treatment, until such memories no longer negatively affect them (Builtin).
- **New medicine and therapies are being developed**—According to a recent report, it takes at least 10 years for a drug to make the journey from discovery to the marketplace, at an estimated cost of $2.6 billion. The probability of a drug entering clinical trials being approved is currently estimated to be less than 12%. Incorporating AI into the drug-discovery process can have a significant impact on the development of safer and more successful drugs (Laal 2012).
- **AI-based Apps as a Scheduler**—To keep a human healthy, two important things need to be done, eat healthily and on time and follow a fitness regime. AI helps in analyzing the medical record and scheduling the fitness routine (Ksolves.com 2021).
- **AI for visually impaired or disabled people—**Many of the researches are going on to help people with any disability. AI-based devices are available in the market from gloves, shoes, etc. which will help and guide the disabled person (Ksolves.com 2021; Patel 2022).

- **AI for old people**—In research conducted in Japan observed that many clinics and old age homes have AI pets to keep the old people engrossed and stay connected with the devices. They have tried to use technology by designing AI robots in form of pets to give emotional support to the patients.

7.3 Most Common AI-Based Healthcare Applications

In recent times, AI has been benefiting the patients, doctors, and admin staff; without human intervention can complete task at a faster pace and are the talk of almost every conversation, especially in the medical domain. It has been observed that AI is often discussed by the modern medical industry to identify and diagnose the disease. There are numerous benefits of using artificial intelligence in today's contemporary medicine, but at the same time there are several issues that are bothering people; one such concern is a miss of the "human touch" in this people-oriented profession where people need to be supported emotionally as well and the trust; the patients have on the doctors mentally strengthen the patient and positive attitude heal the patient faster.

Artificial intelligence (AI) in modern medicine is used to describe the usage of AI software and pre-programmed processes to detect and treat patients that require medical treatment. Besides analysis and cure, there are a number of other processes that must be completed to appropriately take care of a patient designated, which may appear to be trivial tasks, including:

i. Gathering data from patient talks and examinations
ii. Dealing with and analyzing the results
iii. Obtaining precise identification by utilizing a multitude of data sources.
iv. Selecting an acceptable treatment method
v. Organizing and monitoring the treatment plan
vi. Observation by the patient
vii. Rehabilitation and continuing plans.

Healthcare has a wide range of applications, from identifying genetic code connections to powering surgical robots. Predictive, comprehending, reading, and acting machines are being reinvented. Artificial intelligence has been a benefit to the healthcare business in general.

Automation is delivering enormous benefits as creativity continues to flourish. Some applications that help to improve health care accuracy in addition to having a specialist solution that saves time and money in the treatment process (Datta et al. 2019; Nicholson 2019; Pawar et al. 2020).

- **Diagnostic Imaging Interpretation**—Deep learning programs and technology classification is used to provide AI-based imaging systems with algorithms that can read photos swiftly. Buoy Health, one of the most popular AI-based diagnostic checkers, uses an algorithm to help with sickness treatment (Your Team in India 2020). For instance, in the case of Lung cancer screening and to help

detect pulmonary nodules, in many cases, early discovery can save a patient's life. Artificial intelligence (AI) can assist in recognizing and categorizing these nodules as benign or cancerous. Similarly, in the case of abdominal, mammography, brain tumor, and many more cases; artificial intelligence can interpret and evaluate whether they are benign or malignant. In the case of skin cancer, deep learning algorithms are handling and help detect suspicious areas (Hosny et al. 2018).

- **Accuracy**—According to the research diagnosis done by doctors are 71.40% accurate and diagnosis based on AI and ML is 72.52% accurate (Leibowitz 2020). Doctors are adopting a more contemporary strategy that emphasizes prevention and data collecting. This involves genetic data collection, wearable gadgets, and electronic healthcare system developments. Apple watches, Fitbits, Garmin watches, and other fitness trackers monitor your heart rate and activity levels (Luci 2015; Medttech; Your Team in India 2020).
- **Interactive Assistant for Fitness**—AI businesses have created digital health aides that focus on augmented reality, cognitive computing, speech, and body motions. A virtual health assistant is a one-of-a-kind method for reducing the number of trips to the hospital (Rauv 2017; Your Team in India 2020).
- **Bots that provide customer service—**Natural language processing (NLP) and sentiment analysis were used to construct collaborative Chatbots. Patients can ask inquiries regarding bill payment, appointments, and medication refills 24*7 all through the year (Rauv 2017; Xu et. al. 2019; Your Team in India 2020).
- **The Process of Robot-Assisted Surgery**—When acquiring insights, doctors might use pre-op medical data to acquire information. Specific movement, robotic arms, and magnetic imaging are all characteristics of the robotic surgical system (Ksolves.com 2021; Your Team in India 2020).
- **Digital Consultation**—There are numerous health care apps available, which will respond to and would handle the patients' queries in case the concerns raised by the patient need real doctors' intervention, and the call is transferred to the real practitioner (The Medical Futurist 2021). It uses emotional artificial intelligence to provide personalized medical consultations. These days the bots are the first line of primary care (Xu et. al. 2019).
- **Health Observation**—Artificial intelligence (AI) and relevant technologies such as ML and DL are used to monitor the patient's health. When modifications are made, the applications send notifications to the user.
- **Drug Creation**—The search for novel drugs is predicted to become faster, cheaper, and more effective as a result of machine learning and other technologies. When it comes to generating new drugs, clinical trials take a lot of time and money (Ksolves.com 2021; Xu et. al. 2019).
- **Machines that are linked together**—The unique artificial intelligence makes the operation both simpler and more intuitive (Xu et. al. 2019).
- **Electronic Health Records (EHR) Standard**—Traditionally, physicians would manually record or type findings and patient data, and no two were alike. Interactions with patients, clinical diagnoses, and future therapies can all be augmented and reported more reliably (Xu et al. 2019).

7.4 Issues/Concerns of Using AI in Health Care

According to a survey conducted by Accenture, "The AI healthcare market will expand at a compound annual growth rate of 40% by 2021 (Bresnick 2017). However, adoption in healthcare is still in its early stages." Here are some of the AI-related problems and concerns in healthcare.

- **Data Accessibility**: the training of AI systems necessitates a large quantity of data from many sources, including electronic health records. Data is frequently spread across several systems. Handle such a huge data and too fragmented data enhances the chances of inaccuracy, minimizes the database completeness, and also expands the cost of acquiring the data (Patel 2022).
- **Wounds and Fault**: One major concern is that if at any point in time, AI systems diagnoses or handle the patient incorrectly; it will be leading to a disaster or patient injury. If an AI system prescribes the wrong medicine, fails to discover a tumor, or assigns a hospital bed to the wrong patient, then the patient may suffer harm (Patel 2022).
- **Questions about privacy**: Developers are enticed to collect data from a high number of patients while working with enormous datasets. Some patients may be concerned that their privacy would be violated as a result of this data collection. As a result of data sharing between huge health systems and AI companies, lawsuits have been brought (Patel 2022).
- **Bias and inequality**: There is a risk of prejudice and inequity in healthcare AI. The AI system may be biased. As the machines learn from the data they're given and may incorporate biases based on that data (Patel 2022; Dilmegani 2017).
- **Safety and Transparency**: IBM Watson for Oncology has come under fire for allegedly making "unsafe and incorrect" cancer care guidelines. Instead of using actual patient data, the program was only trained with a few "synthetic" cancer cases (Price II 2019; Patel 2022).
- **Technical Debt**: In the last five to seven years, the new AI techniques based on deep neural networks have achieved incredible results. Few individuals possess the technical skills required to solve the full spectrum of difficulties relating to data and software engineering Limited data and changeable data quality will be a common problem for AI solutions (Price II 2019; Patel 2022).
- **Unexplainable AI Models**: In order to produce better performance, most AI models become more complicated. Both healthcare companies and patients are concerned about the lack of logic. To function appropriately.
- **Strict monitoring protocols to avoid diagnostic errors**: Diagnostic mistakes account for 60% of all medical errors, causing 40,000–80,000 fatalities per year (Kelly et. al. 2019; Price II 2019; Patel 2022).

7.5 Why Explainable AI?

Symbolic AI was the first form of AI, in which logic rules represented knowledge. There was no ability to learn and a weak ability to deal with ambiguity. Then there's Statistical AI, which uses huge data to train statistical algorithms for specific areas. There is no contextual capacity and only a bare minimum of explainability. As depicted in the figure below the results of AI creates confusion (Pawar et al. 2020; Turek 2016; Thimbleby 2013) (Fig. 7.1).

Later systems were constructed using Explanatory models (Ahmad 2020). Systems learn and reason with new tasks and situations. As shown in Fig. 7.2, how AI is different from XAI? In AI it just computes and gives end result; why this result has been given no explanation is provided. XAI will give explain the result, which will give better clarity and understanding of the decision. Explainable AI is a field where techniques are developed to clarify AI system predictions. In this chapter, XAI is explored as a strategy for using AI-based systems to analyze and diagnose health data. In the field of healthcare, accountability, outcome tracing, and model improvement are all essential. XAI can be used to achieve transparency in the healthcare industry. It is required to make it easier to share data about a patient's medical history with doctors and practitioners.

In recent years, AI researchers have worked to bring neural networks out of the shadows and make them more apparent. The initial AI Models built were based on a black box in which when the result is produced it is difficult to trace the detailed information about why this result has come. Let us understand with an example—if a company has implemented the AI for fraud detection so the machine will give a scoring or a rank but will not give a detailed explanation for the same. Whereas as depicted in Fig. 7.3, XAI explains the predictions made by the machine, which will help the organization understand with logic and clarity the predictions made by the machine. Similarly, if a customer's loan has been rejected, he can see the details and work on improving his score (Bizarro 2020). Similarly, if a person has a family

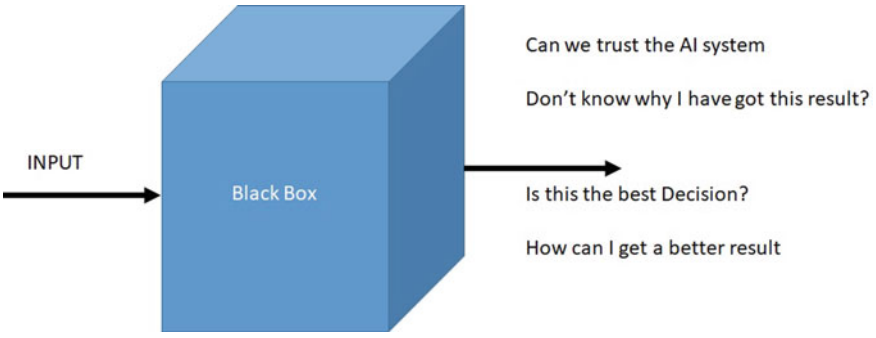


Fig. 7.1 AI may create confusion as a decision is given without any explanation

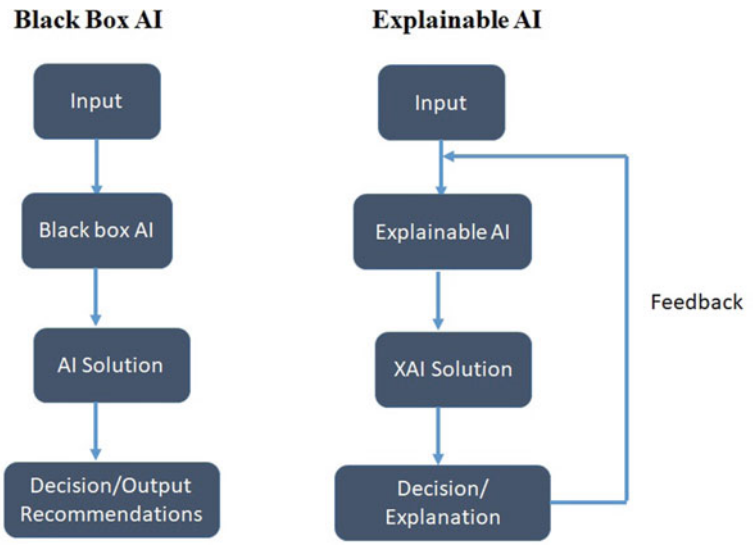


Fig. 7.2 Need for XAI

history of any disease such as Diabetes or Cancer, can get the tests done and based on the analysis, and the results doctors can suggest a lifestyle, diet for prevention.

IBM has announced AI Explainability 360, describing it as "a robust open-source toolset of cutting-edge methods that aid machine learning model interpretability and explanation" (Mojsilovic 2019). AI must be viewed as black boxes, having internal inference procedures that are impenetrable to humans and undetectable to

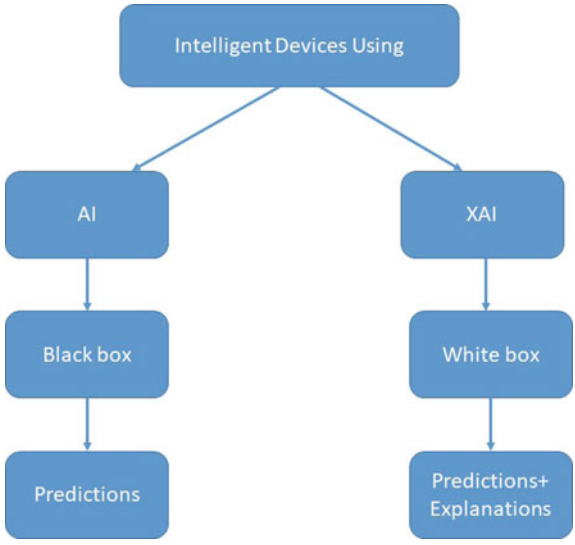


Fig. 7.3 XAI works as a white box that explains the results produced

the observer. The two most important aspects of XAI are transparency and post-hoc interpretation. In the eyes of developers, the transparency architecture shows how a model works, which includes feature relevance and evaluating & comparing the model.

Relevant features (specific data) for the study are provided like transaction amount, location, payment method, etc., and so on over a period of time can be used to detect fraud and prevent the same. The average money withdrawn from an account per month is an example of a feature. XAI assigns a value to each feature. Take a look at the features listed below:

- average monthly credit card charges in dollars.
- This card is likely to be used in this merchant category in this zip code.
- in the last week, the number of unique clients who utilized this IP address;

The model ranked the characteristic 'the average amount of credit card charges per month' in this case. Organizations can monitor and also can alter the priorities to improve the overall fraud detection and prevention. Later organizations can use model evaluation to assess the model's performance before releasing it. XAI model evaluation and comparison, is a critical component as organizations gain from testing the performance of a machine learning model since they can assess how accurate the model is and what the false positive rate is before relying on it for fraud detection and prevention. If the false positive rates do not satisfy the organization's objectives, the team tweaks the model until it performs optimally. Model comparison, as the name suggests, provides a side-by-side comparison of how different models perform when compared. Organizations can choose the best fraud detection models to deploy into production.

Let us shed a light on the components of Explainable AI, to attain transparency and post-hoc interpretation. Post-hoc extracts relationships between feature values and predictions, the behavior of a model. Model-agnostic methods can be used on any type of model, but model-specific methods can only be used on one type of model. They reject ante-hoc approaches, which build explainability into the model's structure, making it explainable even before the training phase is completed.

It attempts to (a) comprehend the model structure, such as decision tree construction; (b) understand single elements, such as a logistic regression limitation and (c) know the strategy for training, such as finding the solution in convex optimization. The post-doc interpretation explains why a result is presumed in the eyes of the customers. It tries to (d) provide analytic assertions, such as why a product is recommended on a shopping website; (e) provide visualizations, such as a saliency map for displaying pixel value in an object classification result; and (f) many transparent, interpretable algorithms such as K-nearest-neighbors, Convolutional Neural Networks, etc. are used to support the results (Dilmegani 2021; Xu et. al. 2019) (Fig. 7.4).

The Explainable AI (XAI) program aims to develop a set of machine learning techniques that will:

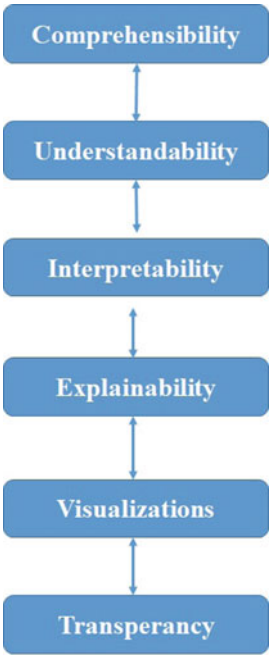


Fig. 7.4 Components of explainable AI

- Enable users to comprehend, effectively manage and adequately trust, the emerging generation of artificially intelligent partners.
- Producing more explainable models while preserving high learning performance (prediction accuracy).

Various XAI solutions have been proposed in the recent past, and have also been applied to the healthcare sector. Some XAI models are self-explanatory, based on the decision sets, which have influenced largely the prediction of several diseases like diabetes, asthma, lung cancer, etc. by seeing the patient's health record. The patient's records are self-explanatory as they are developed by mapping an instance of data to an outcome using IF-THEN rules. For example, decision sets will learn to forecast lung cancer given the following circumstances.

Predict lung cancer: If the individual smokes and has a history of respiratory disease. Self-explainable AI models have the drawback of limiting the amount of AI models that can be used to increase accuracy. There has been a surging interest in XAI techniques that can explain any AI model to address explainability in a larger number of AI models.

Model-agnostic XAI procedures are those that are unaffected by the AI model that needs to be explained. One of the most extensively used model-agnostic approaches, ***Local Interpretable Model-Agnostic Explanation*** (LIME), was presented by researchers as a framework for explaining predictions by quantifying the contribution of all the components involved in calculating prediction.

Researchers utilized LIME to describe how Recurrent Neural Networks (RNNs) forecast heart failure, and their explanations helped them identify the most frequent health problems that raise the risk of heart failure in people, such as renal failure, anemia, and diabetes. In the healthcare arena, various model-independent XAI approaches like Anchors and Shapley values have been developed and are actively used. An outline/ framework was specified for using human reasoning expertise in the development of XAI approaches, to improve details by incorporating the user's cognitive skills. The methodology developed could be used for any specific fields, such as to improve healthcare, and to provide user-friendly comprehension of how AI-based systems that apply XAI techniques at various phases enhances the clinical decision-making work.

There are certain challenges in putting XAI techniques into practice. XAI has created explanations to benefit the end-users, who might be physicians with medical domain knowledge or ordinary people. It is possible to create proper user interfaces for successfully displaying explanations.

7.6 History of XAI

Before talking about the XAI history, let us have a quick glimpse of AI. AI is where the machines depict and mimic human intelligence. Like—Self-driving cars, games using AI like chess, Amazon echo, Alexa, Siri, Google Alpha Go, IBM Watson, chatbots on websites working as virtual assistants (Kalinin 2020), etc. One can see many sci-fi movies like The Terminator, Star Trek, ex Machina, etc. AI is also used by E-Commerce organizations to suggest products based on previous purchasing patterns. Pepper recognizes human faces with a few emotions, Da Vinci Surgical System can perform minimally invasive surgeries, and Google Duplex can make reservations over the phone.

AI has moved from making intelligent machines to Machine learning, which has the ability to learn without being explicitly programmed. ML consists of 3 techniques Supervised learning (An input is mapped with an output, data sets are mapped—help to predict the next value), Unsupervised (Data-driven approach and clusters are formed), and Reinforced learning (learn from Mistakes—like these days gaming apps are based on this). It uses data to detect patterns and adjust accordingly, developing programs that can teach themselves to change and grow when needed and enables computers to find hidden insights using algorithms. In short, ML automates analytical model building. Numerous ML Techniques are available like Classification, Categorization, Clustering, Trend Analysis, Anomaly Detection, Visualization, and Decision Making. ML is used in Image processing, health care, robotics, text analysis, video games, and data mining. ML is used in various applications like—spam filtering, information extracting, and sentiment analysis. ML is a subset of AI and superset of DL. ML models need human intervention to reach the optimal outcome.

Deep Learning makes predictions independent of human intervention. DL makes the computation of multi-layer neural networks, based on the human brain. A few examples of DL, ImageNet a database of 14 million labeled images used to train neural nets, 2012, Google Brain team trained the neural networks by watching unlabeled images of cats from frames of YouTube Videos. In the year 2014, Facebook's Deep face was released identifying the face with 97.35% accuracy with a training set of 4 million images. Alpha Go developed by Google Deep mind defeated the 18-time world champion Lee Sedol in the year 2015. Deep learning is reshaping healthcare through image analytics and diagnosis and drug discovery and precision medicine.

Explainability in XAI derives from a combination of strategies that improve machine learning models' contextual flexibility and interpretability. There is no formal definition, however, it can be defined as the ability to draw conclusions based on conceptions (as a person) rather than probability alone. It can be seen as "contextual reasoning". An explainable Artificial Intelligence generates information or justifications to make its operation understandable or simple to comprehend. The next generation of artificially intelligent partners, with the help of XAI, will develop a set of machine learning algorithms to assist people in comprehending, trusting, and managing the diagnosis.

Explainable AI isn't a new notion. In the first work on explainable AI, which was published in the literature forty years ago, certain expert systems used rules to explain their conclusions. Since the dawn of AI, experts have advocated that intelligent systems should be used to explain AI findings, particularly when it comes to judgments. If a rule-based expert system refuses to accept a credit card charge, it must provide an explanation. The principles and knowledge of expert systems are simple to comprehend and infer since they are explained and established by human experts. A logically structured decision tree is a common strategy as shown in the following figure to demonstrate that why a loan application gets rejected on what parameters using decision tree (Fig. 7.5).

Similarly, XAI, helps in health care, let us see with the following figure, which helps the patient knows about the lump detected is benign or cancerous (Fig. 7.6).

7.7 Explainable AI's Benefits in Healthcare

Health care workers utilize AI to speed up and improve a variety of functions, including forecasting, risk management, decision-making, and even diagnosis, by scanning medical pictures for anomalies and patterns that are undetected to the naked eye. Many health care practitioners now use AI as a critical tool, but it is often difficult to understand, causing dissatisfaction among clinicians and patients, especially when making high-stakes decisions.

Machine learning's lack of explainability limits its use in healthcare applications where decision-makers need to understand the underlying reasoning. If artificial intelligence (AI) is unable to justify itself in the field of business, it will not be implemented on a big scale. If anyone is in charge of healthcare, the risk of taking

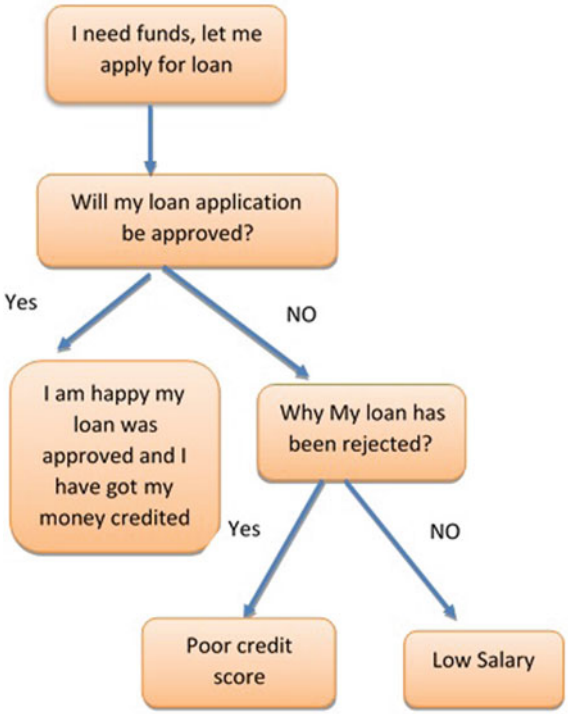


Fig. 7.5 Illustrates an example of a decision tree, which is constructed by working down from the top, level by level, according to the reasoning stated

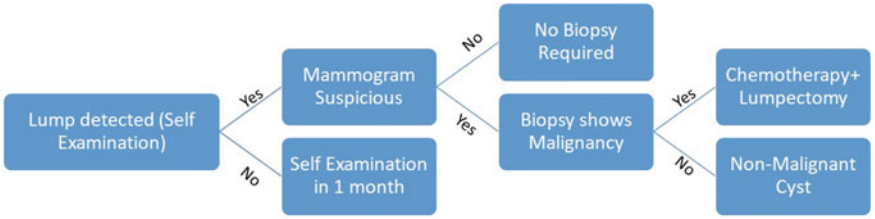


Fig. 7.6 Decision tree for detecting breast cancer

a poor/wrong decision may outweigh the benefits of precision, pace, and decision-making efficiency. This will, harm the environment. As a result, its reach and utility will be severely limited. As a result, it's important to take a close look at these issues. Until a model can be implemented in the healthcare domain, standard tools must be developed. Explainability is one such method (Gulavani and Kulkarni 2010).

XAI increases medical practitioners' and AI researchers' confidence in AI systems, resulting in the more widespread use of AI in healthcare. The heart disease

dataset as an example of how explainability approaches can be used to build trustworthiness when using deep learning systems in healthcare. The adoption of XAI comes with a slew of advantages. No matter what industry you work in, these advantages will help you and your company succeed. These are only a few of the main benefits of using XAI, according to Philip Pilger Storfer, a Quantum Black Data Scientist and XAI specialist:

- Developing user trust,
- Complying with legal obligations,
- Providing ethical reasoning, and
- Obtaining actionable and strong insights (Gill 2001).

Many companies are implementing XAI methodologies and techniques to achieve greater effectiveness, as shown by the findings described above. Explainable AI alleviates AI's problems and offers the following advantages (Gill 2001):

- **Trust and confidence**: Due to the uncertain existence of AI systems, gaining trust and confidence in doctors and patients is difficult. Users ask for answers from computers. In the pursuit of better efficiency, modern machine learning architectures are becoming more complex, often relying on black box-style architectures that provide computational benefits at the cost of model intelligence.
- **Detect and Remove Prejudice**: Since the system lacks clarity, users are unable to identify the system's flaws and biases. As a consequence, identifying and removing bias, as well as offering bias defense, becomes difficult.
- **Model Performance**: Model users are unable to monitor the model's actions due to a lack of knowledge.
- **Regulatory Standards**: Consumers are unable to assess whether or not the device complies with regulatory standards. Otherwise, the device could be affected.
- **Risk and Vulnerability**: It's important to be able to explain how systems deal with threats. Especially in circumstances where the user is unsure of the surroundings. Explainable AI assists in identifying it in a timely manner and taking effective action. But what if the device doesn't tell the consumer how to stop these dangers?

Explainable AI has accelerated the use of AI systems in healthcare. It is difficult for a person to make decisions because AI systems understand trends and make decisions based on Big Data. Explainable AI provides the following features (Gill 2001):

1. **Transparency**: The most important principle of Explainable AI is transparency. It is the algorithm, model, and features that the user can comprehend. Different users can need different levels of transparency. It is including appropriate explanations for appropriate users.
2. **Reliability**: The device offers reliable details. It should be in line with the model's production.
3. **Domain meaning**: The framework offers a user-friendly description that makes sense in the context of the domain. It is elucidating in the proper sense.

4. **Consistency**: For all forecasts, the interpretation should be consistent because different explanations will confuse the consumer.
5. **Generalizability**: The device should be able to explain things in a broad sense. However, it should not be too broad.
6. **Simplicity**: The system's description should be clear. It needs to be as transparent as possible.
7. **Reasonable**: It achieves the objective of each AI system's result.
8. **Traceable**: Explainable AI can track data and logic. The contribution of data in the production is revealed to the users. The user can track and solve logic and data problems.

7.8 XAI Has Proposed Applications for Patient Treatment and Care

Algorithms "that are inherently explainable" are the simplest method to develop functioning XAI in health care. Simpler solutions like decision trees, regression models, bayesian classifiers, and other transparent algorithms can be employed instead of sophisticated deep learning or ensemble approaches like random forests "without sacrificing too much performance or accuracy." XAI has been benefiting the medical practitioners and experts

1. **Assisted or automated diagnosis and prescription**: Chatbots can assist patients in self-diagnosis as well as doctors in diagnosis. Based on the symptoms identified by the patient, many healthcare organizations provide useful health and triage information. They do, however, note that no diagnosis has been made. This is to reduce their legal liability, but if the accuracy of chat bots increases, we may see chatbots delivering diagnoses in the future (https://medicalfuturist.com/top-12-health-chatbots/) Kalinin 2020; Kaushal et al. 2019).
2. **Prescription auditing**: Prescription auditing systems that use artificial intelligence (AI) can assist decrease prescription mistakes.
3. **Pregnancy Management**: Keep an eye on both the mother and the fetus to assuage the mother's fears and enable an early diagnosis.
4. **Real-time case prioritization and triage**: Prescriptive analytics on patient data enables real-time case prioritization and triage.
    - Jvion: The Cognitive Clinical Success Machine accurately predicts danger and comprehensively, providing prescribed actions that enhance outcomes.
    - Well frame: It flips the script by offering interactive care services to patients via their mobile devices. The Care Team's portfolio of clinical modules, which are based on evidence-based care, allows it to give a tailored experience.
    - Enlitic: Patient triaging solutions search incoming cases for various clinical findings, priorities them, and route them to the network's most suitable doctor.
5. **Personalized medications and care**: Assess the most appropriate treatment options based on patient data, decreasing costs and increasing care efficacy.

- GNS Healthcare: The business uses machine learning to match patients with the most effective treatments.
- Oncora Medicals: Software that helps health systems structure, interpret, and learn from their data in order to provide personalized care.

6. **Patient Data Analytics**: Analyze data from patients and/or third parties to uncover information and make recommendations. The organization (hospital, etc.) may use AI to analyze clinical data and derive deep insights into the health of patients. It allows for lower healthcare costs, more effective resource usage, and easier population health management (Daley 2018).

    - Zakipoint Health: The organization uses a dashboard to show all related healthcare data at the member level, allowing users to better understand risk and cost, as well as have personalized services and increase patient engagement.

7. **Surgical robots**: AI and collaborative robots are combined in robot-assisted surgeries. These robots are suitable for procedures that involve the same, repetitive movements because they can operate without being fatigued. AI can detect trends in surgical procedures, helping surgeons to improve their best practices and surgical robot control accuracy to sub-millimetre precision.
8. **Early diagnosis**: Analyze lab results and other diagnostic data to achieve an early diagnosis of chronic diseases.

    - Ezra: Ezra uses artificial intelligence to interpret full-body MRI scans and assist clinicians in cancer detection.

9. **Medical imaging insights**: Advanced medical imaging may be utilized to analyze and manipulate pictures, as well as to simulate future situations.

    - SkinVision: By taking pictures of your skin with your phone and going to a doctor at the right time, you can spot skin cancer early.
    - Medical imaging powered by artificial intelligence is also frequently used to diagnose COVID-19 cases and classify patients that need ventilator support. For example, Huiying Medical, a Chinese company, has developed a 96 percent accurate AI-powered medical imaging solution.

7.9 Future Prospects of XAI in Medical Care

The real healthcare advantage in the future would most likely come from the synergies gained by integrating the power of XAI-related technologies across the entire patient journey. Like, currently wearable devices could monitor heart rates, calories taken, sleep patterns, exercise levels, over time, blood glucose levels, and many more. In the future, this data could be synced to a central monitoring system that uses machine learning to detect irregular or undesirable behavior. The monitoring system will automatically alert the patient's physician and advise the patient to make an appointment (Fig. 7.7).

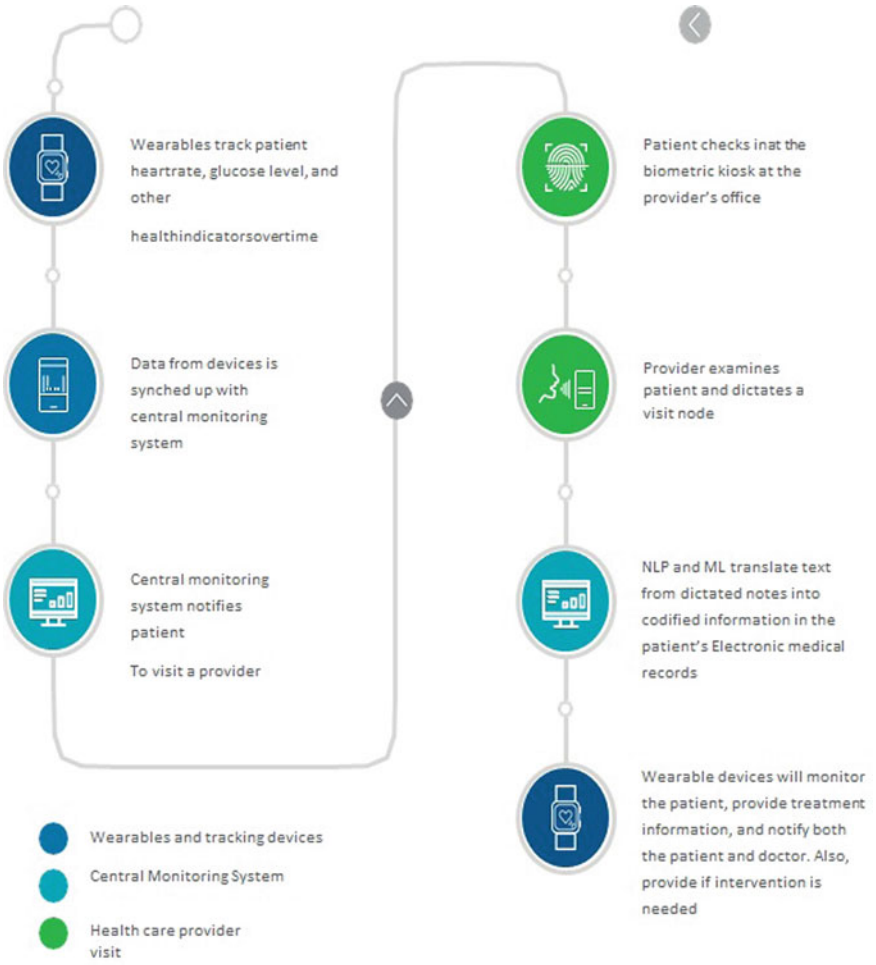


Fig. 7.7 XAI in health care sector

7.10 Case Study on Explainable AI

Explainable AI can help give local or global explanations for single predictions. There are techniques like model agnostic, and model specific. Various algorithms produce non-explainable predictions like the random forest, SVM, NN, etc. There are methods like Intrinsic/Ante-hoc and Post hoc methods. The ante-hoc methods are transparent, can be said that they are based on a white-box approach. The algorithms like Decision Trees, KNN, Fuzzy, Bayesian, etc. are used for transparent and fair predictions. Post-hoc is based on algorithms like PCA, CAM, LIME (**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations), or SHAP (**Sh**apley **A**dditive ex**p**lanations) (Daley 2018).

Diabetes is a very common disease in modern days due to lifestyle and eating habits. Diabetes is a condition in which the patient's blood glucose, often known as sugar, is usually either very high or too low. Glucose is the main source of energy that one gets from the food one eats. Although diabetes has no permanent cure, one can try to control/manage the sugar levels. It is necessary to keep track of the body. Explainable AI explains the data used for the prediction, their correlation, and EDA (Exploratory Data Analysis) to understand the hidden data patterns.

7.11 Framework for Explainable AI

XAI aims to make the model work transparently. It aims to select the right data and preprocess it for the model. It is required to have an accuracy of the model to predict the correct result, we need to be doubly sure when we are dealing with the health care domain as that is impacting the life and health of any individual. A single wrong result can have a threat to human life.

- Which feature influences more diabetes?
- What amount of Glucose do I need to maintain?
- Why did the system say that I can have diabetes in the future*?*

The explainable nature of AI can help doctors

The Glucose value of a person influences the result more while predicting whether a person can have diabetes or not. Doctors and health practitioners can answer what Glucose value an individual need to maintain to have diabetes. The explainable nature of AI can help us look at the change of having diabetes in the future. LIME and SHAP, two of the most common feature-based techniques, are very similar in intent but take very diverse methods.

Unlike LIME, which can only provide local explanations, SHAP can provide both global and local explanations. The dataset is used to generate many plots that illustrate the dataset globally and provide details about the relationships between the features and their significance.

Explainability is a key to producing a transparent, proficient, and accurate AI system. It makes it easy for the enduser to understand the AI systems' complex work. (Fig. 7.8).

Case Study's Conclusion

Explainable AI's contribution to the Diabetes Prediction framework simplifies the intricate workings of AI systems for the end-user. It offers the user a human-centered GUI. Explainability is essential for creating a transparent, competent, and reliable AI system that can assist healthcare practitioners, patients, and researchers in comprehending and using the system.

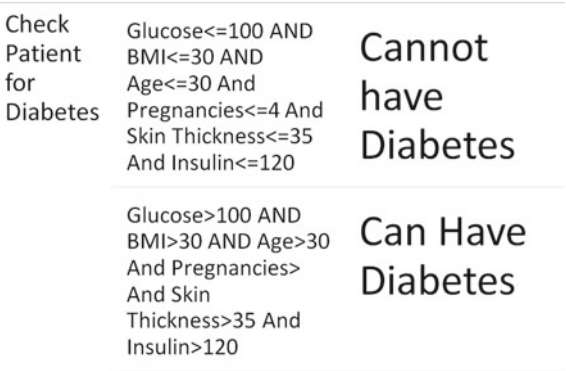


Fig. 7.8 XAI in detecting diabetes

7.12 Conclusion

The disruptive impact of technology on the healthcare business is undeniable. Even though it is a sector that necessitates highly skilled employees with several years of education, it also necessitates a significant amount of infrastructure and instruments. The rise in global life expectancy and the ageing of societies have fueled a surge in healthcare innovation and technology. With the environment changing every year, it looks like field innovation is highly powerful.

The majority of AI systems are not accountable for their outcomes, which can often damage society or users by producing incorrect results. Explainable AI and its values improve the way a system operates by describing the algorithm, model, and features it employs. The domain of XAI must be defined and deployed in AI-based health systems continuously. Most AI system is not answerable for their result, which can sometimes harm society or the user. Explainable AI and its principle bring a change in the system's traditional functioning.

The value of improving explainable AI capabilities has begun to be recognized by the markets. Many companies like Microsoft Azure and Google Cloud Platform have influenced many AI adoption patterns.

References

(n.a).: Top 12 benefits of AI in healthcare in 2021. Ksolves Emerging Ahead Always. https://www.ksolves.com/blog/artificial-intelligence/top-12-benefits-of-ai-in-healthcare-in-2021 (2021)

(n.a).: The top 12 health Chatbots. The Medical Futurist. https://medicalfuturist.com/top-12-health-chatbots/ (2021)

(n.a).: How explainable AI (XAI) for health care helps build user trust—even during life-and-death decisions. Capestart. https://www.capestart.com/resources/blog/how-explainable-ai-for-health-care-helps-build-user-trust/#:~:text=When%20and%20in%20what%20context,undetectable%20by%20the%20human%20eye

Ahmad, R.: Practical explainable AI: unlocking the black-box and building trustworthy AI systems. AI Time Journal. https://www.aitimejournal.com/@raheel.ahmad/practical-explainable-ai-unlocking-the-black-box-and-building-trustworthy-ai-systems-2 (2020)

Becker's Health IT: 10 biggest technological advancements for healthcare in the last decade. https://www.beckershospitalreview.com/healthcare-information-technology/ (2015)

Bizarro, P.: Explainable AI and transaction fraud: moving beyond black box explanations. The Paypers. https://thepaypers.com/thought-leader-insights/explainable-ai-and-transaction-fraud-moving-beyond-black-box-explanations--1241886 (2020)

Bouronikous, V.: Importance of technology in healthcare. Institute of Entrepreneurship Development. https://ied.eu/blog/importance-of-technology-in-healthcare/ (2013)

Bresnick, J.: Artificial intelligence in healthcare market to see 40% CAGR surge. HealthIt Analytics xtelligent Healthcare Media. https://healthitanalytics.com/news/artificial-intelligence-in-healthcare-market-to-see-40-cagr-surge (2017)

BuiltIn: What is healthcare technology? https://builtin.com/healthcare-technology

Daley, S.: 35 examples of AI in healthcare that will make you feel better about the future. Built in. https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare (2018)

Danzig, L.: The case for explainable AI (XAI). infoQ. https://www.infoq.com/articles/explainable-ai-xai/ (2020)

Datta, S., Barua, R., Das, J.: Application of artificial intelligence in modern healthcare system, alginates—recent uses of this natural polymer, Leonel Pereira, IntechOpen. https://doi.org/10.5772/intechopen.90454. https://www.intechopen.com/books/alginates-recent-uses-of-this-natural-polymer/application-of-artificial-intelligence-in-modern-healthcare-system (2019)

Dilmegani, C.: Explainable AI (XAI) in 2021: guide to enterprise-ready AI. AI Multiple. https://research.aimultiple.com/xai/ (2021)

Gill, J.K.: Explainable AI in healthcare industry. Akira.in. https://www.akira.ai/blog/ai-in-healthcare/ (2001)

Gulavani, S.S., Kulkarni, R.V.: Role of information technology in health care. In: Proceedings of the 4th National Conference; INDIACom-2010 Computing for Nation Development, February 25–26, 2010 Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi (2010)

Hosny, A., PArmar, C., Quackenbush, Schwartz, L.H., Aerts, H.J.W.L.: Artificial intelligence in radiology. In: PubMed Central, vol. 18(8), pp. 500–510 (2018). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/; https://www.researchgate.net/publication/228549680_Role_of_Information_Technology_in_Health_Care; https://www.researchgate.net/publication/342600571_Explainable_AI_in_Healthcare

Kalinin, K.: Medical chatbots: the future of the healthcare industry. Topflight. https://topflightapps.com/ideas/chatbots-in-healthcare/ (2020)

Kantarci, A.: 18 AI applications/usecases/examples in healthcare in 2021. AI Multiple. https://research.aimultiple.com/healthcare-ai/ (2021)

Kaushal, A., Abrams, K., Sklar, D., Fera, B.: The future of Artificial Intelligence in Healthcare. https://www.modernhealthcare.com/ (2019)

Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al.: Key challenges for delivering clinical impact with artificial intelligence. BMC Med. **17**, 195 (2019). https://doi.org/10.1186/s12916-019-1426-2. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2#citeas

Laal, M.: Health information technology benefits. In: 2nd World Conference on Innovation and Computer Sciences, AWERProcedia Information Technology & Computer Science, pp. 224–228. https://www.researchgate.net/publication/236216459_Health_information_technology_benefits (2012)

Leibowitz, D.: AI now diagnoses disease better than your doctor, study finds. Towards Data Science. https://towardsdatascience.com/ai-diagnoses-disease-better-than-your-doctor-study-finds-a5cc0ffbf32 (2020)

Luci, A.: The benefits of technology in healthcare: patient care & economic boom. AIMS Education. https://aimseducation.edu/blog/benefits-of-technology-in-healthcare (2015)

Medtech Europe: Value of medical technology. Medtech Europe from Diagnostic to Care. https://www.medtecheurope.org/about-the-industry/value-of-medtech/

Mojsilovic, A.: Introducing AI explainability 360. IBM. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/ (2019)

Nicholson, W.P.: Risks and remedies for artificial intelligence in health care. Brookings. https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/ (2019)

Patel, U.: Artificial intelligence in healthcare: top benefits, risks and challenges. Tristate Technology. https://www.tristatetechnology.com/blog/artificial-intelligence-in-healthcare-top-benefits-risks-and-challenges/ (2022)

Pawar, U., O'Shea, D., Rea, S., O'Reilly, R.: Explainable AI in healthcare. In: 2020 International 413 Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020 (2020). https://doi.org/10.1109/CyberSA49311.2020.9139655

Pulse: Advancements in healthcare technology—benefits for today's healthcare students. https://americancareercollege.edu/pulse/health-e-news/advancements-in-healthcare-technology-benefits-for-todays-healthcare-students.html (2019)

Rauv, S.: The impact of technology in healthcare. Elcom. Blog. https://www.elcom.com.au/resources/blog/the-impact-of-technology-in-healthcare-trends-benefits-examples (2017)

Scherman, J.: 5 Ways technology in healthcare is transforming the way we approach medical treatment. https://www.rasmussen.edu/degrees/healthsciences/blog/technology-in-healthcare-transformation/ (2019)

Thimbleby, H.: Technology and the future of healthcare. J. Public Health Res. **2**(3), e28 (2013). https://doi.org/10.4081/jphr.2013.e28. https://www.ncbi.nlm.nih.gov/pmc/articlmedes/PMC4147743

Turek, M.: Explainable artificial intelligence (XAI). DARPA. https://www.darpa.mil/program/explainable-artificial-intelligence

University of Illionis Chicago: 5 Ways Technology is Improving Health. https://healthinformatics.uic.edu/blog/5-ways-technology-is-improving-health/ (2020)

WHO/UNICEF: Water, sanitation and hygiene in health care facilities: status in low- and middle-income countries. World Health Organization, Geneva. https://www.who.int/news-room/fact-sheets/detail/health-care-waste (2018)

Xu et al. 2019Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: a brief survey on history, research areas, approaches and challenges (2019). https://doi.org/10.1007/978-3-030-32236-6_51

Your Team in India: Top 13 applications of artificial intelligence in the healthcare industry. Your Team in India. https://www.yourteaminindia.com/blog/top-13-applications-ai-healthcare-industry/ (2020)

Chapter 8
Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP

Uche Onyekpe, Yang Lu, Eleni Apostolopoulou, Vasile Palade, Eyo Umo Eyo, and Stratis Kanarachos

Abstract Despite the recent advancements in Autonomous Vehicle (AV) technology, safety still remains a key challenge for their commercialisation and development. One of the major systems influencing the safety of AVs is its navigation system. Road localisation of autonomous vehicles is reliant on consistent accurate Global Navigation Satellite System (GNSS) positioning information. The GNSS relies on a number of satellites to perform triangulation and may experience signal loss around tall buildings, bridges, tunnels, trees, etc. We previously proposed the Wheel Odometry Neural Network (WhONet) as an approach to provide continuous positioning information in the absence of the GNSS signals. We achieved this by integrating the GNSS output with the wheel encoders' measurements from the vehicle whilst also learning the uncertainties present in the position estimation. However, the positioning problem is a safety critical one and thus requires a qualitative assessment of the reasons for the predictions of the WhONet model at any point of use. There is

U. Onyekpe (✉) · Y. Lu · E. Apostolopoulou
School of Science, Technology and Health, York St John University, York YO31 7EX, UK
e-mail: u.onyekpe@yorksj.ac.uk

Y. Lu
e-mail: y.lu@yorksj.ac.uk

E. Apostolopoulou
e-mail: eleni.apostolopoulou@yorksj.ac.uk

U. Onyekpe · V. Palade
Centre for Computational Science and Mathematical Modelling, Coventry University, Priory Road, Coventry CV1 5FB, UK
e-mail: ab5839@coventry.ac.uk

E. U. Eyo
Faculty of Environment and Technology, Civil Engineering Cluster, University of the West of England, Bristol, UK
e-mail: eyo.eyo@uwe.ac.uk

S. Kanarachos
Faculty of Engineering and Computing, Coventry University, Priory Road, Coventry CV1 5FB, UK
e-mail: ab8522@coventry.ac.uk

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_8

therefore the need to provide explanations for the WhONet's predictions to justify its reliability and thus provide a higher level of transparency and accountability to relevant stakeholders. Explainability in this work is achieved through the use of Shapley Additive exPlanations (SHAP) to examine the decision-making process of the WhONet model on an Inertial and Odometry Vehicle Navigation Benchmark Data subset describing an approximate straight-line trajectory. Our study shows that on an approximate straight-line motion, the two rear wheels are responsible for the most increase in the position uncertainty estimation error compared to the two front wheels.

Keywords Wheel odometry · Autonomous vehicles · Inertial navigation system · Deep learning · Explainable machine learning · GNSS outage · Positioning · Neural networks

8.1 Introduction

The potential safety benefits of Autonomous Vehicles (AVs) have long been regarded as the technology's biggest assets (Wang et al. 2020). According to Liu et al. (2019), human error accounts for 75% of traffic-related road accidents in the UK, and 94% in the USA. There is the potential to significantly reduce these road accidents by minimising or eliminating the involvement of humans in the operations of vehicles (Papadoulis et al. 2019). This has been a strong selling point for self-driving vehicles to a public which, so far, seems unwilling to trust the technology (Wang et al. 2020). However, the introduction of AVs could introduce new kinds of accidents. Such safety concerns can affect the customers' intention to use AVs (Lee et al. 2019), and are majorly responsible for the current delay in the commercialisation of the AVs (Wang et al. 2020).

AVs acquire an understanding of their environment through the use of sensory systems (Babak et al. 2017). Ultrasonic systems, LIDARs and cameras are examples of such sensors that can be found on the vehicle's exterior. Cameras and LIDARs are used in the identification of objects, structures, potential collision hazards and pedestrians in the vehicle's path (Onda et al. 2018). Cameras are also essential in identifying signs and road markings on structured roads. An indication of the critical role of imaging systems in the operation of autonomous vehicles is clearly noticeable from the numerous sophisticated versions of these systems currently being employed (Ahmed et al. 2019). Despite the importance of imaging systems in the assessment of the vehicle's environments (e.g., determination of markings, vehicle-object relative positions, etc.), there is the need to localise a vehicle robustly and continuously, with reference to a well-defined coordinate system for real-time positioning and decision making.

8.1.1 Global Navigation Satellite System (GNSS) and Autonomous Vehicles

The GNSS receiver uses signals from at least three satellites orbiting Earth to localise the vehicle to a road (Yao et al. 2017). In spite of its wide acceptance for positioning, as it is unrivalled in terms of cost and coverage, the GNSS is not a perfect positioning system. The GNSS requires a direct line of sight between the GNSS antennae and the satellites to perform localisation; however, in metropolitan areas and similar environments, the line of sight can be blocked by features such as tall buildings, skyscrapers, bridges, dense tree canopies or road tunnels (Yao et al. 2017). Furthermore, the signal from the GNSS could be jammed, leaving the vehicle with no position information (O'Dwyer 2018). Hence, the GNSS receiver cannot serve as a standalone system of vehicle positioning.

The GNSS enables road localisation of the AV. To localise the vehicle to a lane, the GNSS may be used in combination with high accuracy LIDARs, cameras, RADAR and High Definition (HD) maps. There are however instances when the LIDAR and camera could be unavailable for use, or are uninformative. The usage of low-cost cameras and LIDARs could compromise their functions and accuracy especially during extreme weather events, such as blizzards, heavy snow fall, rain, fogs, or sleet (Templeton 2017). These issues are well-known in the field. Whilst the acquisition of LIDARs of high accuracy could render it vulnerable to theft and further increase the cost of the AV (Teschler 2018), camera-based positioning systems may suffer low accuracies depending on the objects in the camera's scene and the external light intensity (Teschler 2018).

According to tests performed by Cruise LLC and Waymo LLC on level 4 self-driving vehicle applications, the LIDARs scan is matched in real-time to a High Definition (HD) map (Teschler 2018). As a result, the system is capable of precisely positioning the vehicle within its surroundings (Teschler 2018). Nonetheless, the downside of this method is its high computational cost. In addition, changes in infrastructures within the driving environment could render an HD map temporarily obsolete and thus not effective for navigation.

Tesla, which is well known for its no LIDAR and HD map autonomy solution, handles GNSS signal outages by relying on its cameras and road markings until the GNSS signal becomes available. But the question is what happens if a decision is needed to be made on navigating to a new road during the signal loss or what happens when the GNSS signal is lost, and the camera is uninformative? Failure Mode and Effect Analysis (FMEA), which is an analysis performed to identify all the ways a system can fail and identify ways to mitigate them, would need to be performed on all the above failure scenarios to provide several fail-safe options to support the safe operation of autonomous vehicles.

8.1.2 Navigation Using Inertial Measurement Sensors

The use of high accuracy Inertial Measuring Unit (IMU) has been proven to be a way to overcome the GNSS reliability issue (Petovello et al. 2004). The IMU measures the AVs rotational rate and linear acceleration in the x, y and z-axis and computes its orientation, position, and velocity information by continuous dead reckoning. The significant cost of such IMU sensors has however hindered their adoption on autonomous vehicles. Even more, low-cost IMU's have accuracies too low to be used independently on autonomous vehicles as they are plagued by noise and biases, which are exponentially cascaded over time, for instance during the double integration from acceleration to position (Chen et al. 2018). In what is usually regarded as a symbiotic relationship, the GNSS can periodically calibrate the Inertial Navigation System (INS) during signal coverage to improve the position estimation accuracy of the INS during the GNSS outages.

Several researchers (Chiang et al. 2008; Noureldin et al. 2011; Fang et al. 2020; Dai et al. 2019) have studied the use of machine learning-based techniques to model the errors and learn the non-linear relationships that exist within the sensor's measurement. Such proposed techniques include Recurrent Neural Networks (RNN) based models (Fang et al. 2020; Dai et al. 2019; Onyekpe et al. 2021a, b; Onyekpe et al. 2020; Onyekpe et al. 2021), Multi Feedforward Neural Network (MFNN) based models (Chiang et al. 2008; Chiang 2003; Sharaf et al. 2005; El-Sheimy et al. 2006; Malleswaran et al. 2011), Radial Basis Function Neural Network (RBFNN) based models (Malleswaran et al. 2013; Semeniuk and Noureldin 2006) and the Input Delay Neural Network (IDNN) (Noureldin et al. 2011). Despite the numerous research into improving the performance of low-cost INS, the issue remains a challenge in need of cost-effective solutions.

8.1.3 Inertial Positioning Using Wheel Encoder Sensors

Modern vehicles are embedded with a number of sensors that support several advanced driver-assist systems, such as the wheel encoder of the Anti-lock Braking System (ABS). The wheel encoder, which operates by measuring the vehicle's wheel or axle speed, has been explored as an alternative to the commonly used low-cost accelerometer of the INS for vehicle positioning (Merriaux et al. 2014). The wheel encoder provides a better position estimation solution compared to the accelerometer, as its resolving requires one less integration step in the computation of the vehicle's position, thus minimising the error propagation. Nevertheless, the wheel encoder-based solution is not a perfect one either. The accuracy of the wheel encoder-based position estimation is affected by factors such as changes in the sizes of the tyres and wheel slippages (Onyekpe et al. 2020). A smaller tyre diameter, due to a reduction in the tyre pressure or a tyre replacement, leads to an underestimation of the

vehicle's displacement (Onyekpe et al. 2020), whereas a larger tyre diameter leads to the vehicle's displacement being overestimated (Onyekpe et al. 2020).

Reference (Onyekpe et al. 2020), showed that the errors present within the position estimation obtained from the wheel speed data can be learned by a Long Short-Term Memory (LSTM) neural network even in complex driving environments, such as roundabouts, successive left and right turns, wet roads, etc. In Onyekpe et al. (2021b), a Wheel Odometry Neural Network (WhONet) was proposed and shown to provide better estimations in both complex driving scenarios and longer-term GNSS outages of up to 180 s, with an accuracy averaging 8.62 m after 5.6 km of travel.

8.1.4 Motivation for Explainability in AV Positioning

Despite the remarkable performance of machine learning on the vehicle positioning problem, there is the requirement of transparency and higher level of accountability from the machine learning based system design. Explanations for machine learning model's decisions and estimations are thus needed to justify their reliability. This requires greater interpretability, often requiring an understanding of the mechanism underlying the operation of the algorithms. Unfortunately, the blackbox nature of neural networks is still unresolved, and many estimations are still poorly understood. Commonly, the eXplainable Artificial Intelligence (XAI) procedures consist of ensemble runs, random sampling and Monte-Carlo simulations, which are quite common methods in engineering. XAI may comprise of a systematic perturbation of some components of the model, which enables it to observe how it affects the model's estimates, mostly using sensitivity analysis. Due to the safety critical nature of the autonomous vehicle navigation systems, interpretability of the vehicular navigation machine learning-based models is necessary and provides sufficient argument for the suitability of a model for use on the road as well as sufficient argument when communicating anomaly behaviours to insurance stakeholders, Original Equipment Manufacturers (OEM) and other relevant stakeholders. We therefore explore in this research the interpretability of the WhONet models in estimating the position of autonomous vehicles in the absence of GNSS signals.

8.2 eXplainable Artificial Intelligence (XAI): Background and Current Challenges

8.2.1 Why XAI in Autonomous Driving?

Neural network-based models are built on complex non-linear functions and are commonly heavily parameterised (Chollet 2017; He et al. 2016; Ren et al. 2015). However, the high non-linearity feature as well as complexity of algorithms makes

it difficult to understand the internal working mechanisms. More importantly, such opaqueness can create distrust in Artificial Intelligence (AI) based applications. For instance, passengers may feel extremely anxious when sitting in the self-driving cars if the behaviours are not self-explanatory, e.g., a car suddenly turns around at an intersection whereas it normally passes it without explanation. Besides, the AI based models can take wrong actions due to biases in training data. This may cause catastrophic and even life-threatening consequences in medical diagnosis and treatment. As a result, the eXplainable Artificial Intelligence (XAI) becomes highly demanded to interpret models' decision-making and working mechanisms.

Explainability can enable good understanding of a model from different aspects, bringing insights that can be adopted by different stakeholders involved (Tintarev and Masthoff 2007). Figure 8.1 shows what positive effects can be brought by explainability to stakeholders. For instance, data scientists can easily debug an AI based model, adjust the parameters so as to improve performance, while business owners may care more about whether a model will fit with the business strategy and investment purpose. Risk analysts will need to check the robustness and decide on the deployment of the model, and regulators can evaluate whether a model is reliable as well as what impact can be triggered by its decision on the customers. Finally, consumers can demand for transparency in terms of how decisions were taken. Specifically, explainability in the development of AI approaches can help address different critical concerns such as in Table 8.1 (Belle and Papantonis 2021). In the example about autonomous vehicles, the passengers can trust the automated decision if the car turns around with an explanation such as "*a car accident is detected 200 m in front of us. We will choose another route from the previous exit which can take 10 min more than the usual route*" (Atakishiyev et al. 2112).

Fig. 8.1 Concerns faced by stakeholders

Table 8.1 Key XAI concerns (Belle and Papantonis 2021)

Features	Implications
Correctness	How confident are we that the variables contributing to the decision making are all and only those of interest? How confident are we that the spurious patterns and correlations have been removed?
Robustness	How confident are we that the model is not vulnerable to minor perturbations, and if so, can it be justified for that outcome? How confident are we that the model does not misbehave in the presence of noisy and missing data?
Bias	Are there any biases in the data that penalises any group of individuals unfairly? And if yes, can they be identified and corrected?
Improvement	How can the model's prediction be improved concretely? What are the effects of having an enhanced feature space and additional training data?
Transferability	How can the prediction model be generalised from one application domain to another? What properties of the model and data are needed to facilitate the transferability of the model to other domains?
Human comprehensibility	Can the algorithmic machinery of the model be explained to an expert and perhaps a lay person? Is the model's explainability needed for a wider deployment of the model

8.2.2 What is XAI?

8.2.2.1 Attributes of Explainability

Explainability attributes should contain the criteria and characteristics that researchers could use to define the explainability constructs. Firstly, it is necessary to make it clear and explicit to the end-users what causal relationships exist between the inputs and the model's predictions (Fox et al. 2017). According to Adadi and Berrada (2018), the explanation of the logic of an inferential system can help to justify, control, discover and improve the learning algorithm. In Montavon et al. (2018), interpretation is referred to as the mapping of an abstract concept (as a predicted class) to a domain that the human can understand. However, an explanation contains all features of a domain that can contribute to making a prediction (Montavon et al. 2018). Given the definition of interpretability or explainability as "*the degree to which a human observer can understand the reason behind a decision made by the model*" Dam et al. (2018), both notions can be interchangeable.

8.2.2.2 Theoretical Approaches for Structuring Explanations

Structuring an explanation for ad-hoc applications can involve making a decision on what information is included or excluded, e.g., causes, contexts, and consequences of the predictions from a model (Vilone and Longo 2021). Some researchers created a classification system for different explanation types, which can be suitable for different learning algorithms in terms of logic interpretation (Ribera and Lapedriza 2019). In addition, de Graaf and Malle (2017) identified that different types of users, problems, and behaviours require different explanations, as illustrated in Fig. 8.2. With the focus on user types, Glomsrud et al. (2019) summarised four explanation categories, ordered by the levels of completeness required by different user groups, i.e., explanations for developer, assurance explanations, explanations for end-users, and external explanations. Haynes et al. (2009), proposed three classifications for the types of explanations: the first, called Mechanistic operation, which attempts to answer the question "How does it work?"; the second was referred to as ontological explanations, which describes the structural properties of the model such as it's attributes and components, and how they relate; the third type was referred to as operational explanations, which attempts to answer the question "How do I use it?". Sheh and Monteath (2017), provided a more articulated classification of the types of explanations of intelligent systems to include teaching explanations, introspective tracing explanations, introspective informative explanations, post-hoc explanations, and executive explanations. Besides, Barzilay et al. (1998) proposed a classification of the knowledge that should be embedded in an explanation (Fox et al. 2017; Langley et al. 2017). Finally, Sohrabi et al. (2011) introduced a formal framework to generate preferred explanations for a given plan. It is necessary to contextualise the explanation preference to the observational patterns. Certain causes may affect the action, and thus require the explanation to reflect on the past, which means that produced explanations should consider the past events and data.

8.2.3 Types of XAI

Six types of post-hoc explanations for opaque models have been listed in Arrieta et al. (2020). In this section we focus on the most prominent types, which include model simplification, feature relevance, and visualisations.

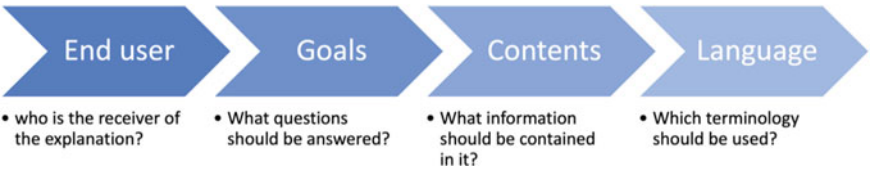


Fig. 8.2 Diagram of the main factors shaping the structure of a machine-generated explanation

8.2.3.1 Explanation by Simplification

Explanation by simplification aims to use a simpler model to approximate an opaque one which can be difficult to interpret. A popular technique is Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016), which can approximate a complex model. For instance, the complex model can be explained using a decision tree model built around the predictions. Ribeiro et al. (2018) designed a similar technique called anchors, which aims to approximate a model locally by using "if-then" rules. Krishnan and Wu (2017) proposed another simplification approach which seeks to partition the training dataset into similar instances while using a decision tree to structure the explanations. Similarly, Bastani et al. (2017) formulated simplification as an extraction process of using a transparent model to approximate a complex model. Particularly, the proposed method suggests building a greedy decision tree based on the predictions from a black-box model to obtain more insights about the original model while inspecting the surrogate one. Tan et al. (2018) considered simplification as a way of inspecting if the variable set is sufficient to restore the original one with the same accuracy. Wachter et al. (2017) proposed the counterfactual explanations for creating instances closed to those users who are interested in explaining. Through comparing the new data point to the original point, users can obtain insights on what minimal changes should be considered to change the decision made based on the original point.

8.2.3.2 Explanation by Feature Relevance

Feature relevance explanations attempts to measure the influence of each input, and provides a ranking of importance scores, showing which corresponding variables are more important than others for the model. One of the most significant contributions in this area is SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017). The Shapley feature values refer to the average expected marginal contributions of the features to the model's decisions. Shapley value has proven to be highly influential in the XAI community. Henelius et al. (2014) designed another method based on feature permutation to identify significant variables, or variable interactions that are picked up by the model. Additional ways to assess the significance of a feature can be quantifying the feature importance, transforming all features of a dataset, and achieving a new dataset without the influence of a certain feature (Adebayo and Kagal 1611). Datta et al. (2016) proposed QII (Quantitative Input Influence) to quantify the influence by estimating the performance change with the use of an original dataset, and the new dataset with the feature replaced by a random value. In this research, the explainability of the localisation model is investigated based on explanation by feature relevance using SHAP.

8.2.3.3 Explanation by Visualisation

Visual explanation aims to generate visualisations that allow a better understanding of a model. Existing approaches can help in obtaining insights about the decisions as well as how features interact with each other. Consequently, visualisations can be used to appeal to a non-expert audience. For this purpose, Cortez and Embrechts (2011) proposed a series of plots and discussed additional techniques (Cortez and Embrechts 2013), such as the Sensitivity Analysis approaches. Goldstein et al. (2015) introduced the ICE (Individual Conditional Expectation) and PD (Partial Dependence) plots, which can show insights into the relationship between the interested feature and the outcome (whether it is monotonic or linear, for example) (Molnar 2020). However, the average effect of the feature on the model's decision can be misleading and thus affect the identification of the interactions of the variables. Therefore, a more complete approach would be to utilise both ICE and PD plots, given that a relationship exists between these two plots.

8.3 XAI in Autonomous Vehicle and Localisation

Autonomous vehicles (AVs) have achieved significant milestones in research and development over the last decade (Atakishiyev et al. 2112). The significance of the need for XAI has been emphasised as advanced artificial intelligence techniques are applied in self-driving scenarios (Li et al. 2020). Currently, most of the advanced models of AVs are employ different state-of-the-art machine learning methods (Bojarski et al. 1604). Therefore, one of the research streams involves constructing a knowledge base into the AV systems, such as making text-based explanations for the vehicle's behaviour (Kim and Canny 2017; Kim et al. 2018). Meanwhile, some other researchers focus on trust computing of explainable AV models (Mittu et al. 2016; Petersen et al. 2017; Haspiel et al. 2018; Cysneiros et al. 2018). For example, the trustworthiness levels of AV systems can be calculated as a reference for insurance companies and customers (Hengstler et al. 2016).

The demand for explainable AVs creates diverse concerns and issues. Specifically, the occurrence of car accidents is considered a fundamental practical concern. According to Riberio et al., users will not adopt a model or a decision if they do not trust the machine (Ribeiro et al. 2016). With an empirical case study, Holliday et al. also showed that providing explanations can significantly increase users' trust towards a system (Holliday et al. 2016), however regaining the trust in an intelligent system can be onerous if it is damaged (Kim and Song 2021). Besides, trustworthiness in the decisions made by AVs can support transparency in the system. Such a positive factor can further develop fairness enabling good ethical analysis and causal reasoning of the decisive behaviours (Arrieta et al. 2020), achieving public approval of automated vehicles. In particular, the real-time decisive actions of AVs involve interconnected operational stages of sensing, localisation, planning and control as discussed below.

1. Sensing: As a primary requirement for the self-driving car, sensing refers to road surface extraction and object detection (Pendleton et al. 2017). Depending on the environment and the type of information to be captured, perception data can be collected by using devices such as the RADAR, LIDAR, ultrasonic sensors and cameras (Yeong et al. 2140; Ahangar et al. 2021).

2. Localisation: Localisation enables an AV to locate its position accurately in the physical world (Woo et al. 2018; Grigorescu et al. 2020) by comparing the location of reflected objects to the high-definition maps. One effective way is to use satellites to get the position of self-driving car, such as determining a global location of a car by using the Global Navigation Satellite System (GNSS) (Miguel et al. 2020). In places like underground tunnels and canyons, alternative sensor technologies like Inertial Measurement Units (IMUs) are combined with GPS, in order to navigate, control, and direct a car.

3. Planning: Based upon real-time environmental perception and localisation, an AV can plan its trajectory from the starting point to the destination. Particularly, the motion planning needs to consider the interaction with other vehicles, the dynamics of the environment such as people met on a trajectory, as well as available navigating resources and infrastructure. Geisberger et al. proposed the contraction hierarchies in fast routing (Geisberger et al. 2012). Studies on AV's planning use a variety of different terms for relevant components of the planning process.

4. Control: A feedback controller in an AV can read inputs from an actuator, fulfil the motion and correct errors brought in by actuation variables. With the aim of calculating the optimal solution for the prediction horizon, the feedback controller can make prediction on motions within a short time interval. Model predictive control has been successfully applied in several control applications, including the combined steering and braking, lane-keeping and navigating in adverse conditions dynamically (Liu et al. 2015; Falcone et al. 2007; Borrelli et al. 2006).

In this research, we however focus on XAI for AV localisation.

8.4 Methodology

In this section, we discuss the dataset employed in this study, the mathematical formulation of the target of the machine learning model, the details characterising the optimisation and evaluation of the WhONet model, and the SHapley Additive exPlanation method.

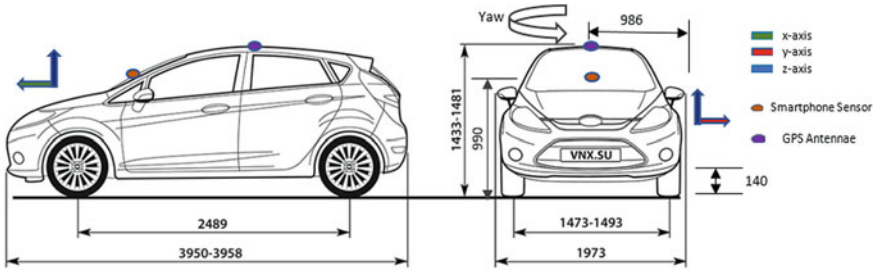


Fig. 8.3 Data collection vehicle, showing sensor locations (Onyekpe et al. 2021)

8.4.1 Dataset: IO-VNBD (Inertial and Odometry Vehicle Navigation Benchmark Dataset)

The IO-VNBD, publicly available at (Onyekpe et al. 2021c) is a large scale inertial and odometry dataset created to facilitate the benchmarking, development, and evaluation of positioning algorithms. The dataset is made up of several simple and complex driving scenarios such as residential road drives, sharp cornering hard brakes, dirt roads, roundabout, town drives, dirt roads, etc., and was collected over 5700 km and 98 h of driving. A Ford Fiesta Titanium vehicle as illustrated in Fig. 8.3 was used to collect the data on public roads within the United Kingdom. The Ford Fiesta is a front-wheel driven car. The dataset contains information describing the dynamics and position of the vehicle such as the speed of the vehicle's wheel (in rad/s) and GPS coordinates (in degrees) which were extracted from the vehicle's Electronic Control Unit (ECU) at a sampling frequency of 10 Hz. In this research, the V-Vw12 IO-VNB data subset, which describes driving on a motorway within the UK, is used. For more information on the IO-VNBD, please see (Onyekpe et al. 2021).

8.4.2 Mathematical Formulation of the Learning Problem

The angular velocity of the vehicle's wheels at any time (t) are measured by the wheel encoder. However, there can be uncertainties in the wheel encoder's measurements and the state of the tyres due to tyre wearing, changes in tyre pressure and wheel slips. The accuracy of the displacement estimation from the wheel encoder's measurements ω are affected by these uncertainties.

Equations (8.1), (8.2), (8.3) and (8.4) considers the errors that could affect the calculation of the vehicles wheel speed.

$$\hat{\omega}^b_{whrl} = \omega^b_{whrl} + \varepsilon^b_{whrl} \tag{8.1}$$

$$\hat{\omega}^b_{whrr} = \omega^b_{whrr} + \varepsilon^b_{whrr} \tag{8.2}$$

$$\hat{\omega}^b_{whfl} = \omega^b_{whfl} + \varepsilon^b_{whfl} \tag{8.3}$$

$$\hat{\omega}^b_{whfr} = \omega^b_{whfr} + \varepsilon^b_{whfr} \tag{8.4}$$

where $\hat{\omega}^b_{whfr}$, $\hat{\omega}^b_{whfl}$, $\hat{\omega}^b_{whrl}$, and $\hat{\omega}^b_{whrr}$ are the noisy wheel speed measurements of the front right, front left, rear left, and rear right wheels; whereas ε^b_{whfr}, ε^b_{whfl}, ε^b_{whrl}, and ε^b_{whrr} are the corresponding errors (uncertainties); and, ω^b_{whfr}, ω^b_{whfl}, ω^b_{whrl}, and, ω^b_{whrr}, are the respective error-free wheel speed measurements.

The calculation of the angular velocity of the rear axle is as shown in Eqs. (8.5) and (8.6).

$$\hat{\omega}^b_{whr} = \frac{\omega^b_{whrr} + \omega^b_{whrl}}{2} + \frac{\varepsilon^b_{whrr} + \varepsilon^b_{whrl}}{2} \tag{8.5}$$

Expressing $\frac{\varepsilon^b_{whrr}+\varepsilon^b_{whrl}}{2}$ as ε^b_{whr} and $\frac{\omega^b_{whrr}+\omega^b_{whrl}}{2}$ as ω^b_{whr}

$$\hat{\omega}^b_{whr} = \omega^b_{whr} + \varepsilon^b_{whr} \tag{8.6}$$

The vehicle's linear velocity in the body frame can be found from $v = \omega r$, where r is a constant which maps the angular velocity of the rear axle to its linear velocity:

$$v^b_{wh} = \omega^b_{whr} r + \varepsilon^b_{whr} r \tag{8.7}$$

Take $\varepsilon^b_{whr} r$ as $\varepsilon^b_{whr,v}$

$$v^b_{whr} = \omega^b_{whr} r + \varepsilon^b_{whr,v} \tag{8.8}$$

The vehicle's displacement in the body frame can be found through the integration of its velocity from Eq. (8.8) and incrementally updated for continuous tracking. $\varepsilon^b_{whr,x}$ in Eq. (8.9) is the integral of $\varepsilon^b_{whr,v}$ from Eq. (8.8)

$$x^b_{whr} = \int_{t-1}^{t} (\omega^b_{whr} r) + \varepsilon^b_{whr,x} \tag{8.9}$$

The uncertainty in the position estimation can be found through Eq. (8.10) during the presence of the GNSS signal. The task thus becomes that of estimating $\varepsilon^b_{whr,x}$ during GNSS outages needed to correct the vehicles displacement x^b_{whr}.

$$\varepsilon^b_{whr,x} \approx x^b_{whr} - x^b_{GNSS} \tag{8.10}$$

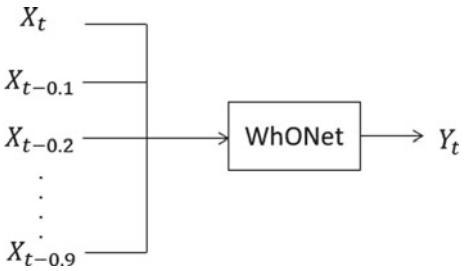


Fig. 8.4 WhONet's learning scheme

where x^b_{GNSS} refers to the true displacement of the vehicle measured according to reference (Onyekpe et al. 2020) using Vincenty's formula for geodesics on an ellipsoid based on the latitudinal and longitudinal positional information of the vehicle as implemented (Vincenty 1975; Pietrzak 2016). The accuracy of x^b_{GNSS} is however limited to the accuracy of the GNSS which according to VBOX Video HD2 (2019), is defined as ± 3 m.

8.4.3 WhONet's Learning Scheme

We adopt the WhONet model developed and evaluated in Onyekpe et al. (2021b) based on the simple Recurrent Neural Network proposed by Rumelhart et al. (1985). The WhONet's learning scheme is as presented in Fig. 8.4, where for any time t, the Neural Network's (NN's) input, $X_{t|t-0.9}$, is made up of the wheel speed information of all four wheels of the vehicle: $\hat{\omega}^b_{whrl}$, $\hat{\omega}^b_{whrr}$, $\hat{\omega}^b_{whfl}$ and $\hat{\omega}^b_{whfr}$ from every tenth of a second within the previous second; X_t, $X_{t-0.1}$ … and $X_{t-0.9}$.

The NN is then tasked with predicting Y_t, which is defined as the error $\varepsilon^b_{whr,x}$ between the GNSS-derived displacement x^b_{GNSS} and the wheel-speed-derived displacement x^b_{whr}. See (Onyekpe et al. 2021b) for more information on the WhONet model.

8.4.4 Performance Evaluation Metrics

The Cumulative Root Square Error (CRSE) metric as adopted Onyekpe et al. (2021a) is used to evaluate the performance of the WhONet model in this work. The CRSE describes the cumulative root mean squared of the prediction error for every one second of the total duration of the GNSS outage. The mathematical definition of the CRSE is as presented in Eq. (8.12)

$$\text{CRSE} = \sum_{t=1}^{N_t} \sqrt{e^2_{pred}} \tag{8.11}$$

Where N_t is the length of the GNSS defined as 18 s, e_{pred} refers to the prediction error, and t is the sampling period.

We also adopt the Average Error Per Second (AEPS) metric from Fang et al. (2020) in evaluating the performance of the WhONet model. The AEPS measures the average error of the prediction every second of the GNSS outage and is defined mathematically in Eq. (8.12) below.

$$AEPS = \frac{1}{N_t} . \sum_{t=1}^{N_t} e_{pred} \tag{8.12}$$

8.4.5 Training of the WhONet Models

The training of the WhONet model is done according to Onyekpe et al. (2021b). The model is trained using the Keras–Tensorflow version 1.15 platform (Google Brain 2017), in order to ensure compatibility with the SHAP library (Lundberg 2021). The training dataset is made up of the first 80% of the V-Vw12 data subset of the IO-VNB dataset as presented in Table 8.2. The V-Vw12 training set used to train the WhONet Model is characterised by motion on an approximate straight-line trajectory on the motorway over a distance of 2.12 km. The model was optimised with the Adamax optimiser using an initial learning rate of 0.0007 and a mean absolute error loss function. Table 8.3 highlights the parameters characterising the training of the WhONet model.

Table 8.2 IO-VNB V-Vw12 data subset (Onyekpe et al. 2021b)

Scenario	IO-VNB data subset	Total time driven, distance covered, velocity and acceleration
Motorway	V-Vw12	1.75 min, 2.64 km, 82.6–97.4 km/h, −0.06 to +0.07 g

Table 8.3 WhONet's training parameters

Parameters	Displacement estimation
Learning rate	0.0007
Dropout rate	0.05
Time step	1
Hidden layers	1
Hidden neurons	72
Batch size	128

8.4.6 WhONet's Evaluation

The WhONet model is evaluated on the last 20% of the V-Vw12 IO-VNBD data subset characterised by approximately 18 s.

WhONet's performance is evaluated on an approximately straight-line travel scenario on the motorway. Although this scenario appears relatively easy, it can be considered challenging because of the large distance per second the vehicle travels.

GPS outages are assumed on the test scenarios, for the purpose of the investigation with a prediction frequency of one second.

8.4.7 SHapley Additive exPlanations (SHAP) Method

The SHAP method, which was proposed in 2017 (Lundberg and Lee 2017), is a unified framework for the interpretation of the predictions of machine learning models. It is regarded as the only locally accurate and consistent method for feature attribution based on expectations. As well as being able to provide interpretable predictions, SHAP also interprets feature importance scores from complex models. SHAP values present a unified measure of feature importance by assigning an importance value φ_i to each feature, as such describing the effects of having that feature included in the model's prediction. SHAP values in a cooperative game theory could be represented mathematically as follows:

$$\varphi_i = \sum_{S \subseteq F, \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right] \tag{8.13}$$

Here F refers to a set of all the features, S is a subset of all features from F after the ith value has been removed. Consequently, two models f_S and $f_{S\cup\{i\}}$ are retrained and then a comparison is made between the predictions from these models and the current input $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)$, where x_S describes the input feature's values in the set S. The estimation of φ_i from $2^{|F|}$ differences is done by the approximation of the Shapley value by either Shapley quantitative influence or performing Shapley sampling.

We employ the SHAP approach in this study to interpret the predictions of the WhONet model. The SHAP analysis on the WhONet model is presented and discussed in Sect. 8.5.

8.5 Results and Discussions

The results from the evaluation of the WhONet model on the test data subset are presented on Table 8.4. The results reported, shows that the adapted WhONet model

achieves a CRSE of 0.54 m and AEPS of 0.03 m/s after about 500 m and 18 s of travel.

To further understand the predictions of the WhONet model trained on the motorway dataset, the SHAP values were calculated for the test observations. Figure 8.5 shows each feature ordered according to its average absolute SHAP value, with the feature that contributes the most to the model's output identified as the top placed bar in the illustration. Similar to Fig. 8.5, Fig. 8.6 presents a SHAP summary plot which also ranks the features according to their influence on the model's output. However, for each feature captured on the summary plot, there are multiple-coloured dots which each represents the SHAP value for that feature, and for each observation in the test dataset. The colour of the dots indicate the magnitude of the value of the feature, with the red dots indicating high wheel speed values and blue dots representing low wheel speed values. A greater distance from zero shows a greater influence on the model's prediction whilst a smaller distance shows less impact. The SHAP summary plot, essentially, reveals the effect an increase or a decrease of a specific feature affects the WhONet's position error estimation and to what degree. Figures 8.5 and 8.6 show that the top 5 features with the greatest impact on the WhONet's output are the wheel speed of the front right wheel at time $t - 0.1s$, the front left wheel speed at $t - 0.2s$, the front right wheel speed at $t - 0.3s$, the front right wheel speed at $t - 0.9s$, and the rear left wheel speed at $t - 0.1s$. Of these 5 features, an increase in the value of the wheel speed of the front right wheel at $t - 0.1s$ and $t - 0.9s$ leads to a decrease in the position error prediction. Conversely, an increase in the value of the wheel speed of the front left wheel speed at $t - 0.2s$, the front right wheel speed at $t - 0.3s$ and the rear left wheel speed at $t - 0.1s$ lead to an increase in the position error estimation. We also notice that the higher the speed of both front wheels at $t - 0.1s$ the higher the accuracy of the position error estimation. Similarly, these behaviors are repeated for the front right wheel speed at $t - 0.9s$, and $t - 0.8s$ and for the front left wheel speed at $t - 0.4s$, $t - 0.7s$, and $t - 08s$. These observations hint at a greater connection between the two front wheels and the accuracy of the predicted position error compared to the two rear wheels. We investigate these observations further by looking at the SHAP waterfall plot and the SHAP decision plot.

The SHAP waterfall plot, shown in Fig. 8.7, offers more insight into how each feature affects the WhONet's position error estimation. Starting from the expected value, when no features are taken into account (see the bottom of Fig. 8.7), the waterfall plot shows how applying one feature at a time affects the predicted position error, increasing or decreasing the error in the model's estimations, until it reaches the final model output, after all features have been taken into account. The features are sorted according to their influence on the predicted position error with those that have

Table 8.4 Performance measures of the WhONet model on the test dataset

Performance evaluation metrics	Results
CRSE	0.54 m
AEPS	0.03 m/s

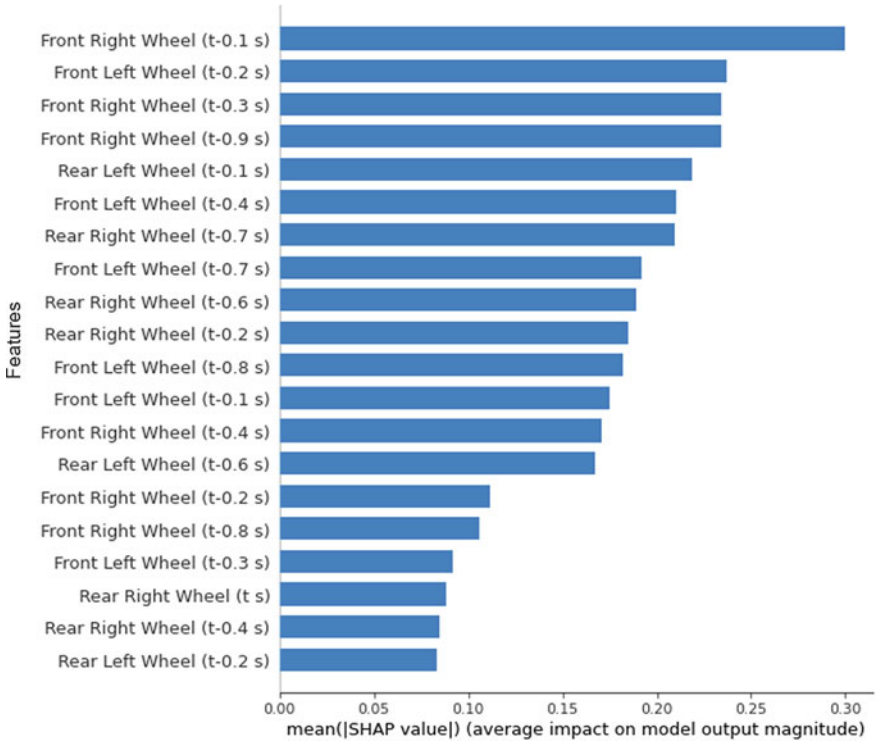


Fig. 8.5 Average absolute SHAP value per feature

the least impact on the model's output shown at the bottom of the plot and those with the most impact positioned at the top. The colour of the arrows indicates the direction of the change with red arrows showing an increase in the model's predicted position error and blue arrows showing a decrease. In this specific test observation, the wheel speed of the front left wheel at $t-0.2s$ increases the predicted position error the most. However, considering the SHAP values for each wheel speed across the timesteps within the sequence as presented on Table 8.5, it is observed that overall, the front left wheel contributes significantly to the reduction of the error in the predicted positional uncertainty compared to the other three wheels. It is further observed that the rear left wheel increases the error in the position uncertainty estimation the most. Table 8.6 reports the aforementioned behaviours across 4 additional test observations.

Figure 8.8 provides a collective demonstration on how the model arrives at its estimations across all the test observations. The decision plot uses a line graph to illustrate how the WhONet model navigates through the decision for each test observation, consequently demonstrating the effect of each feature on the decisions made by the model. Starting from the bottom, much like the waterfall plot, the decision plot shows how the SHAP values of the 20 most important features accumulate to move the model's predicted position error from the expected value to the final prediction.

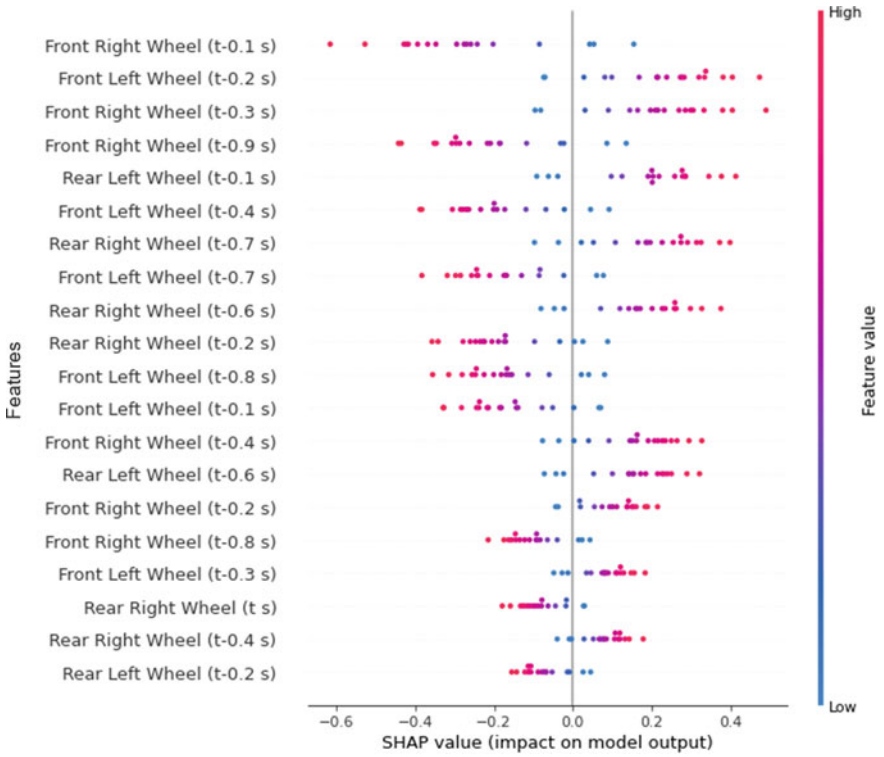


Fig. 8.6 SHAP summary plot showing individual feature contributions

The y-axis, again, shows each feature ordered from lowest at the bottom of the plot to highest at the top, according to their influence of the model's predicted position error. The movement of the line on the x-axis is the result of the SHAP value for that feature. The results in the decision plot confirms what was shown in the waterfall plot. In the majority of the cases, the rear wheels seems to increase the value of the predicted position error while the front wheels decreases it.

8.6 Conclusions

In this work, we have examined the interpretability of the WhONet model on a relatively simple scenario: an approximate straight-line trajectory on the motorway. Our study shows that overall, the two rear wheels are responsible for the most increase in the position error estimation, with the rear left being the most prevalent of the two. Although the reason for this is not immediately clear, the contributions of the measurements from the front wheels compared to the rear wheels could be attributed to the vehicle being a front wheel drive. Nevertheless, these behaviours have been

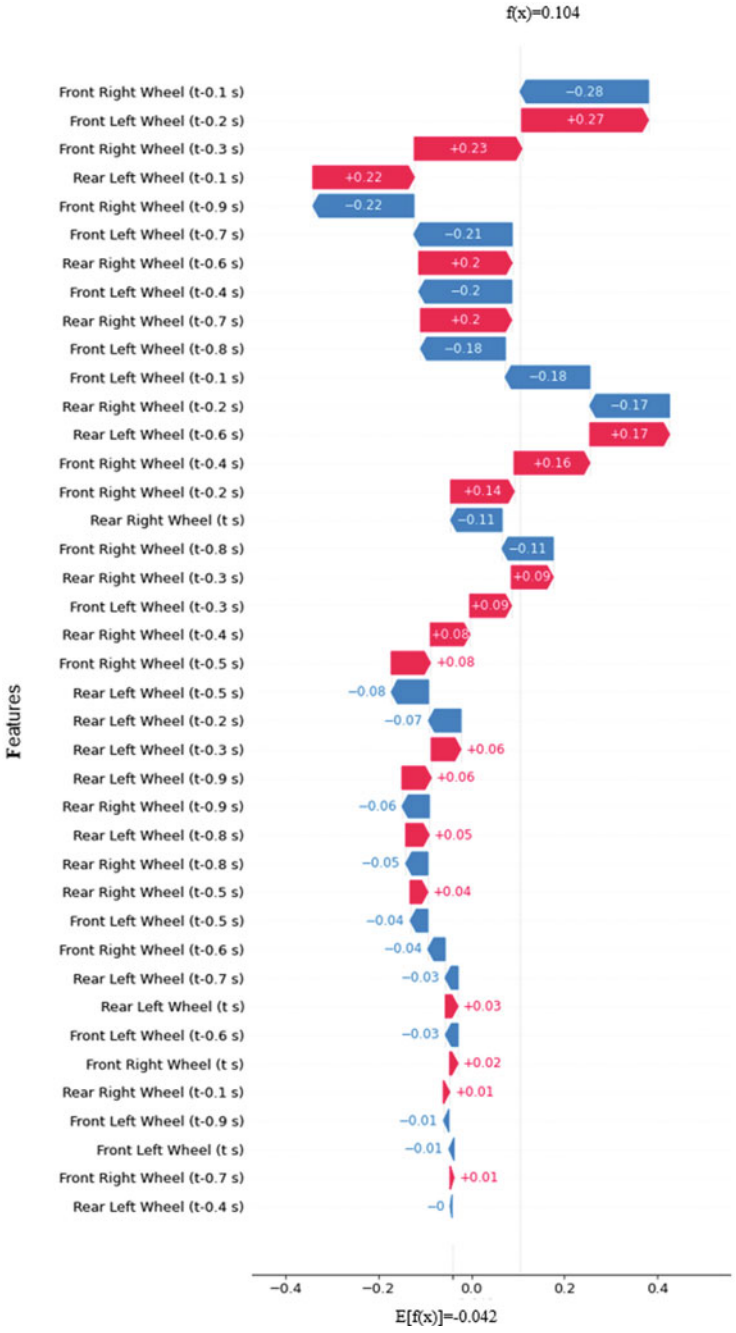


Fig. 8.7 Sample SHAP waterfall plot for a single observation

Table 8.5 Total SHAP values for wheel each across the timesteps

	Front left	Front right	Rear left	Rear right
t	−0.01	0.02	0.03	−0.11
t-1	−0.18	−0.28	0.22	0.01
t-2	0.27	0.14	−0.07	−0.17
t-3	0.09	0.23	0.06	0.09
t-4	−0.20	0.16	0.00	0.08
t-5	−0.04	0.08	−0.08	0.04
t-6	−0.03	−0.04	0.17	0.20
t-7	−0.21	0.01	−0.03	0.20
t-8	−0.18	−0.11	0.05	−0.05
t-9	−0.01	−0.22	0.06	−0.06
Grand total	−0.50	0.00	0.41	0.23

Table 8.6 Average SHAP value for 5 test observations

Test observations	Front left	Front right	Rear left	Rear right
Observation 1	−0.50	0.00	0.41	0.23
Observation 2	0.13	0.18	−0.05	−0.04
Observation 3	−0.72	−0.01	0.61	0.35
Observation 4	−0.70	−0.09	0.47	0.35
Observation 5	0.22	0.10	−0.18	−0.10
Average	−0.31	0.04	0.25	0.16

observed on a motion on an approximate straight line. Future research would involve exploring the generalisation of these behaviors to more complex scenarios, especially those characterised by a difference in the wheel speed of the left and right front wheels, such as on a roundabout, successive left right turns, etc. The output of this study could provide insights on how to improve the performance of the WhONet model for safer autonomous vehicle navigation.

Furthermore, gaining a deeper insight into how the features influence the model's predicted position error offers transparency into the decision making of the model. This can be valuable for the different stakeholders. For insurance companies, for example, explainability can offer a deeper understanding of the underlying causes in case of an accident. This information can also help manufacturers improve autonomous vehicles so that they take into account features that increase the predicted position error during the manufacturing process, consequently reducing the chance of an accident happening in the first place. By understanding the model, car retailers can have better knowledge of the vehicles they have available and can highlight their strengths and weaknesses to potential customers. Finally, for consumers knowing how features affect the predicted position error reduces the element of the unknown

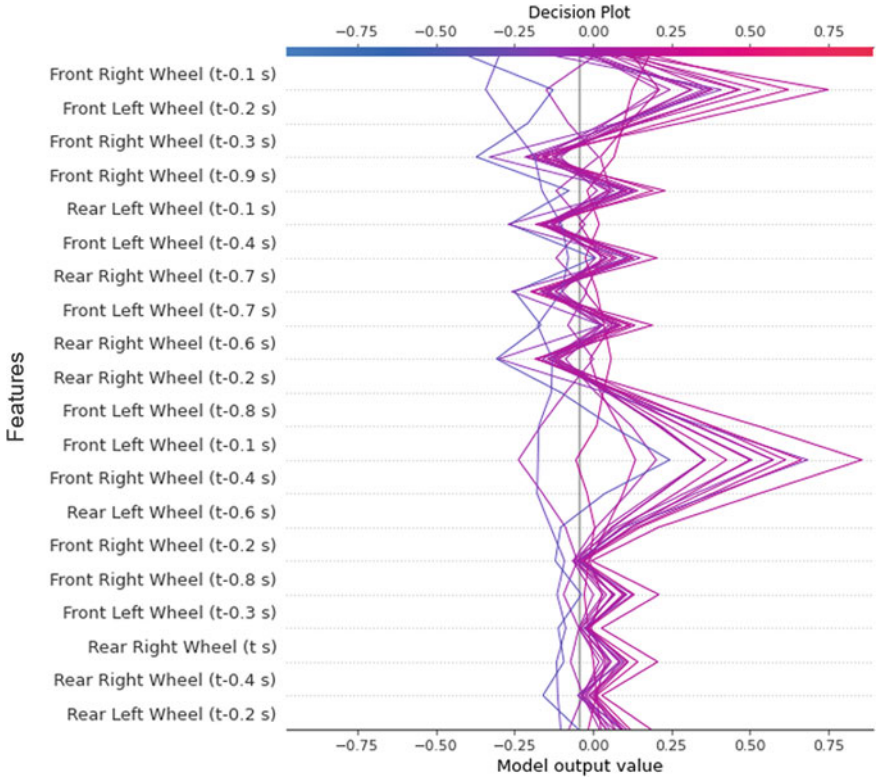


Fig. 8.8 Decision plot across all test observations

and provides some transparency into how the autonomous vehicle makes certain decisions.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Adebayo, J., Kagal, L.: Iterative orthogonal feature projection for diagnosing bias in black-box models. arXiv preprint arXiv:1611.04967 (2016)

Ahangar, M.N., Ahmed, Q.Z., Khan, F.A., Hafeez, M.: A survey of autonomous vehicles: enabling communication technologies and challenges. Sensors **21**(3), 706 (2021)

Ahmed, S., Huda, M.N., Rajbhandari, S., Saha, C., Elshaw, M., Kanarachos, S.: Pedestrian and cyclist detection and intent estimation for autonomous vehicles: a survey. Appl. Sci. **9**(11), 2335 (2019). https://doi.org/10.3390/app9112335

Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)

Atakishiyev, S., Salameh, M., Yao, H., Goebel, R.: Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. arXiv preprint arXiv:2112.11561 (2021)

Babak, S.-J., Hussain, S.A., Karakas, B., Cetin, S.: Control of autonomous ground vehicles: a brief technical review—IOPscience (2017). https://doi.org/10.1088/1757-899X/224/1/012029. Accessed 22 Mar 2020

Barzilay, R., McCullough, D., Rambow, O., DeCristofaro, J., Korelsky, T., Lavoie, B.: A new approach to expert system explanations (1998)

Bastani, O., Kim, C., Bastani, H.: Interpretability via model extraction. arXiv preprint arXiv:1706.09773 (2017)

Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Front. Big Data 39 (2021)

Bojarski, M., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)

Borrelli, F., Bemporad, A., Fodor, M., Hrovat, D.: An MPC/hybrid system approach to traction control. IEEE Trans. Control Syst. Technol. **14**(3), 541–552 (2006)

Chen, C., Lu, X., Markham, A., Trigoni, N.: IONet: Learning to Cure the Curse of Drift in Inertial Odometry, pp. 6468–6476 (2018)

Chiang, K.-W.: The Utilization of Single Point Positioning and Multi-Layers Feed-Forward Network for INS/GPS Integration, pp. 258–266 (2003)

Chiang, K.W., Noureldin, A., El-Sheimy, N.: Constructive neural-networks-based MEMS/GPS integration scheme. IEEE Trans. Aerosp. Electron. Syst. **44**(2), 582–594 (2008). https://doi.org/10.1109/TAES.2008.4560208

Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

Cortez, P., Embrechts, M.J.: Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 341–348 (2011)

Cortez, P., Embrechts, M.J.: Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. **225**, 1–17 (2013)

Cysneiros, L.M., Raffi, M., do Prado Leite, J.C.S.: Software transparency as a key requirement for self-driving cars. In: 2018 IEEE 26th International Requirements Engineering Conference (RE), pp. 382–387 (2018)

Dai, H.F., Bian, H.W., Wang, R.Y., Ma, H.: An INS/GNSS integrated navigation in GNSS denied environment using recurrent neural network. Def. Technol. (2019). https://doi.org/10.1016/j.dt.2019.08.011

Dam, H.K., Tran, T., Ghose, A.: Explainable software analytics. In Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, pp. 53–56 (2018)

Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617 (2016)

de Graaf, M.M.A., Malle, B.F.: How people explain action (and autonomous intelligent systems should too) (2017)

de Miguel, M.Á., García, F., Armingol, J.M.: Improved LiDAR probabilistic localization for autonomous vehicles using GNSS. Sensors **20**(11), 3145 (2020)

El-Sheimy, N., Chiang, K.W., Noureldin, A.: The utilization of artificial neural networks for multisensor system integration in navigation and positioning instruments. IEEE Trans. Instrum. Meas. **55**(5), 1606–1615 (2006). https://doi.org/10.1109/TIM.2006.881033

Falcone, P., Borrelli, F., Asgari, J., Tseng, H.E., Hrovat, D.: A model predictive control approach for combined braking and steering in autonomous vehicles. In: 2007 Mediterranean Conference on Control & Automation, pp. 1–6 (2007)

Fang, W., et al.: A LSTM algorithm estimating pseudo measurements for aiding INS during GNSS signal outages. Remote Sens. **12**(2), 256 (2020). https://doi.org/10.3390/rs12020256
Fox, M., Long, D., Magazzeni, D.: Explainable planning." arXiv preprint arXiv:1709.10256 (2017)
Geisberger, R., Sanders, P., Schultes, D., Vetter, C.: Exact routing in large road networks using contraction hierarchies. Transp. Sci. **46**(3), 388–404 (2012)
Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S., Smogeli, Ø.: Trustworthy versus explainable AI in autonomous vessels. In: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC), pp. 37–47 (2019)
Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. **24**(1), 44–65 (2015)
Google Brain: tensorflow 1.15 (2017)
Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. J. Field Robot. **37**(3), 362–386 (2020)
Haspiel, J., et al.: Explanations and expectations: trust building in automated vehicles. In: Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction, pp. 119–120 (2018)
Haynes, S.R., Cohen, M.A., Ritter, F.E.: Designs for explaining intelligent agents. Int. J. Hum.-Comput. Stud. **67**(1), 90–110 (2009)
He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
Henelius, A., Puolamäki, K., Boström, H., Asker, L., Papapetrou, P.: A peek into the black box: exploring classifiers by randomization. Data Min. Knowl. Discov. **28**(5), 1503–1529 (2014)
Hengstler, M., Enkel, E., Duelli, S.: Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. Technol. Forecast. Soc. Change **105**, 105–120 (2016). https://doi.org/10.1016/j.techfore.2015.12.014
Holliday, D., Wilson, S., Stumpf, S.: User trust in intelligent systems: a journey over time. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 164–168 (2016)
Kim, J., Canny, J.: Interpretable learning for self-driving cars by visualizing causal attention. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2942–2950, (2017)
Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 563–578, (2018)
Kim, T., Song, H.: How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telemat. Inform. **61**, 101595 (2021)
Krishnan, S., Wu, E.: Palm: machine learning explanations for iterative debugging. In: Proceedings of the 2nd Workshop on Human-in-the-Loop Data Analytics, pp. 1–6 (2017)
Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems (2017)
Lee, J., Lee, D., Park, Y., Lee, S., Ha, T.: Autonomous vehicles can be shared, but a feeling of ownership is important: examination of the influential factors for intention to use autonomous vehicles. Transp. Res. Part C: Emerg. Technol. **107**, 411–422 (2019). https://doi.org/10.1016/J.TRC.2019.08.020
Li, X.-H., et al.: A survey of data-driven and knowledge-aware explainable AI. IEEE Trans. Knowl. Data Eng. **34**(1), 29–49 (2020)
Liu, P., Yang, R., Xu, Z.: How safe is safe enough for self-driving vehicles? Risk Anal. **39**(2), 315–325 (2019). https://doi.org/10.1111/risa.13116
Liu, C., Carvalho, A., Schildbach, G., Hedrick, J.K.: Stochastic predictive control for lane keeping assistance systems using a linear time-varying model. In: 2015 American Control Conference (ACC), pp. 3355–3360 (2015)

Lundberg, S.: shap 0.40.0 (2021)
Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions (2017). https://github.com/slundberg/shap. Accessed 02 May 2022
Malleswaran, M., Vaidehi, V., Saravanaselvan, A., Mohankumar, M.: Performance analysis of various artificial intelligent neural networks for GPS/INS integration. Appl. Artif. Intell. **27**(5), 367–407 (2013). https://doi.org/10.1080/08839514.2013.785793
Malleswaran, M., Vaidehi, V., Deborah, S.A.: CNN based GPS/INS data integration using new dynamic learning algorithm. In: International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp. 211–216 (2011). https://doi.org/10.1109/ICRTIT.2011.5972270
Merriaux, P., Dupuis, Y., Vasseur, P., Savatier, X.: Wheel odometry-based car localization and tracking on vectorial map (extended abstract) (2014)
Mittu, R., Sofge, D., Wagner, A., Lawless, W.F.: Robust Intelligence and Trust in Autonomous Systems. Springer, Berlin (2016)
Molnar, C.: Interpretable machine learning. Lulu.com (2020)
Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. **73**, 1–15 (2018)
Noureldin, A., El-Shafie, A., Bayoumi, M.: GPS/INS integration utilizing dynamic neural networks for vehicular navigation. Inf. Fusion **12**(1), 48–57 (2011). https://doi.org/10.1016/j.inffus.2010.01.003
O'Dwyer, G.: Finland, Norway press Russia on suspected GPS jamming during NATO drill (2018). https://www.defensenews.com/global/europe/2018/11/16/finland-norway-press-russia-on-suspected-gps-jamming-during-nato-drill/. Accessed 04 Jun 2019
Onda, K., Oishi, T., Kuroda, Y.: Dynamic environment recognition for autonomous navigation with wide FOV 3D-LiDAR. IFAC-PapersOnLine **51**(22), 530–535 (2018). https://doi.org/10.1016/j.ifacol.2018.11.579
Onyekpe, U., Palade, V., Kanarachos, S., Christopoulos, S.-R.G.: A quaternion gated recurrent unit neural network for sensor fusion. Information **12**(3), 117 (2021). https://doi.org/10.3390/info12030117
Onyekpe, U., Palade, V., Kanarachos, S., Szkolnik, A.: IO-VNBD: inertial and odometry benchmark dataset for ground vehicle positioning. Data Brief **35**, 106885 (2021). https://doi.org/10.1016/j.dib.2021.106885
Onyekpe, U., Palade, V., Herath, A., Kanarachos, S., Fitzpatrick, M.E.: WhONet: wheel odometry neural Network for vehicular localisation in GNSS-deprived environments. Eng. Appl. Artif. Intell. **105**, 104421 (2021b). https://doi.org/10.1016/J.ENGAPPAI.2021.104421
Onyekpe, U., Kanarachos, S., Palade, V., Christopoulos, S.-R.G.: Vehicular localisation at high and low estimation rates during GNSS outages: a deep learning approach. In: Wani, M.A., Khoshgoftaar, T.M., Palade, V. (eds.) Deep Learning Applications, Volume 2. Advances in Intelligent Systems and Computing, vol. 1232, pp. 229–248, V. P. M. Arif Wani, Taghi Khoshgoftaar (eds.). Springer, Singapore, (2020). https://doi.org/10.1007/978-981-15-6759-9_10
Onyekpe, U., Kanarachos, S., Palade, V., Christopoulos, S.-R.G.: Learning uncertainties in wheel odometry for vehicular localisation in GNSS deprived environments. In: International Conference on Machine Learning Applications (ICMLA), pp 741–746 (2020). https://doi.org/10.1109/ICMLA51294.2020.00121
Onyekpe, U., Palade, V., Kanarachos, S.: Learning to localise automated vehicles in challenging environments using inertial navigation systems (INS). Appl. Sci. **11**(3), 1270 (2021a). https://doi.org/10.3390/app11031270
Onyekpe, U., Palade, V., Kanarachos, S., Szkolnik, A.: IO-VNBD: inertial and odometry benchmark dataset for ground vehicle positioning. Data Brief **35**, (2021c). https://doi.org/10.1016/j.dib.2021c.106885
Papadoulis, A., Quddus, M., Imprialou, M.: Evaluating the safety impact of connected and autonomous vehicles on motorways. Accid. Anal. Prev. **124**, 12–22 (2019). https://doi.org/10.1016/j.aap.2018.12.019

Pendleton, S.D., et al.: Perception, planning, control, and coordination for autonomous vehicles. Machines **5**(1), 6 (2017)
Petersen, L., Tilbury, D., Yang, X.J., Robert, L.: Effects of augmented situational awareness on driver trust in semi-autonomous vehicle operation (2017)
Petovello, M.G., Cannon, M.E., Lachapelle, G.: Benefits of using a tactical-grade IMU for high-accuracy positioning. Navig., J. Inst. Navig. **51**(1), 1–12 (2004). https://doi.org/10.1002/J.2161-4296.2004.TB00337.X
Pietrzak, M.: vincenty· · · PyPI (2016). https://pypi.org/project/vincenty/. Accessed 12 Apr 2019
Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28** (2015)
Ribeiro, M.T., Singh, S., Guestrin, C.: 'Why should I trust you?' Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. Proc. AAAI Conf. Artif. Intell. **32**(1) (2018)
Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In: IUI Workshops, vol. 2327, p. 38 (2019)
Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation (1985)
Semeniuk, L., Noureldin, A.: Bridging GPS outages using neural network estimates of INS position and velocity errors. Meas. Sci. Technol. **17**(10), 2783–2798 (2006). https://doi.org/10.1088/0957-0233/17/10/033
Sharaf, R., Noureldin, A., Osman, A., El-Sheimy, N.: Online INS/GPS integration with a radial basis function neural network. IEEE Aerosp. Electron. Syst. Mag. **20**(3), 8–14 (2005). https://doi.org/10.1109/MAES.2005.1412121
Sheh, R., Monteath, I.: Introspectively assessing failures through explainable artificial intelligence. In: IROS Workshop on Introspective Methods for Reliable Autonomy, pp. 40–47 (2017)
Sohrabi, S., Baier, J., McIlraith, S.: Preferred explanations: theory and generation via planning. Proc. AAAI Conf. Artif. Intell. **25**(1), 261–267 (2011)
Tan, S., Caruana, R., Hooker, G., Lou, Y.: Distill-and-compare: auditing black-box models using transparent model distillation. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 303–310 (2018)
Templeton, B.: Cameras or lasers? (2017). http://www.templetons.com/brad/robocars/cameras-lasers.html. Accessed 04 Jun 2019
Teschler, L.: Inertial measurement units will keep self-driving cars on track (2018). https://www.microcontrollertips.com/inertial-measurement-units-will-keep-self-driving-cars-on-track-faq/. Accessed 05 Jun 2019
Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: 2007 IEEE 23rd International Conference on Data Engineering Workshop, pp. 801–810 (2007)
VBOX Video HD2 (2019). https://www.vboxmotorsport.co.uk/index.php/en/products/video-loggers/vbox-video. Accessed 26 Feb 2020
Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)
Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Surv. Rev. **23**(176), 88–93 (1975). https://doi.org/10.1179/sre.1975.23.176.88
Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. J. Law Technol. **31**, 841 (2017)
Wang, J., Zhang, L., Huang, Y., Zhao, J.: Safety of autonomous vehicles. J. Adv. Transp. **2020** (2020). https://doi.org/10.1155/2020/8867757
Woo, A., Fidan, B., Melek, W.W.: Localization for autonomous driving. In: Handbook of Position Location: Theory, Practice, and Advances, 2nd edn, pp. 1051–1087 (2018)

Yao, W., et al.: GPS signal loss in the wide area monitoring system: prevalence, impact, and solution. Electr. Power Syst. Res. **147**(C), 254–262 (2017). https://doi.org/10.1016/j.epsr.2017.03.004
Yeong, D.J., Velasco-Hernandez, G., Barry, J., Walsh, J., et al.: Sensor and sensor fusion technology in autonomous vehicles: a review. Sensors **21**(6), 2140 (2021)

Chapter 9
A Smart System for the Assessment of Genuineness or Trustworthiness of the Tip-Off Using Audio Signals: An Explainable AI Approach

Sirshendu Hore and Tanmay Bhattacharya

Abstract Assessment of the genuineness or trustworthiness of a Tip-off is a challenging research area as it depends on the mental state and perception of the Tip-off providers. Thus, in the proposed work an attempt has been made to help the Law Enforcement (LE) personnel to assess the legitimacy of a Tip-off from a voice call. For the aforesaid objective, four widely used mental states such as 'Anger', 'Happy', 'Sadness', and 'Neutral' have been considered. To placate our goal, a few classical Machine Learning (ML) models, as well as a few latest ML models, have been employed. Regional, international, and a combination of both audio sets have been engaged for an in-depth study. The novelty of this work is to, select a set of Important 26 or 13 Mel-Frequency Cepstral Coefficients (MFCCs) using Explainable AI (XAI) approaches (Mean Decreased Impurity based Gini and Permutation), whereas most of the researchers had employed either the First 26 or 13 MFCCs in their works. The proposed model shows the supremacy over the conventional approach of using sequential MFCCs feature vector result analysis shows the supremacy of XAI-based features over conventional approaches thereby making our system better and smarter. Among the employed models, 1D CNN has shown its supremacy over other employed models for this study. Hence, the 1D-CNN-based Machine learning approach has been proposed.

Keywords Mental awareness · Machine learning · Smart system · Tip-off · Explainable AI approach

S. Hore (✉)
Department of CSE, Hooghly Engineering and Technology College, Pipulpati, Hooghly, West Bengal, India
e-mail: shirshendu.hore@hetc.ac.in

T. Bhattacharya
Department of IT, Techno Main, Salt Lake, Kolkata, India

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_9

9.1 Introduction

The purpose of a Tip-off is to send confidential information or to give an early warning to an individual or organization so that preventive measures can be initiated (https://dictionary.cambridge.org/dictionary/english/tip-off. xxxx). A timely Tip-off may prevent blood shade, and violence may bring down communal tension on the other hand it helps the LE personnel to capture a criminal who is involved in a heinous act such as murder or rape. In today's society, every citizen has the right to feel safe and stay safe. It is the collective responsibility of the citizens to become the ear and eyes of the local law enforcement official. If they observed any suspicious activity or crime then it's their basic responsibility to bring such incidents before the law enforcement officials. However, it has been observed that sometimes prank calls and intentional miss reporting for vengeance may distract the law enforcement officials, which may lead to a waste of time and effort (https://www.criminallawyersandiego.com//crimes-police-government/false-report/). By doing so these criminals and their associates can keep the police personnel busy in an area where no crime has been observed or taken place, at the same time allowing them to carry out criminal activity in other parts of the locality. In this way, they succeed to mislead the law enforcement agencies. Therefore, before initiating any action based on a Tip-off received, probably it will be better if these agencies tried to measure the genuineness or trustworthiness of the Tip-off. One way to measure the genuineness or trustworthiness of the Tip-off is to check the mental state of the Tip-off provider from their voice or speech. It has been reported in some literature that we can get people's mental state from their voices or speeches. Understanding people's mental state from their voices is one of the important areas of research. Consequently, lots of research work have been carried out in this direction (Basharirad and Moradhaseli 1891; Ayaida et al. 2011; Poria et al. 2017). The use of ML models has been used widely in SER based approach (Pinto et al. 2020; Akçay and Oguz 2020). It has been observed that RAVDESS, an English audio corpus (Livingstone et al. 2018), and EMODB, a German audio corpus (EMO-DB) have been used mostly in the SER-based system (Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019). Between the two methods, used popularly in SER based approach, isolated or static label-based systems are most commonly used in the SER because it is easier to implement. It has been found in various studies that MFCC is the most trusted audio feature among researchers (Shashidhar et al. 2018; Boles and Rad 2017; Panwar et al. 2017). Nowadays, several modern methods have been used to make ML-based AI models more explanatory. Collectively, these approaches are called XAI approaches. (McDermid 2021).One of the methods used in XAI (Saarela and Jauhiainen 2021; Fisher et al. 2019) is to look for important features from feature vectors. Thus, in the proposed work, we have generated 26 and 13 MFCCs using Mean Decrease in Impurity (MDI) and employing Permutation Importance with Correlated Features. RF-based XAI has been employed to achieve the stated objectives.

Motivation: As the legitimacy of the Tip-off in the form of a voice call depends highly on the mental state of the providers. It is our responsibility to help the LE

personnel by suggesting a smart XAI-based model, to assess the Genuineness of a voice call. This motivates us to pursue this work.

Objective: The objective is to build a smart system to assess the genuineness or trustworthiness of the Tip-off from a voice call based on Important MFCC(s) features employing XAI based approach.

Highlights: Following are some of the takeaways of the proposed work:

- Successfully able to assess the genuineness or trustworthiness of Tip-off by analyzing the provider's mental state from their voice sample.
- Employed RF-based XAI as a tool to find important MFCCs (26, or 13) from a set of 40 MFCCs
- Use of the Regional, international, and a combination of both audio sets
- Employed Conventional and Latest ML Models.
- For comparative analysis mean decreased based Gini and Permutation based approaches have been employed to find important MFCCs
- Execution Time for all the experimental works have been compared

The ***abbreviations*** used are Ar = Anger, Bm = Boredom, Cm = Clam, Dt = Disgust, Fr = Fear, Hy = Happy, Nl = Neutral, Sd = Sad, Su = Surprise; Perm = Permutation, CSL = Classical, LT = Latest, ML = Machine Learning, IMP = Important, CoD = EMODB + RAVDESS.

The rest of the paper is organized as follows: The background of the study has been done in Sect. 9.2, The methodology adopted has been discussed in Sect. 9.3 which is followed by the results and discussion section, Finally, the conclusion, limitation, and future scope of the study have been given.

9.2 Background

Nowadays Artificial Intelligence (AI) and Machine Learning are being used extensively, to assist the human decision-making process. These include some conventional ML models such as SVM, RF, kNN, and some advanced state of art approach-based ML models like CNN, DNN, LSTM, etc. Researchers are making use of these models because these models show good performance under different circumstances. With the advancement of cloud computing researchers have easier access to high-performance machine instances having higher throughput. Therefore the presence of ML-based solutions has been observed in various sectors of our life, such as audio and speech processing (Basharirad and Moradhaseli 1891; Ayadia et al. 2011; Poria et al. 2017; Pinto et al. 2020; Akçay and Oguz 2020; Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019; Shashidhar et al. 2018; Boles and Rad 2017; Panwar et al. 2017; Sinith et al. 2016), disorders during the growing time (Silva et al. 2020), image classification (Bendre et al. 2020), as well as in cyber-security (Parra et al. 2020). Pinto et al. (2020) proposed one 1D CNN model to govern the human emotional state using MFCCs as a feature set. The model has attained reasonable accuracy. Yang et al. (2020) suggested smart home assistance

using scaled MFCC as features to acquire the consumer's psychological state. In their work author(s) have engaged classical ML Models such as SVM, BPNN, ELM, etc. The proposed model has attained 92.4% accuracy. In the year 2021 Chatterjee et al. (2021) suggested one smart assistant system using 1D CNN. The suggested system used MFCC as an input to obtain higher accuracy. In their work 1D, CNN Model had been engaged. Lalitha et al. (2014) and Iqbal and Barua (2019) employed SER to determine human psychology in real-time, using both forms of the ML Model. In both works, they have employed MFCCs as a feature vector. Research works of Akçay and Oguz (2020), Shashidhar et al. (2018) suggested in detail the psychological models, dataset, classifier, pre-processing, and feature to be used in the SER system. At the same time, these models have given birth to too many queries because of a lack of interpretability (Gunning et al. 1973). XAI approaches have been used widely to mitigate some queries (McDermid 2021; Saarela and Jauhiainen 2021; Fisher et al. 2019). According to Bellotti (Bellotti 2009), there can be two levels of explainability. It can be either local or global or it could be Time based. The time-based interpretability can be further divided into three parts, Prior (what had been done), Contemporary (what is going on), and Post (what it has planned to do in the next). XAI methods help us to remove blind faith and bring transparency to the system (Zarsky 2016). It has been reported in various studies that transparency improves user awareness (Ananny and Crawford 2018), minimizes bias (Diakopoulos 2014), detects discrimination (Sweeney 2013), makes the user more accountable (Diakopoulos 2017), and helps to understand the functionalities of the intelligent system more in details (Lim and Dey 2009). Table 9.1 briefly describes some of the works carried out by the researchers.

9.3 Proposed Methodology

Figure 9.1 depicted the overall system in a nutshell.

9.3.1 Dataset Used

In the proposed work one regional language-based (EMO-DB), one international or mostly spoken language-based dataset (Livingstone et al. 2018), as well as a combined dataset combined these two datasets have been considered.

9.3.1.1 Regional Language-Based Audio Dataset

Berlin EMODB, a popularly used audio dataset used by the researchers to find participants' mental states. 5 Germans males and 5 Germans females have contributed to building the dataset. The dataset has seven emotion labels with a total size is 535. In

Table 9.1 A few studies to assess the mental state based on different parameters

Ref No (year)	Dataset language type	Employed features	Emotions labels used and count	Classifier used	Results
Lalitha et al. (2014)	Berlin EMODB Regional (German)	Frequency MFCCs, Scaled MFCCs,	Ar, Bm, Dt, Fr, Hy, Nl, Sd (7)	ANN	85.7%
Sinith et al. (2016)	Berlin EMODB Regional (German) SAVEE (English)	MFCCs, Pitch, Intensity	Ar, Hy, Nl, Sd (4)	SVM	Males: 67.5%, Females: 70%, Both: 75%,
Iqbal and Barua (2019)	RAVDESS SAVEE (English)	MFCCs, energy, spectral entropy	Ar, Dt, Fr, Hy, Nl, Sd, Sur (7)	GBM, SVM, KNN	Satisfactory
Yang et al. (2020)	Berlin EMODB Regional (German)	Scaled MFCCs	Ar, Hy, Nl, Sd (4)	SVM, BPNN, ELM, PNN	92.4%,77.8%, 7881%
Pinto et al. (2020)	RAVDESS (English)	MFCCs	Ar, Dt, Fr, Hy, Nl, Sd, Sur (7)	1D CNN	91%
Chatterjee et al. (2021)	RAVDESS, TESS (English)	MFCCs	Ar, Dt, Fr, Hy, Nl, Sd, Cm, Sur (8)	1D CNN	90.48%, 95.79%

the proposed work we have considered only four emotion labels with a total size of 337, Table 9.2.

9.3.1.2 International or Mostly Spoken Audio Dataset

It has been found in the literature that most of the researchers are using RAVDESS as their preferred dataset. In the dataset, there are eight emotion labels with a total size of 1440. In our proposed work we have only considered four emotion labels with a total size of 672. Twelve men and an equal number of women have contributed to building the dataset, Table 9.3.

9.3.1.3 Hybridization

This dataset has been obtained by combing regional audio datasets i.e. EMODB with international audio datasets i.e. RAVDESS, Table 9.4.

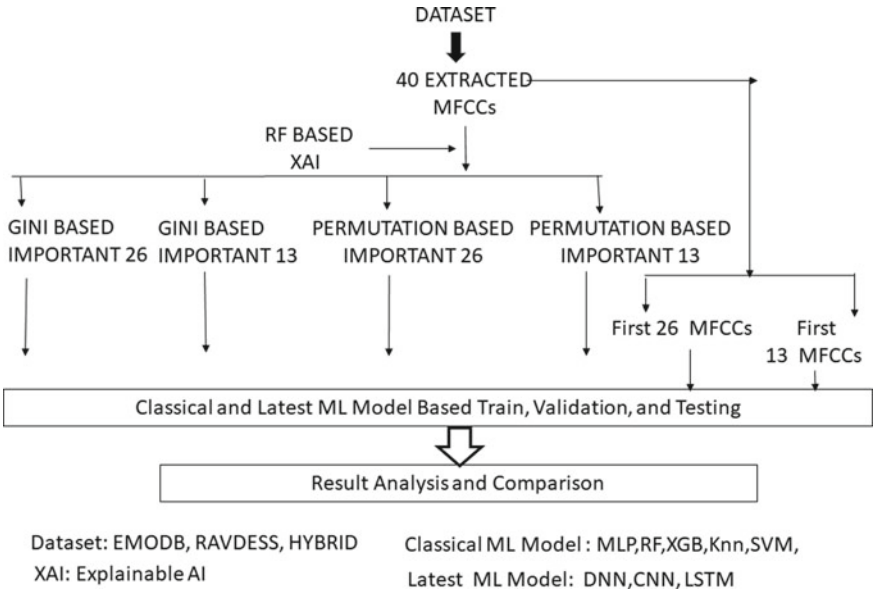


Fig. 9.1 Block diagram of proposed System

Table 9.2 Employed dataset and considered emotion labels. Here triple dots (…) de notes the emotion labels that have not been considered

Dataset	Emotion labels							
EMODB	*Ar*	Dt	Fr	*Hy*	*Nl*	*Sd*	Bm	Total
	127	48	69	69	79	62	61	535
	127	–	–	69	79	62	–	337

Table 9.3 Employed Dataset and considered emotion labels. Here triple dots (…) de notes the emotion labels that have not been considered

Dataset	Emotion labels								
RAVDESS	*Ar*	Dt	Fr	*Hy*	*Nl*	*Sd*	Su	Cm	Total
	192	192	192	192	96	192	192	192	1440
	192	–	–	192	96	192	–	–	672

Table 9.4 Hybridization of EMODB and RAVDESS

Dataset	Emotion labels				
CoD	Ar	Hy	Nl	Sd	Total
	319	261	175	254	1009

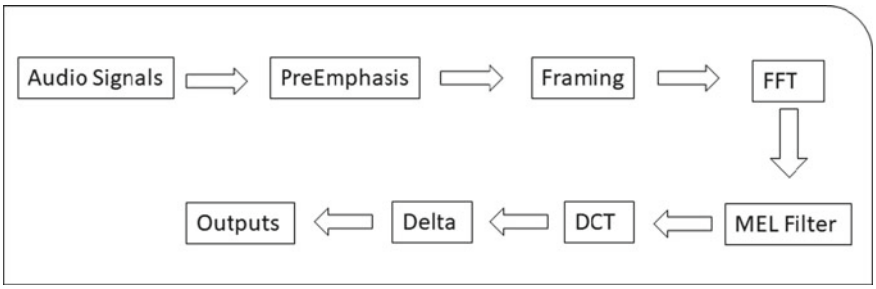


Fig. 9.2 MFCCS extraction process

9.3.2 Pre-processing

To make the system better and more robust pre-processing has been introduced. Through this process, we tried to increase the standard of voice or speech. Normalization, noise removal, trimming, etc. are some widely used pre-processing techniques.

9.3.3 Feature Extracted

It has been reported in various studies that MFCC is the most trusted audio feature among researchers (Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019). MFCC 40, 26, and 13 have been used widely to determine human mental state from speech or voice. Figure 9.2 shows the general process followed to extract MFCCs from an audio signal.

9.3.4 Feature Selected

It is quite obvious that to make the system smart we need to reduce the size of the number of features. There have been various ways we can reduce the size of the feature vector (Velliangiria et al. 2019). With the advancement of XAI, people are making use of the XAI approach to reduce the number of features and included those features having more impact on the classification process.

The Mean Decrease in Impurity based on important features can be obtained by calculating the node probability using the following equation:

$$n_{ij} = w_j C_j - wl(j)Cl(j) - wr(j)Cr(j) \tag{9.1}$$

Here, n_{ij} is the importance of node j, wj is the weighted number of samples reaching node j, Cj represents the impurity value of node j, l(j) is the child node on the left of node j, r(j) is the child node on right of node *j*.

The Permutation Importance with Correlated Features can be obtained using the following equation below

$$i_j = s - \frac{1}{K}\sum_{k=1}^{k} s_{kj} \tag{9.2}$$

Where i_j the permutation of importance, S is the reference score of the employed RF model, J is the feature to be evaluated, K number of iterations, s_{kj} is the computed score based on the employed model.

9.3.5 Machine Learning in SER

It has been observed that Machine learning algorithms have been used widely to determine a human mental state from his/her voice or speech. It can broadly be categorized into the following two categories.

9.3.5.1 Classical ML Models

In the proposed work some of the widely used ML models such as Multilayer perceptron with the feed-forward network, Random forest, XGBoosing, K nearest Neighbour, and support vector machine have been engaged.

9.3.5.2 Latest ML Models

With the advisement of new technology, storage capacity, and advanced processing systems people have shifted their focus from the Conventional ML Approach to these new approaches. Alex Net, RNN, CNN, AutoML, etc. have been used extensively to solve the real-time problem.

9.3.6 Performance Index

Judging the performance of the employed model is one of the required parameters in all classification processes. In the proposed work following yardsticks have been engaged.

$$Precision = \frac{tp}{tp + fp} \tag{9.3}$$

$$Recall = \frac{tp}{tp + fn} \tag{9.4}$$

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{9.5}$$

Where tp = True Positive, fp = False Positive, tn = True Negative, fn = False Negative.

9.4 Results and Discussion

Detail experiments have been conducted to test the impact of important features to determine the Tip-provider's mental state. Doing so will help the LE personnel to protect citizens from possible threats or untoward incidents To satisfy the objective, two popularly used audio/speech datasets as well as one dataset, combining these two datasets have been produced and employed. Scikit-learn Python library has been employed for all ML-based experimental processes. It is also used to find the important features. Google Colaboratory (Colab) has been used as an environment. For CSL-based ML, the underline is used to indicate the best accuracy while bold font represents the same for LT-based ML. Table 9.5 illustrates the parameter of the RF-based approach to find the important features.

9.4.1 Figure 9.3 shows the outline view of the employed DNN model while Figs. 9.4 and 9.5 show the architecture of the employed CNN and LSTM models. Table 9.6 tried to analyze the performance of CSL and LT models employing 40 MFCCs as a feature vector. Tables 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12, tabulated the findings of CSL and LT ML models. In all such experimental findings, different MFCCs feature vector sizes such as 26 and 13 in sequence; MDI, and permutation-based important 26 and 13

Table 9.5 Parameters used to find important MFCCs using the XAI approach

Parameter name	Value	Parameter name	Value
Bootstrap	True	class_weight	None
Criterion	Gini	max_depth	None
max_features	Auto	max_leaf_nodes	None
min_impurity_decrease	0.0	min_impurity_split	None
min_samples_leaf	1	min_samples_split	2
min_weight_fraction_leaf	0.0	n_estimators	100
n_jobs	None	oob_score	False
random_state	0	Warm_start	False

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	896
dense_1 (Dense)	(None, 128)	8320
dense_2 (Dense)	(None, 256)	33024
dense_3 (Dense)	(None, 4)	1028

Total params: 43,268
Trainable params: 43,268
Non-trainable params: 0

Fig. 9.3 The architecture of the employed DNN model for the proposed study

have been engaged. These MFCCs have been extracted from EMODB, RAVDESS, and combined datasets. Tables 9.13, 9.14 and 9.15 shows the comparative analysis of the findings. Tables 9.16, 9.17 and 9.18 demonstrate the time taken to execute different models for different MFCCs in the employed datasets. Figures 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, 9.12, 9.13, 9.14, 9.15, 9.16 and 9.17 show the selected 26 and 13 MFCCs, obtained using MDI, and Permutation based approach from the employed dataset. Figures 9.18, 9.19 and 9.20 display the confusion Matrix of adopted 1D CNN while Fig. 9.21 shows the accuracy and loss for the adopted 1D CNN models. Table 9.19 shows the performance of the adopted 1D CNN model based on the benchmark score.

9.4.2 Based on Table 9.6, it can be recorded that in the case of EMODB, MLP has shown its supremacy over other CSL models employed for this study. For the other two datasets, RF is comparatively better. In the case of LT-based ML models, CNN has a clear mandate over the other two. The result of Table 9.7 shows that for CSL models, the important 26 and 13 MFCCs obtained using MDI based approach are the most appropriate since both of them have achieved 95.59% accuracy. The kNN and MLP are the most suitable CSL ML models. CNN is the best LT-based ML model for the proposed work, Table 9.8. MFCCs 13 in sequence and MDI-based important 13 MFCCs have shown their supremacy over the rest. Both have achieved 98.88% accuracy using CNN.

9.4.3 Based on Table 9.9 it can be recorded that among the CSL models, MLP is the most suitable ML model. The model has achieved the best accuracy (71.53%) and MDI-based Important 26 MFCCs are the most appropriate feature set compared to other MFCCs involved. The MLP-based model also shows its supremacy. It has achieved 72.99% accuracy also MDI based MFCCs 13 is the most suitable feature vector. In the case of LT-based ML models, Table 9.10, CNN has a clear mandate

```
Model: "CNN"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 Conv1 (Conv1D)               (None, 26, 32)            192

 Activation1 (Activation)     (None, 26, 32)            0

 Dropout1 (Dropout)           (None, 26, 32)            0

 Conv2 (Conv1D)               (None, 26, 32)            5152

 Activation2 (Activation)     (None, 26, 32)            0

 Dropout2 (Dropout)           (None, 26, 32)            0

 Conv3 (Conv1D)               (None, 26, 64)            10304

 Activation3 (Activation)     (None, 26, 64)            0

 Dropout3 (Dropout)           (None, 26, 64)            0

 Flat1 (Flatten)              (None, 1664)              0

 Dense1 (Dense)               (None, 4)                 6660

 Activation4 (Activation)     (None, 4)                 0

=================================================================
Total params: 22,308
Trainable params: 22,308
Non-trainable params: 0
```

Fig. 9.4 The architecture of the employed CNN model for the proposed study

over the other two. The result of Table 9.9 also shows that MDI-based important 26 MFCCs (accuracy 78.11%) and MDI-based important 13 MFCCs (accuracy 74.72) are the most suitable feature sets among the feature sets employed for this study.

9.4.4 Table 9.11 shows that among the employed Conventional ML models RF is the most appropriate. The RF-based models have achieved 76.60% and 73.68% accuracy for MFCCs 26 and MFCCs 13 respectively. The MDI-based important approach also shows its supremacy. Table 9.12 again shows the supremacy of 1D CNN among the employed LT models. It has achieved 83.46 and 81.18% of accuracy for MFCCs 26 and MFCCs 13 respectively. The MDI-based important approach also shows its supremacy.

9.4.5 Table 9.12 shows that for MFCCs 40, both CSL and LT-based ML models' best accuracy value achieved is 94.12% and 98.52% respectively, and the employed

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm (LSTM)                  (None, 13, 64)            16896

 activation (Activation)      (None, 13, 64)            0

 dropout (Dropout)            (None, 13, 64)            0

 flatten (Flatten)            (None, 832)               0

 dense_4 (Dense)              (None, 4)                 3332

 activation_1 (Activation)    (None, 4)                 0

=================================================================
Total params: 20,228
Trainable params: 20,228
Non-trainable params: 0
```

Fig. 9.5 The architecture of the employed LSTM model for the proposed study

Table 9.6 Evaluation of results by employing CSL and LT algorithms in terms of **accuracy (%)** based on 40 MFCCs, datasets employed are EMODB, RAVDESS, and COMBINED

Employing 40 MFCCs Accuracy (%)								
Dataset	CSL ML models					LT ML models		
	MLP	RF	XGB	kNN	SVM	DNN	CNN	LSTM
EMODB	*94.12*	92.71	91.18	92.65	91.18	95.50	**98.52**	92.64
RAVDESS	69.34	*70.80*	68.61	68.61	66.42	69.89	**75.91**	74.45
CoD	73.66	*75.12*	73.66	75.61	71.71	78.61	**80.00**	78.07

Table 9.7 Evaluation of result by employing CSL algorithms in terms of accuracy (%) based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) dataset engaged EMODB

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	94.12	94.18	94.12	94.12	*95.59*	94.18
RF	89.71	89.24	89.71	89.66	92.65	89.24
XGB	92.42	92.65	92.65	92.02	94.12	92.58
kNN	92.65	*95.59*	94.12	91.18	92.65	91.18
SVM	91.18	92.65	94.12	86.76	89.71	89.71

Table 9.8 Evaluation of result by employing LT algorithms in terms of accuracy (%) based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged EMODB

LT ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	92.65	97.06	97.06	98.53	98.53	97.95
CNN	97.05	**98.52**	**98.52**	98.80	**98.88**	98.82
LSTM	96.05	**98.52**	97.05	95.58	95.58	95.58

Table 9.9 Evaluation of result by employing CSL algorithms in terms of accuracy, based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged RAVDESS

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	68.61	*71.53*	70.07	69.14	*72.99*	69.20
RF	69.34	70.80	68.61	70.45	70.51	*70.58*
XGB	66.15	66.42	66.17	62.04	65.69	64.96
kNN	65.69	65.77	65.75	64.03	64.20	64.13
SVM	65.69	66.15	66.02	66.40	66.42	65.98

Table 9.10 Evaluation of result by employing LT algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged RAVDESS

LT ML Models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	75.18	75.72	76.64	72.26	72.53	71.88
CNN	78.10	**78.11**	77.64	74.56	**74.72**	74.38
LSTM	73.72	75.91	73.72	72.86	73.99	72.96

dataset is EMODB. Table 9.13 displays that the highest accuracy achieved is 95.59% using EMODB as a dataset. MFCCs involved are MDI-based IMP-26 and IMP-13 respectively. For RAVDESS best accuracy achieved is 72.99 and MFCCs involved are MDI-based IMP-13. Finally, for the combined dataset best accuracy achieved is 76.60 and MFCCs involved are MDI-based IMP-26. Table 9.14 demonstrates that for EMODB best accuracy achieved is 98.88 and the MFCCs involved are MDI-based

Table 9.11 Evaluation of result by employing CSL algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged combined

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	71.24	74.63	74.25	70.73	71.19	68.34
RF	76.59	*76.60*	76.58	73.66	*73.68*	70.73
XGB	71.22	71.42	71.32	66.83	66.83	69.27
kNN	72.68	73.66	72.68	69.27	69.76	66.34
SVM	70.24	70.26	70.24	72.68	72.82	70.24

Table 9.12 Evaluation of result by employing LT algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged combined

LT ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	78.05	79.02	78.21	79.89	80.08	77.10
CNN	83.41	**83.46**	81.51	81.12	**81.18**	79.53
LSTM	82.43	81.97	81.54	77.07	77.34	74.18

Table 9.13 Comparative Findings of 40 MFCCs, using CSL and LT Models based on the best accuracy (%) score achieved

Dataset	Accuracy (%) MFCCs 40	
	Classical ML models accuracy (%)	Latest ML models accuracy (%)
EMODB	*94.12*	**98.52**
RAVDESS	70.80	75.91
CoD	75.61	80.00

Table 9.14 Comparative findings of different MFCCs(First26 and 13 or Important 26 and 13 using MDI and Prem) based on the best accuracy(%) obtained, by employing CSL Model, engaged dataset EMODB, RAVDESS, and combined

Dataset	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	94.12	*95.59*	94.12	94.12	*95.59*	91.18
RAVDESS	69.34	71.53	70.07	70.45	*72.99*	70.58
CoD	76.59	*76.60*	76.58	73.66	73.68	70.73

Table 9.15 Comparative findings of different MFCCs(First26 and 13 or Important 26 and 13 using MDI and Prem) based on the best accuracy(%) obtained, employed Model (LT), engaged dataset EMODB, RAVDESS, and combined

Dataset	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	97.05	98.52	98.52	98.80	**98.88**	98.52
RAVDESS	78.10	**78.11**	77.64	74.72	74.72	74.38
CoD	83.41	**83.46**	81.54	81.12	81.18	79.53

Table 9.16 Comparative findings of the execution time of the CSL model based on the maximum time taken

MFCCs 40		
Dataset	CSL ML models Time in seconds	LT ML models Time in seconds
EMODB	0.0180	83.13
RAVDESS	0.0294	145.89
CoD	*0.0330*	**213.45**

Table 9.17 Comparative findings of the execution time using different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) of the CSL Models based on the maximum time taken

Dataset	Execution time in seconds					
	IMP-26 MFCCs (MDI)	26 MFCCs	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	0.0120	0.0120	0.0120	0.0040	0.0041	0.0040
RAVDESS	0.0173	0.0172	0.0170	0.0018	0.0018	0.0018
CoD	0.0290	0.0300	0.0280	0.0230	0.0230	0.0230

Table 9.18 Comparative findings of the execution time using different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) of the LT Model based on the maximum time taken

Dataset	Execution time (in seconds)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	48.66	48.73	45.23	39.41	39.46	39.71
RAVDESS	52.23	52.23	52.11	49.08	49.08	49.90
CoD	147.59	147.85	147.12	103.45	104.22	104.00

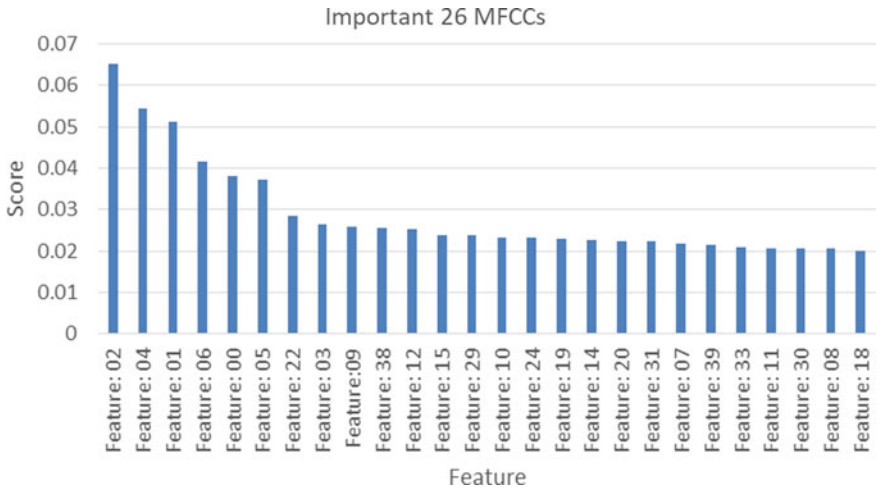


Fig. 9.6 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed EMODB

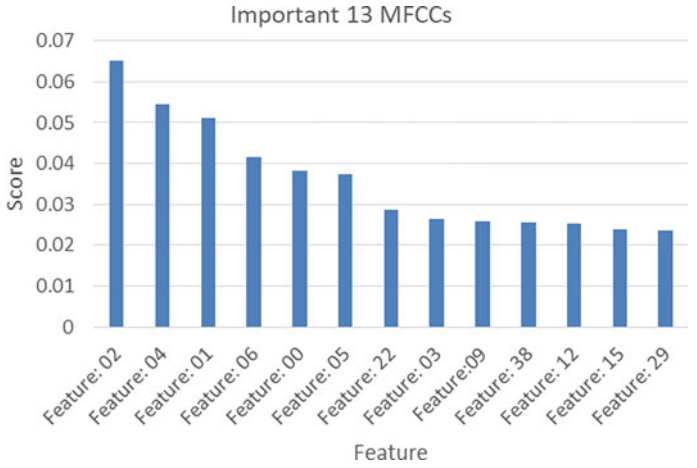


Fig. 9.7 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed EMODB

IMP-13. For RAVDESS and Combined best accuracy achieved is 78.11 and 83.46 respectively. In both cases, MFCCs involved are MDI-based IMP-26.

9.4.6 Tables 9.16, 9.17 and 9.18 shows the time taken to execute increases as the size of the dataset increases, as well as the number of MFCCs increases. It also shows that the execution time of LT-based ML is much higher compared to CSL ML models.

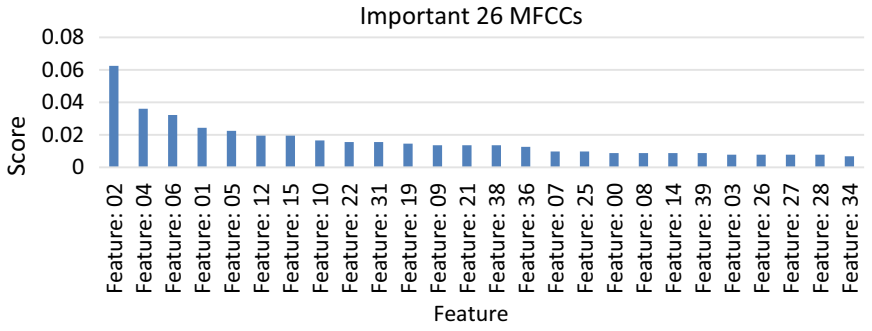


Fig. 9.8 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed EMODB

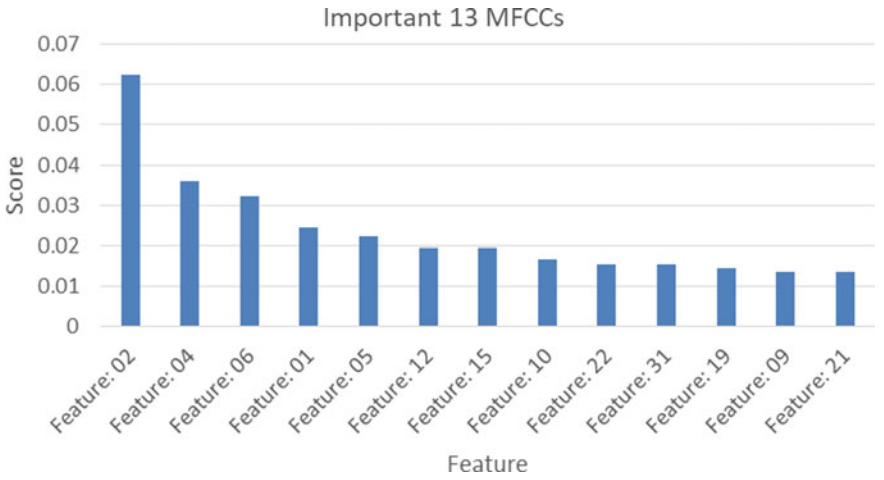


Fig. 9.9 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed EMODB

9.5 Conclusion

The result analysis shows the supremacy of XAI based (MDI, Permutation) approach over the traditional approach by selecting the important MFCCs. This helps to make the system better and smarter. It also shows that among the employed ML models 1D CNN-based model achieved the best accuracy. For the regional dataset 1D, CNN achieved 98.88% accuracy, for the International dataset it has achieved 78.11% accuracy, and for the combined dataset 83.46%accuracy has been achieved. It further shows that MDI based approach outperforms the permutation-based approach. The LT-based ML models show better performance though they are a bit expensive in

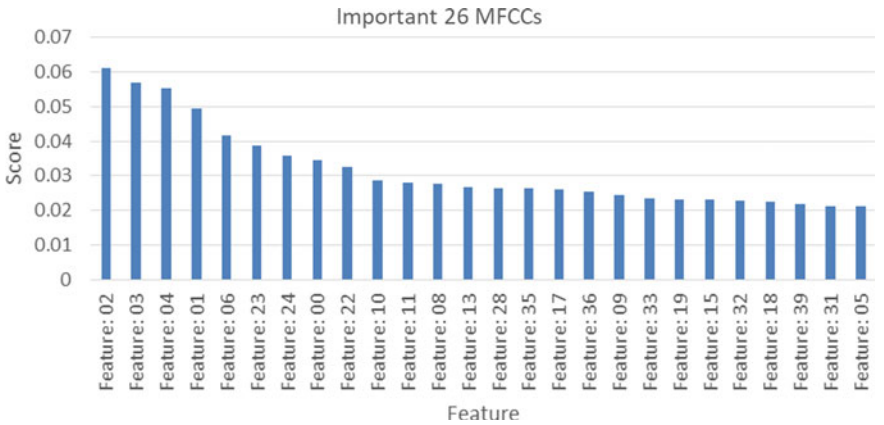


Fig. 9.10 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed RAVDESS

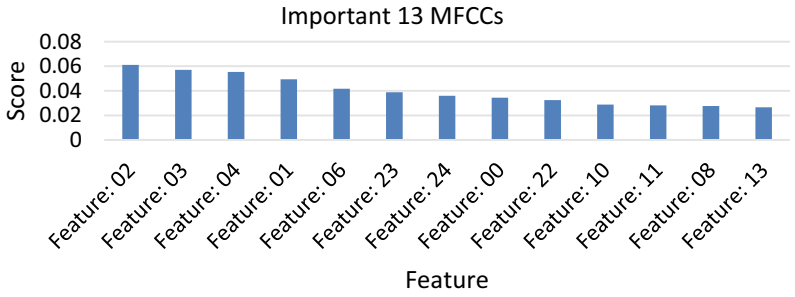


Fig. 9.11 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed RAVDESS

terms of execution times. Thus, a 1D CNN model has been proposed, where important features are extracted and used employing XAI based (MDI) approach.

Limitation and Future Scope: This study has a few limitations. Firstly as a feature, only MFCC(s) have been considered. Secondly, we have only considered two audio datasets and one combined dataset. The inclusion of a more diversified audio dataset may make the system more robust and better. Important MFCCs have been obtained using the RF algorithm. Thus for comparative analysis, other algorithms can be employed in the future for selecting the important MFCCs.

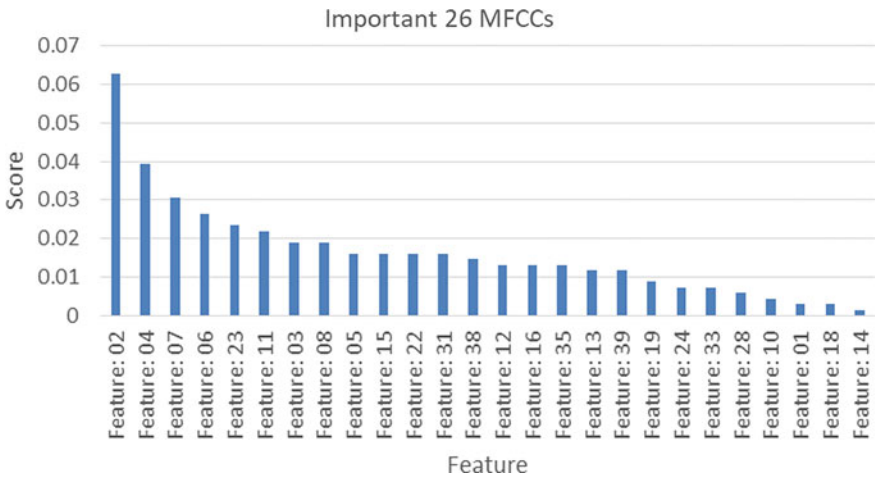


Fig. 9.12 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed RAVDESS

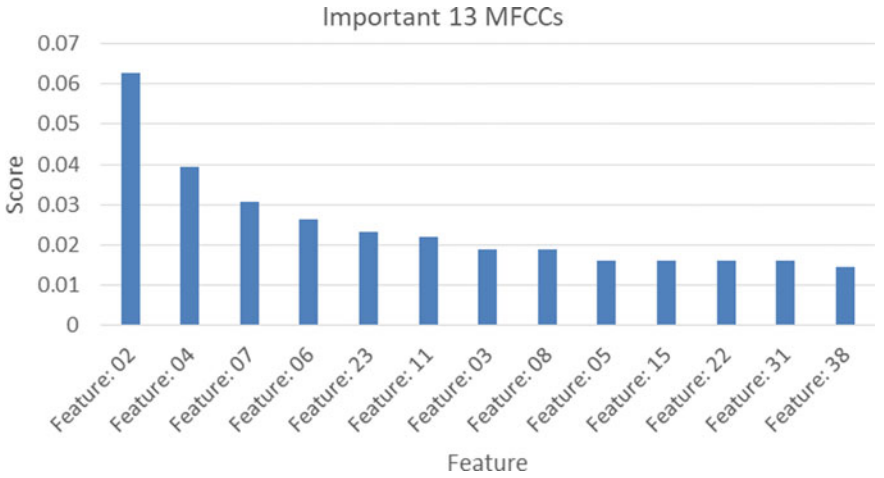


Fig. 9.13 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed RAVDESS

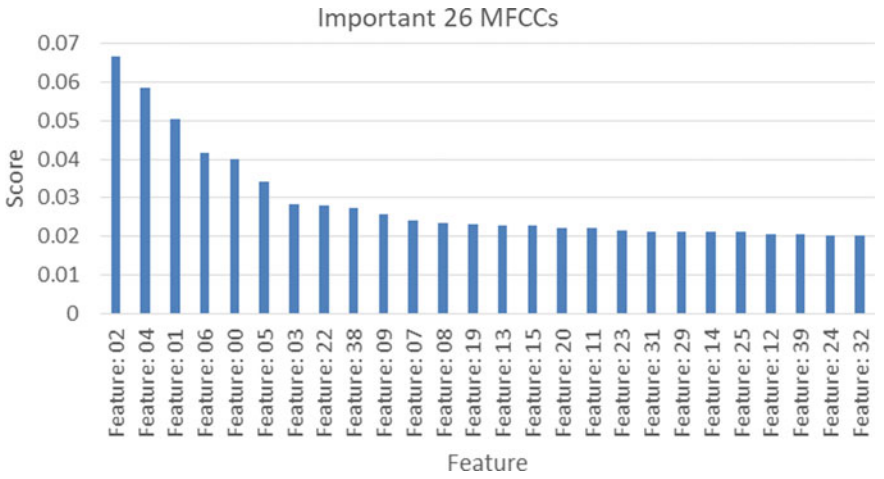


Fig. 9.14 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed combined

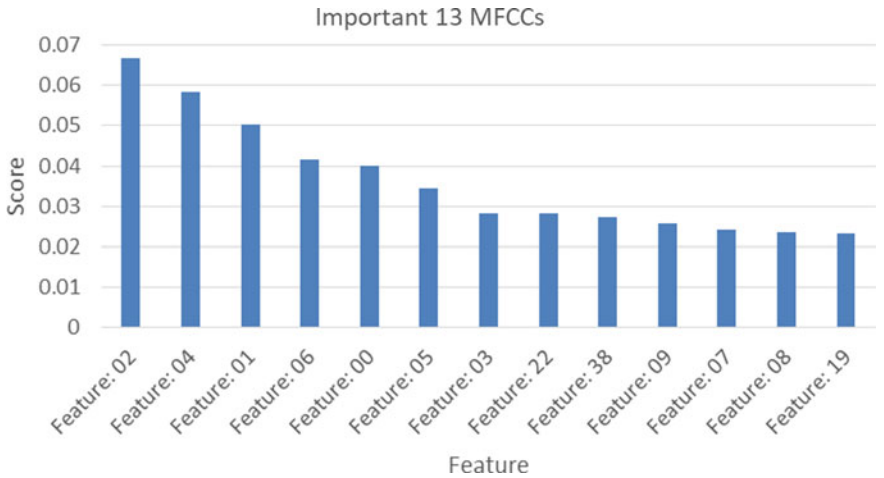


Fig. 9.15 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed combined

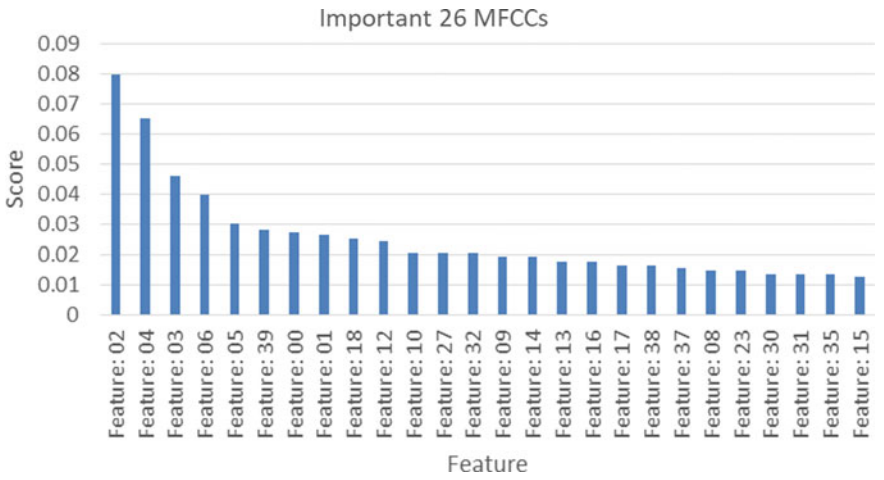


Fig. 9.16 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed combined

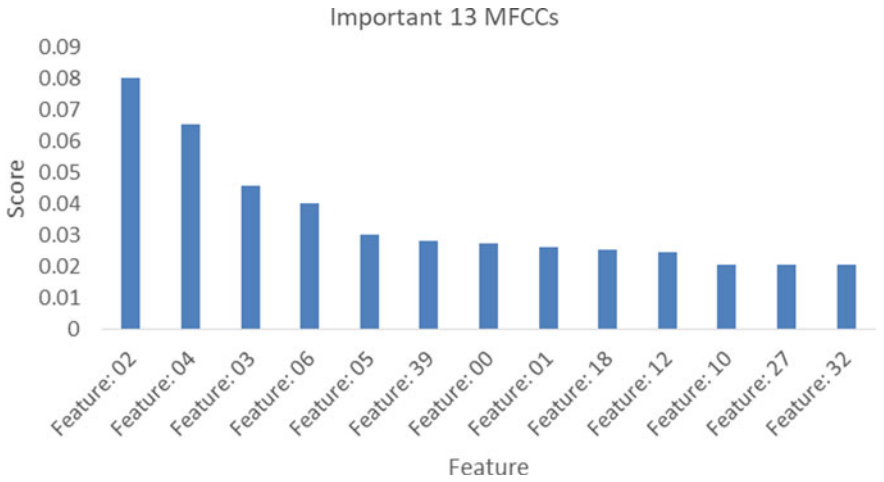


Fig. 9.17 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed combined

Fig. 9.18 Confusion Matrix of adopted 1D CNN. 26 Important MFCCs selected using MDI based approach, Dataset employed RAVDESS

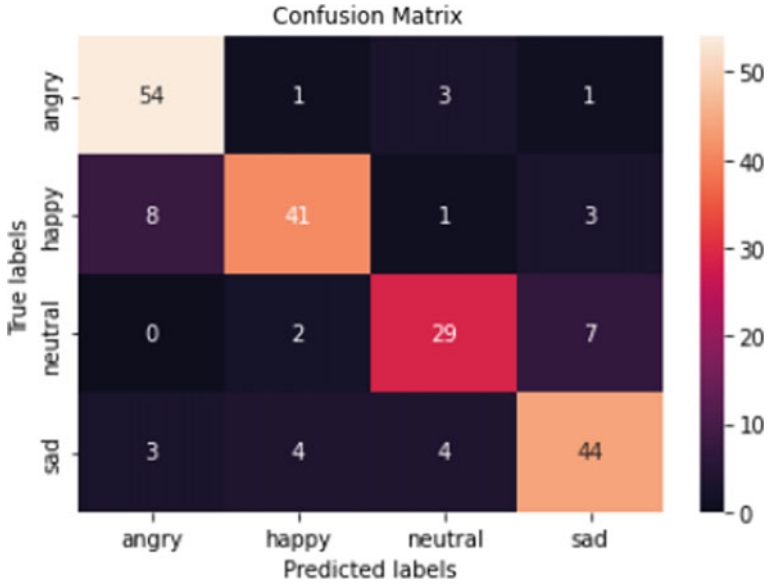


Fig. 9.19 Confusion Matrix of adopted 1D CNN 26 Important MFCCs selected using MDI based approach, Dataset employed combined

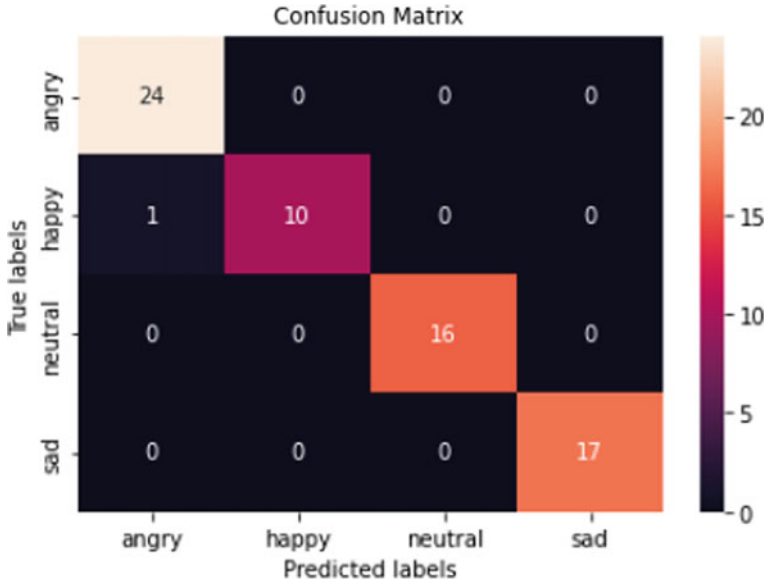


Fig. 9.20 Confusion Matrix of adopted 1D CNN. 13 Important MFCCs selected using MDI based approach, Dataset employed EMODB

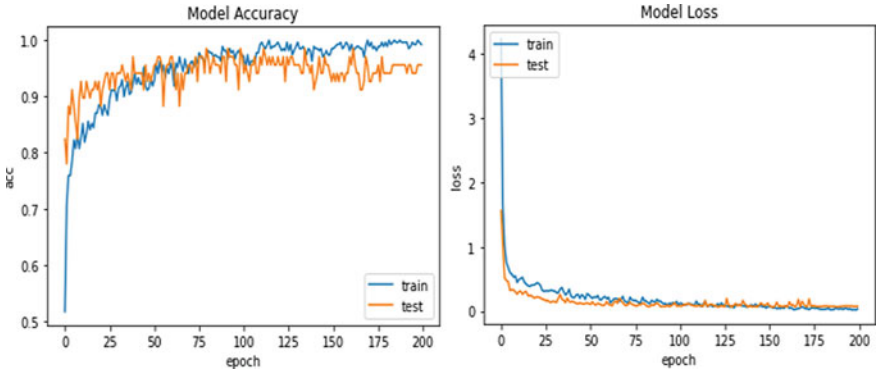


Fig. 9.21 Model accuracy and Model Loss of adopted 1D CNN. 13 Important MFCCs selected using MDI based approach, Dataset employed EMODB

Table 9.19 Evaluation of result based on employed 1D CNN using three benchmark scores against FOUR mental health states. Feature vector employed important 13 MFCCs MDI based

Mental state	Benchmark score			
	Precision	Recall	F1 Score	Support
Angry	0.96	1.00	0.98	24
Happy	1.00	0.91	0.95	11
Neutral	1.00	1.00	1.00	16
Sad	1.00	1.00	1.00	17
Accuracy	–	–	**0.99**	**68**
Macro avg	0.99	0.98	**0.98**	**68**
Weighted avg	0.99	0.99	**0.99**	**68**

References

Akçay, M.B., Oguz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020). https://doi.org/10.1016/j.specom.2019.12.001

Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media and Soc. **20**, 3 973–989 (2018)

Ayadia, E.M., Kamel, S., M, Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**, 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020

Basharirad, B., Moradhaseli, M.: Speech Emotion Recognition Methods: A Literature Review. In: AIP Conference Proceedings vol. 1891, pp. 020105. (2017).https://doi.org/10.1063/1.5005438

Bellotti, K.: Edwards: Intelligibility and accountability: human considerations in context-aware systems. Hum. Comput. Interact. **16**, 193–212 (2009)

Bendre, N., Ebadi, N. Prevost, J.J., Najafirad, P.: Human action performance using deep neuro-fuzzy recurrent attention model. IEEE Access **8**, 57 749–57 761 (2020)

Boles, A., Rad, P.: Voice biometrics: deep learning-based voiceprint authentication system. In: 12th System of Systems Engineering Conference (SoSE), pp. 1–6. IEEE. (2017).

Chatterjee, R., Majumder, S., Sherratt, R.S., Halder, R., Maitra, T., Giri, D.: Real-time speech emotion analysis for smart home assistants. IEEE Trans Consum Electronics **67**(1), 68–76 (2021). https://doi.org/10.1109/TCE.2021.3056421

Diakopoulos, N.: Algorithmic-accountability: the investigation of black boxes. Tow Cent. Digit. JIsm. (2014).

Diakopoulos, N.: Enabling accountability of algorithmic media: transparency as a constructive and critical lens. In: Transparent Data Mining for Big and Small Data, pp. 25–43. Springe. (2017)

EMO-DB: Berlin Database of Emotional Speech, [Online]. 671. http://emodb.bilderbar.info/start.html

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn Res. **20**(177), 1–81 (2019)

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. ORCID: 0000-0001-6482-1973 and Yang, G-Z.: XAI-Explainable artificial intelligence. Sci. Robot **4**(37), eaay7120, (2019). https://doi.org/10.1126/scirobotics.aay7120V

https://dictionary.cambridge.org/dictionary/english/tip-off.

https://www.criminallawyersandiego.com/crimes-police-government/false-report/.

Iqbal, A., Barua, K.: A real-time emotion recognition from speech using gradient boosting. In: International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–5. IEEE (2019)

Koolagudil, S.G., Srinivasa Murthy1, Y.V., Bhaskar1, S.P.: Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. Int. J. Speech Technol. (2018). https://doi.org/10.1007/s10772-018-9495-8

Lalitha, S., Madhavan, A., Bhushan, B., Saketh, S.: Speech emotion recognition. In: Proceedings of the International Conference on Advances in Electronics, Computers and Communications, ICAECC 2014, pp. 1–4. IEEE (2015b). http://doi.org/https://doi.org/10.1109/ICAECC.2014.7002390

Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proceedings of the 11th International Conference on Ubiquitous Computing, pp. 195–204. ACM (2009)

Livingstone, S.R., Thompson, W.F., Wanderley, M.M., Palmer, C.: Common cues to emotion in the dynamic facial expressions of speech and song. Q. J. Exp. Psychol. 1–19 (2018). https://doi.org/10.1371/journal.pone.0196391

McDermid, J.A., Jia, Y., Porter, Z., Habli, I.: Artificial intelligence explainability: the technical and ethical dimensions. Phil. Trans. R. Soc. A **379**, 20200363 (2021). https://doi.org/10.1098/rsta.2020.0363

Panwar, S., Das, A., Roopaei, M., Rad, P.: A deep learning approach for mapping music genres. In: 12th System of Systems Engineering Conference (SoSE) , pp. 1–5. IEEE. (2017)

Parra, G.D.L.T., Rad, P., Choo, K.-K.R., Beebe, N.: Detecting internet of things attacks using distributed deep learning. J. Netw. Comput. Appl. 102662 (2020)

Pinto, MGD. Polignano, M., Lops, P., Semeraro, G.: Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In: EAIS, IEEE (2020). https://doi.org /978-1-7281-4384-22020

Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017). https://doi.org/10.1016/j.inffus.2017.02.003

Saarela, M., Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. **3**, 272 (2021). https://doi.org/10.1007/s42452-021-04148-9

Silva, S.H., Alaeddini, A., Najafirad, P.: Temporal graph traversals using reinforcement learning with proximal policy optimization. IEEE Access, **8**, 63 910 (2020)

Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S., 2016. Emotion recognition from audio signals using support vector machine. In: Proceedings of the IEEE Recent Advances in Intelligent Computational Systems, RAICS, pp. 139–144. IEEE. (2015). http://doi.org/https://doi.org/10.1109/RAICS.2015.7488403

Sweeney, L.: Discrimination in online ad delivery. Commun. ACM **56**(5), 44–54 (2013).

Velliangiria, S., Alagumuthukrishnan, S., Iwin, S., Joseph, T.: A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. **165**, 104–111 (2019). https://doi.org/10.1016/j.procs.2020.01.079

Yang, N., Dey, N., Sherratt, S., Shi, F.: Emotional state recognition for AI smart home assistants using Mel-frequency Cepstral coefficient features. J. Intell. Fuzzy Syst. **39**(2), 1925–1936 (2020). ISSN 1875–8967 (E)

Zarsky, T.: The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci. Technol. Human Values **41**(1), 118–132 (2016)

Chapter 10
Face Mask Detection Based Entry Control Using XAI and IoT

Yash Shringare, Anshul Sarnayak, and Rashmi Deshmukh

Abstract Today it has become mandatory for all the citizens to wear a face mask to protect them from COVID-19. Also taking two doses of vaccine is a must to visiting public places and currently, the only method to verify whether a person is fully vaccinated is by showing a vaccine certificate. The proposed application is helpful for elderly people who find it difficult to use smart phones. The shop owners, offices, banks, or any public place can check for restrictions of entry if anyone is not wearing a mask. As a result, no need for any guard to keep an eye on people. Machine learning techniques with Explainable AI (XAI) can solve these problems easily and results are made understandable to end-users because of the explaining ability and interpretability of neural network models. The system performs well for prediction and gives more accurate and trustworthy predictions. Hence XAI is more reliable in healthcare systems. The proposed system is implemented completely on Raspberry Pi allowing a complete embedded application. The application is developed using Python and HTML. PyCharm/Visual Studio Code with the help of an open-source library is used for training, defining, etc. Machine learning models used for the system are Tensorflow.js, Keras, OpenCV, etc. The whole application can run on a microcontroller such as Raspberry Pi, which allows one to simply plug and play the system at any time.

Keywords CNN · XAI · OpenCV · Mask detection · Raspberry Pi · Bootstrap · HOG

Y. Shringare · A. Sarnayak
Department of Technology, Shivaji University, Kolhapur, Maharashtra, India

R. Deshmukh (✉)
Department of Technology, Shivaji University, Kolhapur, India
e-mail: rvm_tech@unishivaji.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_10

10.1 Introduction

Emerging studies show that wearing a mask can protect us from the spread of COVID-19. Multi-layered masks have good performance in blocking exhaled small droplets. Hence no one without a mask is allowed to enter the mall. The proposed model for face mask detection can solve this type of problem. The proposed model not only detects a face mask but also controls entry points in any mall, Bank, or cinema hall. A neural network with explainability helps to improve the performance by detecting the system well. It finds where the model fails. Due to its interpretability, XAI is used in the application of image processing (Wells and Bednarz 2021). Also, explanations of the prediction made by the neural network model make them more reliable and trustworthy. XAI plays an important role to solve this issue with the help of CNN (Samek et al. 2019).

The problem is the traditional way of ensuring the mask is done by individual check. This is difficult for people especially old people who don't have a smartphone to always show their vaccine certificate. Also, this problem has negatively impacted business owners because it is hectic work to constantly remind people to wear a mask.

Proposed system with web-based application is designed to collect the vaccine certificate and Government ID. Then verification is done using AI and a link is provided to register face encodings also known as embeddings are taken through the web application and stored in a central database using pickle these encodings can be accessed from any scripts. Embeddings are imported to Raspberry Pi from the database RPi uses the Pi Camera and then prompts the user for facial scanning and checks whether the person is already registered on the web portal if successful prompts the user for wearing a mask and upon successful detection of mask send a command through GPIO's to 12 V DC Solenoid lock to open the door.

As per ruralindiaonline.org, there are 137.9 million elderly people (people above 60) living in India and these people often have problems using smartphones, and finding certificates on smartphones proves to be a difficult task. Our application is designed to help such people through our web application and Raspberry Pi. This application is a deep learning model using CNN. XAI provides more visual interpretability than CNN which uses a separate filter to represent the specific object. XAI builds a black box mode first and simultaneously gives a post hoc explanation of how CNN classifies the images and also gives verification of the method used for classification. Hence with XAI, it is possible to design and develop a neural network model which removes bias and presents better predictive performance (Oh et al. 2021). Further, it is integrated and deployed using the library Tensorflow. The web application for this proposed model uses the face-recognition library which is built upon dlib a C++ toolkit.

10.2 Literature Review

Radzi et al. use Convolutional Neural Network (CNN) technique consists of eight layers. Purposed IoT-based face recognition Home security system uses Raspberry Pi. To capture some additional information and to improve accuracy and processing time of the given system had used various activation functions. System is designed with open source software Python and Keras. System uses hundred epocs at the start and for testing twenty epocs per iterations are executed. For IoT Blynk app was used. Blynk is an app that is used to control Raspberry pi and many other microcontrollers. Simply drag and drop the widgets in the app giving a notification to the owner whenever the doorbell was pressed and giving option to open the lock (Syafeeza Ahmad Radzi et al. 2020).

Another work related to this field was by Amritha Nag et al. (2018) in creating system for controlling door access with the use of IOT and face recongition system. It uses OpenCV based face recognition system using Haar classifiers for a face. The pi camera is employed to capture the image. The main aim of this system was to implement entire processing on the Pi computer and sensors for detection of emotions of the face. GSM module is used by installing Subscriber Identification Module (SIM). To control the lock mechanism. This paper describes the use of the Viola-Jones Face Detection Method in three steps. It is implemented with first evaluating the features of image with Haar function and some of the most relevant features are selected for classification with the use of Adaboost algorithm. Further cascaded classifier is used to give the door access (https://docs.opencv.org/).

Adam et al. portrayed a method for advanced facial recognition which firstly involved the detection of faces using HOG (Histogram of Oriented Gradients) for each pixel. Here surrounding pixels are found and gradient painting is done by drawing an arrow that points to a darker region. A further similar image was found which is matching mostly with the HOG pattern. Paper proves the effectiveness of the Jones classifier for matching images with the face landmark estimation algorithm. System uses ML dlib library which detects face on the basis of sixty eight different points on the face which will give uniqueness to face. These specific points are given as input parameters to CNN. After having measurements for each faces to compare them Adam used a SVM classifier. If measurements were similar the persons in the two images were marked identical (Geitgey 2017).

Detection of different poses and emotions of the face is done by open source platform of python as TensorFlow tools and libraries (Tensorflow 2019). Smaller DNN like MobileNet are used to clarify facial images (Friedhoff and Alvarado 2018).

System represents a real time detector for face masks. It uses MobileNetV2 for classification of facial images (Sanjaya and Adi Rakhmawan 2020).

More et al. Uses face recognition for monitoring the attendance (More et al. 2021). System gives attendance analysis for teaching staff.

Anand et al. demonstrated Face Recognition based Attendance system. System uses controlling door access with various DNN tools. Different tools are used by the

system like small and chip computer Raspberry Pi, 360 degree scanner RPLIDAR A1 and deep learning USB Movidius NCS2 etc. (Anand 2021).

Another research improves the performance of the system with the use of RGB image data for avoiding false positive results and attcks of face spoofing. It gives very high accuracy and F1 score for recognition of facial images (Ko et al. 2021).

10.3 Methodology

Following section gives explanation of methodology used by the proposed system (Fig. 10.1).

The integration of the webcam with the website is the first stage in the flow diagram. Further user registers on the web app by providing their Vaccine Certificates and Government ID. Then the user has to provide face encodings which are stored in a database. This completes the use of the web application module. For training in face mask detection, two classes of images were taken with mask and without the mask. Images are pre-processed and then given as input for the CNN model which is implemented using TensorFlow. After the training process, the trained model has loaded into a script this script. The further script imports the face encodings from the web app and scans the video feed from the camera. If the face is found in the database it prompts the user to wear a mask and sends commands to Raspberry Pi's web server to open the door using GPIO.

10.3.1 Web Application Execution

The backend used for this web application is Django and front end tools used are HTML, JavaScript, bootstrap, and CSS etc. Web application collects the vaccine certificate and Government ID. It provides a link to register face encodings and saves them in the database using pickle.

Fig. 10.1 Flow diagram

ML model is implemented with Django.ai ML framework. It gives a set of tools and libraries to develop ML project with ease.

Pickel: A data structure used by complex object hierarchies is serialized when it is converted into character stream with serialization process and transferred across the network. This is done by Pickel module of Python.

10.3.2 Implementation

This contains detail explanation of the different steps used for implementation.

First step is collection of image data for face recognition; further model is trained and tested with various activation functions.

i. **Detection of facial image and data collection**

Face detection is performed based on the following 4 parameters.

a. Haar Features

Square shape function is scaled with the use of Haar function. Considering all the variations in size, and position of all these features, we have calculated about 160,000+ features. This involves a 24 × 24 window consisting of many redundant features and many of them are not useful.

b. Integral Image

Internal image contain each pixel with summarized form of pixels present to left and above.

c. Adaboost

Adaboost eliminates all the redundant features and narrows them down to several thousand very useful features. For final classification weighted combination of selected best feature is done.

d. Cascading

Let us have an input image of 640 × 480 pixels resolution; we need to move this 24 × 24 window all through the image. We must analyze 2500 features obtained by Adaboost for each 24 × 24 window. Further, a linear combination of all those 2500 outputs is used to see whether it exceeds a given threshold. Finally, a decision is taken whether a face is detected or not. Instead of applying 2500 features on every single 24 × 24 window at a time, we use cascades.

We have defined two class paths as with mask and without the mask. Then each frame captured by the camera is read and faces are detected in the capture frame by using the above cascade classifiers. A further rectangle is drawn around the face to crop that part and the images are saved in respective folders (Fig. 10.2).

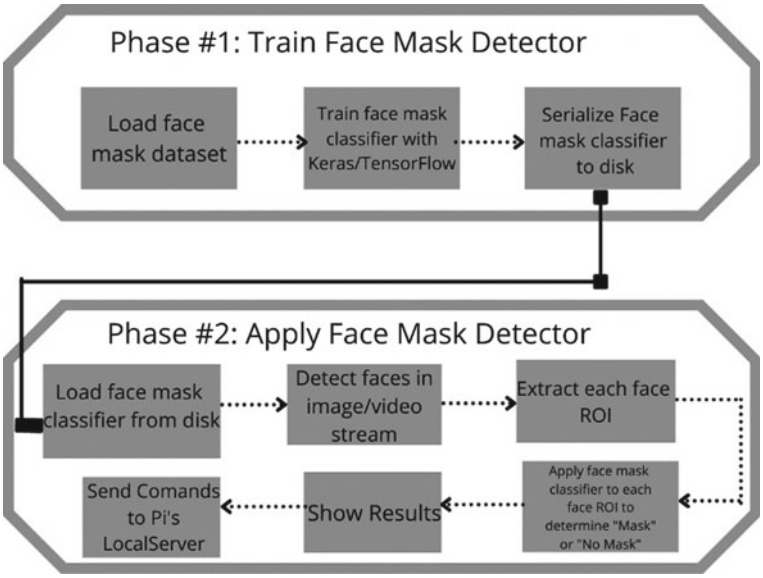


Fig. 10.2 Train and apply

ii. Model Training

Model Training was done using Tensor flow. Some layer modules from the Tensor flow library were used in the proposed system.

Dense: It gives fully connected NN layer where each neuron gets output from previous layer and further passes one output to next layer.

MaxPooling2D: MaxPooling2D is useful for down sampling the input by taking the maximum value for each channel of the input over an input window (Fig. 10.3).

Flatten: To flatten multi-dimensional data into one-dimensional data, utilize the Flatten layer.

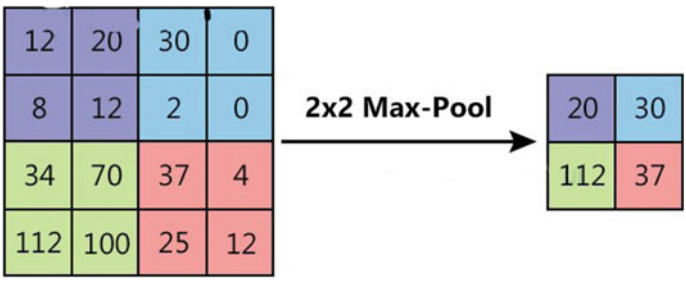


Fig. 10.3 Down sampling of input image

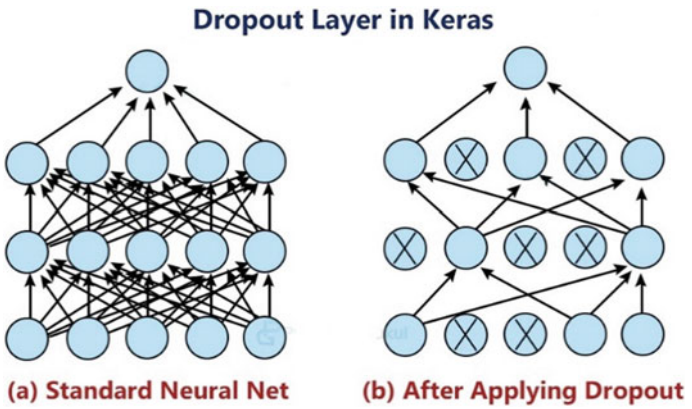


Fig. 10.4 Droupout layer

Dropout: Dropout is a technique for preventing over fitting in neural networks by disregarding neurons at random during the training phase (Fig. 10.4).

i. **Data Preprocessing**

In the data preprocessing; part images from both directories (class index) with_mask and without mask were collected and stored in list x (for image data) and y (for class index) respectively. Resize function from OpenCV is used to resize image to 100 × 100 sizes.

Resized image array is divided by '255' sothat deep learing performs well. Further all the lists were converted to NumPy array as the model feeds on more number of images simultaneously. Generally this process is known as a batch. Further reshaping of the array is needed.

To create model architecture sequential API is used. This plain stack of layers contains each layer with one input and one output. Model. Add () function adds a layer to the model this Model architecture contains four Convolution layers and three Dense Layers. A batch of arrays is fed into the first convolution layer. 'input shape' determines the shape of the input array. Final layer has two output so data is classified into two different classes. 'Softmax' is the last layer's activation function as the categorical data was used. Finally, the model was compiled, and the loss function "sparse_categorical_crossentropy" is used.

10.3.3 Activation Functions

Activation functions comprise mathematical formulas for finding the output of a neural network in Deep learning models. Each neuron in the network has an Activation Function that determines whether or not to stimulate that neuron. The relevance of each neuron's input to the model's prediction is used to make this decision.

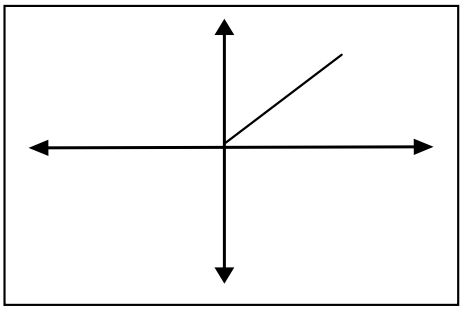


Fig. 10.5 Graph of RELU activation function

The activation functions employed in the proposed model are listed below:

i. **Softmax Activation Function**

The Softmax function is frequently used to solve classification problems involving several classes. It's used to convert the output of neural networks into a number between 0 and 1. It's utilized to signify a certain level of "probability" in the network's output. As we are dealing with probabilities, the Softmax function's scores will all add up to 1.

- Formula of Softmax Activation Function:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{10.1}$$

ii. RELU Activation Function (Rectified Linear Unit):

The RELU layer finds and eliminates any negative values from the filtered image and replaces them with zeros while leaving only positive values. To keep the values from aggregating to zero, RELU is utilized (Fig. 10.5).

$$f(x) = \begin{cases} 0 \; if \; x < 0 \\ x \; if \; x \geq 0 \end{cases} \tag{10.2}$$

10.3.4 Raspberry Pi Webserver

The web server is implemented on Raspberry Pi HTML page was developed to control GPIO over the internet. If a mask is detected the above model will execute a command on Pi's HTML page using Selenium Web Driver. Admin connected to the local LAN network will also be able to override the system to control the GPIO (Fig. 10.6).

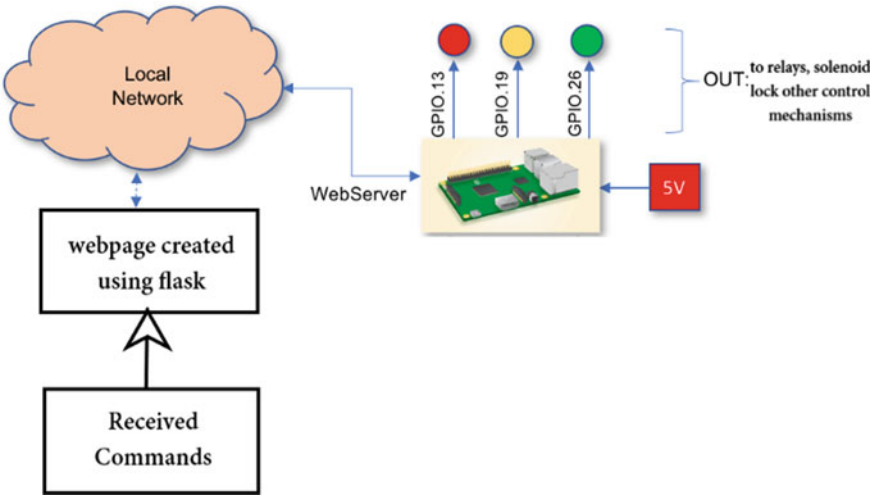


Fig. 10.6 Flow diagram of Pi webserver

10.4 Results

10.4.1 Dataset

We have created our database for the training. One with the mask as shown in Fig. 10.7 and one without the mask shown in Fig. 10.8

First, images for the mask class were captured and stored in the class_path as a mask. Then get_detection (frame) function. İs used to detect a face. Then img[y:y + h, x:x + w] extracts some of the portion of image and cv2.imwrite() used for storing the image details.

The count was set to "count >=500" so 500 images were collected and the loop closed automatically. After collecting with the mask shown in Fig. 10.8. We have collected data without mask and with mask in Fig. 10.7.

10.4.2 Model Summary

The model was built using layers API: Sequential model which is a linear stack of layers made by passing a list of layers to the add() function. Validation is one of the key advantages of working with a LayersModel as it compels to provide the input shape, which will be further used to validate the input.

Figure 10.9 gives a summary of model.summary() function. This includes, all layers in the model with name and type, each layer's output form, each layer's

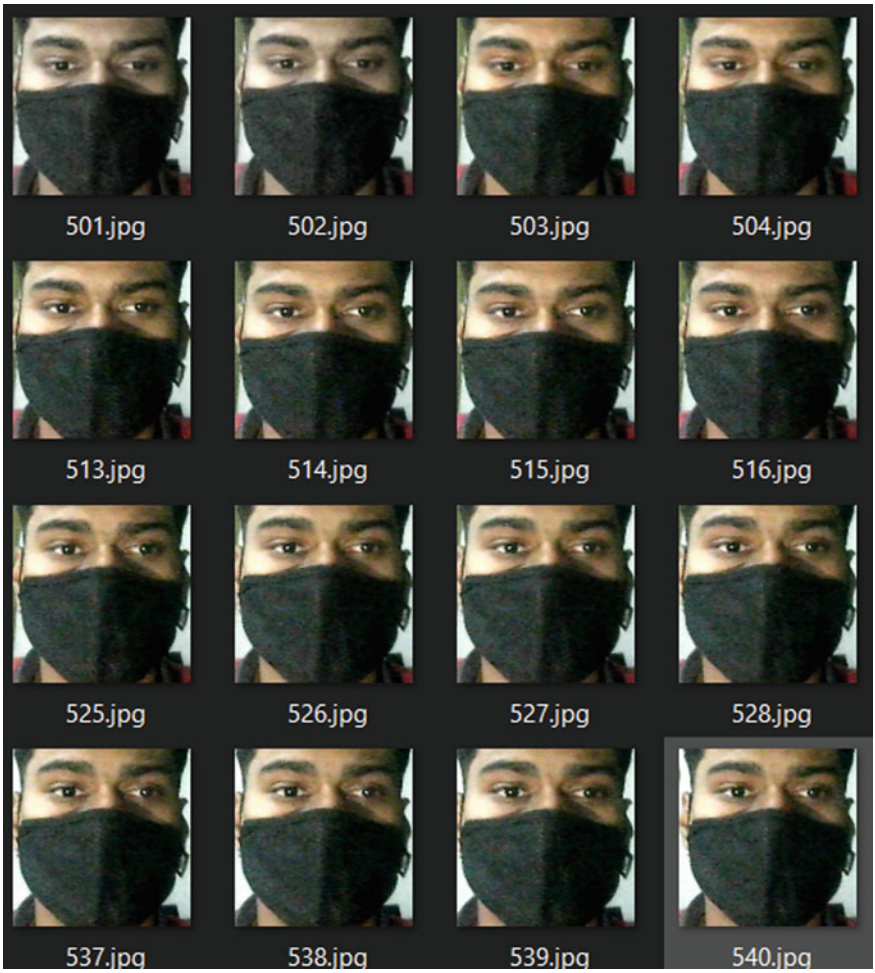


Fig. 10.7 Dataset with mask

weight parameters, generic topology of model and model's total number of trainable and untrainable parameters.

10.4.3 Model Evaluation

When we run the above statement we get: Output: 0.0282 loss–0.9950 accuracy. In the test dataset, we can observe that the model has an accuracy of above 99%. This shows a test accuracy of 99%, which should be acceptable. What it means to us that in 1% of the cases, the mask or without mask would not be classified correctly.

Fig. 10.8 Dataset without mask

10.5 Conclusion

Due to COVID-19, most malls and public places need to check a vaccine certificate and mask of each individual at the entrance. It becomes difficult for elderly people to find their certificates on mobile phones and many of them don't even have smart phones. The proposed application helps such people by using facial recognition. With the proposed system, we can detect the mask on a person's face and allow entry. Deep learning and machine learning models are used for image classification. In many cases, DNN results are not easily known to the end-user. But more understandable

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 98, 98, 64)        1792
_________________________________________________________________
activation (Activation)      (None, 98, 98, 64)        0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 49, 49, 64)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 47, 47, 256)       147712
_________________________________________________________________
activation_1 (Activation)    (None, 47, 47, 256)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 23, 23, 256)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 21, 21, 128)       295040
_________________________________________________________________
activation_2 (Activation)    (None, 21, 21, 128)       0
_________________________________________________________________
dropout (Dropout)            (None, 21, 21, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 19, 19, 32)        36896
_________________________________________________________________
activation_3 (Activation)    (None, 19, 19, 32)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 9, 9, 32)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 9, 9, 32)          0
_________________________________________________________________
flatten (Flatten)            (None, 2592)              0
_________________________________________________________________
dense (Dense)                (None, 100)               259300
_________________________________________________________________
dense_1 (Dense)              (None, 16)                1616
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 34
_________________________________________________________________
activation_4 (Activation)    (None, 2)                 0
=================================================================
Total params: 742,390
Trainable params: 742,390
Non-trainable params: 0
```

Fig. 10.9 Model.summary()

results are achieved with explainable AI (XAI) which will give the most trustworthy and interpretable results.

Most of the XAI uses RNN and the classification of an image is done with CNN. With the use of XAI, the CNN model results in an output image that shows the most relevant features needed for prediction. Local interpretable model- Agnostic Explanation algorithm is used with CNN to give explanations of prediction used by any classifying model. Explanations are given in the form of feature relevance or contribution to the prediction of a certain sample dataset. CNN Algorithm used in the proposed system is divided into four stages Haar Features Selection, Integral Images, AdaBoost and Cascading Classifier. The project is the epitome of IoT with Raspberry pi. Experimental results show that the proposed system with CNN gives more accurate and reliable results for image detection and classification. This Application has many advantages over the traditional way of enforcing strict lockdown rules and regulations in a country like India where the population is highly dense. The proposed application will provide an easy way of COVID prevention. With the use of XAI, it is possible to present more interpretable results of face mask detection and classification.

References

Anand, P.V.: Facial-attendance-on-Pi-with-LIDAR. (2021). https://github.com/AdroitAnandAI/FacialAttendance-on-Pi-with-LIDAR

Friedhoff, J., Alvarado, I.: Move mirror: an AI experiment with pose estimation in the browser using tensorflow.js. TensorFlow Medium. (2018). https://medium.com/tensorflow/move-mirror-an-ai-experiment-with-pose-estimation-in-the-browser-using-tensorflow-js-2f7b769f9b23/

Geitgey, A.: Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning. (2017). https://github.com/ageitgey/face_recognition

Introduction to OpenCV. https://www.geeksforgeeks.org/introduction-to-opencv/, https://docs.opencv.org/

Ko, K., Gwak, H., Thoummala, N., Kwon, H., Kim, S.H.: SqueezeFace: integrative face recognition methods with LiDAR sensors. J. SensS. **2021** (2021). https://doi.org/10.1155/2021/4312245

More, S., Kadam, N., Savla, S., Shelar, S., Shah, A., Mali, S.: CaptureIt!-A web-based attendance system using face recognition. In: IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), pp. 91–96. (2021). https://doi.org/10.1109/RTSI50628.2021.9597350

Nag, A., Nikhilendra, J.N., Kalmath, M.: IOT based door access control using face recognition. In: 3rd International Conference for Convergence in Technology (I2CT), pp. 1–3. (2018). https://doi.org/10.1109/I2CT.2018.8529749

Oh, J.H., Choi, W., Ko, E., Kang, M.: PathCNN: interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma. Bioinformatics, **37**(1), i443–i450 (2021). https://doi.org/10.1093/bioinformatics/btab285

Syafeeza Ahmad Radzi, M.K., Mohd Fitri Alif, Y., Nursyifaa Athirah, Jaafar, A.S.: IoT based facial recognition door access control home security system using raspberry pi. Int. J. Power Electron. Drive Syst. **11**(1), 417 (2020). https://doi.org/10.11591/ijpeds.v11.i1.pp417-424

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K.: Explainable AI: interpreting, explaining and visualizing deep learning. In: Lecture Notes in Computer Science book series (LNCS), vol. 11700. (2019). https://doi.org/10.1007/978-3-030-28954-6. ISBN 978-3-030-28953-9

Sanjaya, S.A., Adi Rakhmawan, S.: Face Mask Detection Using MobileNetV2 in The Era of COVID-19 Pandemic. In: International Conference on Data Analytics for Business and Industry: Way towards a Sustainable Economy (ICDABI), pp. 1–5. https://doi.org/10.1109/ICDABI51230.2020.9325631

Tensorflow, J.S: Machine learning for the web and beyond. https://proceedings.mlsys.org/paper/2019/file/1d7f7abc18fcb43975065399b0d1e48e-Supplemental.pdf

Wells, L., Bednarz, T. Explainable AI and reinforcement learning—a systematic review of current approaches and trends. (2021). https://doi.org/10.3389/frai.2021.550030

Chapter 11
Human-AI Interfaces are a Central Component of Trustworthy AI

Markus Plass, Michaela Kargl, Theodore Evans, Luka Brcic, Peter Regitnig, Christian Geißler, Rita Carvalho, Christoph Jansen, Norman Zerbe, Andreas Holzinger, and Heimo Müller

Abstract This chapter demonstrates the crucial role that human-AI interfaces play in conveying the trustworthiness of AI solutions to their users. Explainability is a central component of such interfaces, particularly in high-stake domains where human oversight is essential: justice, finance, security, and medicine. To successfully build and communicate trustworthiness, a user-centered approach to the design and development of AI solutions and their human interfaces is essential. In this chapter, we explain how proven methods for stakeholder analysis and user testing from human-computer interaction (HCI) research can be adapted to human-AI interaction (HAII) in support of this goal. The practical implementation of a user-centric approach is described within the context of AI applications in computational pathology.

11.1 Introduction

The prevalence of Artificial Intelligence (AI) in daily life is ever-increasing. It is integrated into smartphones and consumer goods, transforming the role of the user (Harper et al. 2020) and the human-machine interface. While the traditional human-computer interface simply represents the input-output (I/O) surface of a device (Holzinger 2004), or web page (Ebner et al. 2007), human-AI interfaces transcend the simple I/O paradigm. Besides enabling intelligent interaction via voice or facial recognition, human-AI interfaces can learn from users' behavior, react adaptively, and make predictions about future actions (Holzinger et al. 2022). Accordingly, the scope and challenges of Human–AI Interaction (HAII) research (Xu et al. 2021) differs from that of the traditional field of Human–Computer Interaction (HCI) (Dix et al. 1993). For example: AI chatbots can express human-like communication behavior (Przegalinska et al. 2019); AI-based natural language translation systems show contextual understanding (LeCun et al. 2022); AI-based programs for music co-

M. Plass (✉) · M. Kargl · T. Evans · L. Brcic · P. Regitnig · C. Geißler · R. Carvalho · C. Jansen · N. Zerbe · A. Holzinger · H. Müller
Medical University Graz, Graz, Austria
e-mail: markus.plass@medunigraz.at

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_11

creation can generate non-deterministic output (Louie et al. 2020); AI systems can collaborate with humans in teams (Calero Valdez et al. 2012; Robert et al. 2016), augment human intelligence (Crisan and Correll 2021; Holzinger 2016), and continuously learn from user behavior (Ortigosa et al. 2014).

AI has the potential to bring about a range of benefits to society, support individual and social well-being, enhance innovation and progress, and help to realize sustainable development goals (European Commission, Directorate-General for Communications Networks, Content and Technology, 2019). Regarding the case in point, AI applications in healthcare support personalized and precision medicine, drug development, critical surgeries, clinical decision and diagnosis support, medical image processing, and early detection of disease (Rajpurkar et al. 2022).

However, alongside these opportunities, the broadening application of AI brings novel risks and side effects. Fear of negative consequences, whether misplaced or valid, may also result in underuse and/or over-regulation of AI systems, leading to opportunity costs for individuals and societies (Floridi et al. 2018). Therefore, both benefits and risks must be addressed adequately to give people and societies the confidence to accept AI-based solutions, and to trust in their development, deployment, and usage, even in areas where stakes are high, such as medicine, justice, finance, and security. The trustworthiness of AI systems is a prerequisite for their uptake (European Commission 2021).

According to the *High-Level Expert Group on AI*, established by the European Commission in 2018, trustworthy AI has three components (European Commission, Directorate-General for Communications Networks, Content and Technology 2019):

(a) it should be compliant with the law
(b) it should be robust (i.e., safe, secure, and reliable to not cause unintentional harm)
(c) it should be in alignment with the four ethical principles respect for human autonomy, prevention of harm, fairness, and explicability.

This book chapter illustrates the central role that explainability and human-AI interfaces play in realizing, communicating, and verifying the trustworthiness of AI systems and the importance of a user-centered approach to the design and development of these components. The next section describes regulatory requirements for trustworthy AI and the role of human-AI interfaces in fulfilling these. Section 11.3 discusses explicability as one of the core components of trustworthy AI, and demonstrates that explainable AI is key to building trustworthiness. Section 11.4 explains why a user-centered approach is essential for achieving highly explainable and trustworthy AI systems and introduces stakeholder analysis, personas, and user-testing as valuable methods aiding user-centered design and development of AI solutions. Section 11.5 shows, with the aid of the use-case of AI applications in computational pathology, how these methods can be applied to develop human-AI interfaces that support trustworthiness.

11.2 Regulatory Requirements for Trustworthy AI

As described above, one of the three components of trustworthy AI is compliance with the law (European Commission, Directorate-General for Communications Networks, Content and Technology 2019). The *Artificial Intelligence Act* (European Commission 2021) proposed by the European Commission in 2021 is the first legal framework aimed specifically at fostering AI trustworthiness. It sets out requirements that are mandatory for all AI systems that pose significant risks to the health and safety or fundamental rights of persons (European Commission 2021). Communication to the users is a recurring feature in many of the requirements stipulated. Thus, human-AI interfaces play a central role in the fulfillment of these requirements, as illustrated in the following paragraphs:

Communicate the AI system's intended purpose and associated risks: Point 4 of article 9 'Risk management system' of the Artificial Intelligence Act specifies: "*The risk management measures ... shall be such that any residual risk ... is judged acceptable, provided that the AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user*" (European Commission 2021). To meet this requirement, human-AI interfaces must clearly communicate to users the intended purpose of an AI system as well as the residual risks associated with its usage.

Communicate the AI system's result and all information needed for its interpretation: Point 1 of article 13 'Transparency and provision of information to users' of the Artificial Intelligence Act states: "*AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and the provider ....*" (European Commission 2021). To support these demands, human-AI interfaces must clearly communicate to users the AI system's output together with all information needed for the correct interpretation of this output.

Communicate instructions for use of the AI system: Point 2 of article 13 of the Artificial Intelligence Act demands: "*AI systems shall be accompanied by instructions for use ... that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users*" (European Commission 2021); and point 3 of article 13 of the Artificial Intelligence Act specifies in detail all information that shall be included in these instructions for use, such as for example "*identity and the contact details of the provider ... characteristics, capabilities, and limitations of performance of the high-risk AI system ... human oversight measures ... expected lifetime of the high-risk AI system and any necessary maintenance and care measures to ensure the proper functioning of that AI system ....*" (European Commission 2021). Human-AI interfaces can help to fulfill this requirement either by providing information on how to access the instructions for use or by conveying all information constituting instructions for use to the user directly.

Support human oversight of the AI system: Article 14 of the Artificial Intelligence Act calls for 'human oversight', and explicitly mentions the important role of the human-machine interface as a tool enabling humans to complete this task: "*AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools that they can be effectively overseen by natural persons during the period in which the AI system is in use*" (European Commission 2021). Point 4 of article 14 describes in detail all functionalities and features that human-AI interfaces must provide to support human oversight: According to point 4 of article 14, AI systems *"shall enable the individuals to whom human oversight is assigned to do the following:*

(a) *fully understand the capacities and limitations of the high-risk AI system ...;*
(b) *remain aware of the possible tendency of automatically relying or over-relying on the [AI system's] output ('automation bias'), ...*
(c) *be able to correctly interpret the AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;*
(d) *be able to decide, in any particular situation, not to use the AI system or otherwise disregard, override or reverse the output of the AI system;*
(e) *be able to intervene on the operation of the AI system or interrupt the system"* (European Commission 2021).

Support the AI system's cybersecurity: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act requires that "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of ... cybersecurity*" (European Commission 2021). Human-AI interfaces have important functions with respect to the AI system's vulnerability to cyber-attacks, for example, by enabling user authentication or by conveying security alerts to the user.

Communicate the AI system's accuracy: Article 15 'Accuracy, robustness and cybersecurity' of the Artificial Intelligence Act specifies "*AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy .... The levels of accuracy and the relevant accuracy metrics ... shall be declared in the accompanying instructions of use*" (European Commission 2021). This means that human-AI interfaces shall always provide the user with information about the system's current accuracy so that the user can assess whether or not this level of accuracy is appropriate for the task at hand.

Support robustness of the AI system and prevent user errors Point 3 of article 15 of the Artificial Intelligence Act calls for an AI system's robustness and fault tolerance also specifically with respect to user errors: *"AI systems shall be resilient as regards errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems"*(European Commission 2021). To fulfill this requirement, the human-AI interface on the one hand plays an important role in providing clear but

graceful feedback to the user when a user-error has occurred. On the other hand, as we know from usability research (Norman 2013) that the human-machine interface is also crucial for preventing the user from both conscious mistakes and unconscious slips during their interaction with the system.

11.3 Explicability—An Ethical Principle for Trustworthy AI

Compliance with the law is only one of the three components of trustworthy AI. Another is adherence to ethical principles and values, of which four are explicitly named by the High-Level Expert Group on AI: three traditional bioethics principles (human autonomy, prevention of harm, and fairness), which are in turn based on those described in the Charter of Fundamental Rights of the European Union (European Parliament, the Council and the Commission 2012), and a fourth: *explicability* (European Commission, Directorate-General for Communications Networks, Content and Technology 2019).

Explicability is a new ethics principle specifically relating to AI. It relates to the tendency for AI systems to act on the basis of complex internal processes that are invisible and/or unintelligible to humans (Floridi et al. 2018), rendering their decision-making processes difficult to understand, interpret, and explain (Holzinger et al. 2017). These are crucial issues for trustworthiness, validation, and acceptance of AI (Ziefle et al. 2013). According to Floridi et al. (2018), explicability recognizes the need to understand and hold to account the decision-making processes of AI.

To address the challenge of explicability, the field of *explainable AI* (XAI) research strives to provide insights into how a given AI model works and why it generates a particular result (Holzinger et al. 2018; Longo et al. 2020). There is a jumble of terms related to this concept in the XAI literature: with the terms explainability and interpretability often being used interchangeably (Zhou et al. 2021). Moreover, a variety of terms, including *transparency*, *accountability*, *intelligibility*, *understandability*, and *interpretability*, *comprehensibility* are used, sometimes interchangeably, sometimes with subtle differences in meaning that vary according to author. Other times, these terms are used without defining their specific meaning, or with one same term used for different meanings, or many different terms all referring to the same concept (Lipton 2018).

Gilpin et al. (2018) describe the concept of explainability as a combination of *interpretability* and *fidelity*, both of which are needed to achieve explainability. Here, interpretability refers to how understandable an explanation is for a human, and fidelity describes how accurately an explanation depicts the behavior of the AI model over the entire feature space. However, this often entails a trade-off between these two qualities, whereby it is difficult to simultaneously achieve both high interpretability and high fidelity: The most comprehensive explanation may not be easily interpreted by a human, and an intuitive explanation may not be sufficiently complete in its

coverage of other usage scenarios (Gilpin et al. 2018). To reach optimal explainability, it is, therefore, necessary to assess the relative importance of each of these explainability properties in a specific application context.

Miller (Miller 2019) states that we know from social sciences that usually "*people ask for 'everyday' explanations of why specific events occur, rather than explanations for general scientific phenomena*" and he argues that this holds also in the context of Artificial Intelligence (Miller 2019). To be useful, any explanation must fit the tasks and goals of the receiver of this explanation. Therefore, for an efficient and effective explanation component in an AI system, it is crucial to take into account **who** uses **which** type of AI-solution for **what** purpose, and **how** the human-AI interface is designed (Müller et al. 2022).

11.4 User-Centered Approach to Trustworthy AI

For achieving explainability, as a precondition to trustworthiness, it is critical to develop a profound and comprehensive understanding of the purpose and context of the AI application in question. This includes detailed knowledge of the stakeholders who need to understand and interpret the results provided. With respect to this deep understanding of stakeholders, the article 9 of the Artificial Intelligence Act mandates that *"due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used"* (European Commission 2021). To this end, the following section describes methodologies for generating the rich stakeholder profiles necessary for meeting these requirements.

11.4.1 Stakeholder Analysis and Personas for AI

To achieve the aforementioned requirements for trustworthy AI, it is necessary to focus on users and use-cases throughout the conception, scoping, and implementation stages of AI application development. For traditional computer applications, such a user-centered approach (Holzinger et al. 2005) has gradually been adopted over the past four decades. There is a large set of proven tools and methodologies available for the user-/human-centered design of conventional computer systems (Vredenburg et al. 2002). However, due to the specific characteristics of AI systems, many of these existing HCI tools and methods will need to be adapted and extended to effectively support their human-centered design and development (Xu et al. 2021).

One of the existing methods successfully applied in user-centered design of conventional computer applications is that of *Personas*. This method was introduced for user-centered interaction design by Alan Cooper in 1999 (Cooper and Saffo 1999). Personas are hypothetical user archetypes that help designers and developers to empathize with the target users, to focus on the needs and goals of these users

throughout the product development process (Miaskiewicz and Kozar 2011; Nielsen 2018), and to ultimately design and develop effective, easy-to-use products (Nielsen 2019).

In traditional HCI, personas comprise the following aspects: context and environment, tasks and workflows, skills and knowledge, personal traits, goals and values, motivations and frustrations. To adapt the personas method to the context of HAII, three additional aspects describing the user's attitude to AI solutions should be taken into account (Holzinger et al. 2022):

(a) *Trust*—How much trust does the user have in the decisions/output of the AI system?
(b) *Acceptance*—Does the user accept (and follow) the decision of the AI system?
(c) *Assent*—Is the user willing to accept and use the support of the AI system?.

Furthermore, user requirements for AI applications go beyond the needs and requirements known to traditional HCI: i.e., those related to functionality, physiology, psychology, safety, usability, and user experience (Law et al. 2009). In HAII, additional needs related to explainability, decision-making authority, ethical issues, and emotion are also taken into account (Müller et al. 2021; Xu et al. 2021).

Based on the procedures for the creation of personas described in literature (Cooper and Reimann 2003; Holzinger et al. 2022; Nielsen 2019), a 5-step process can be applied to develop user personas for AI (see Fig. 11.1).

As described by Holzinger et al. (2022), a large part of the process of developing personas for AI is similar to 'traditional' (i.e., HCI-oriented) persona development. Details of this process dealing with aspects specific to personas for AI are described in the following paragraphs:

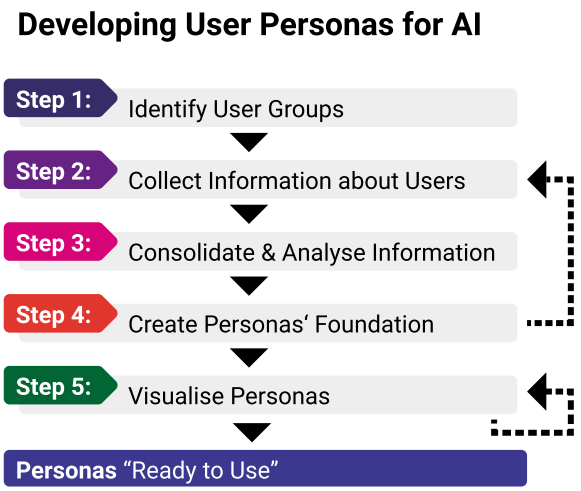


Fig. 11.1 A 5-step process, quite similar to the persona development in human-computer-interaction (HCI), can be applied to develop user personas for Artificial Intelligence (AI) based products

Step 1: Identification of (Potential) User Groups

The first step in developing personas for AI is to compile a comprehensive list of groups of people, who will potentially use the AI application. For AI applications in given a business domain, these user groups may align with job descriptions; in a consumer domain, with lifestyles (Cooper and Reimann 2003). Since each identified user group may be the seed for a distinct persona, instead of restricting the list to the most obvious end-users, a wide view should be applied in this step of the process.

For AI applications in domains where close human oversight is needed, it is necessary to include as potential users, all persons who are required to interpret and understand its results. To avoid misleading outcomes, the identification of (potential) user groups for an AI application should be data-driven. Where this is not feasible, the initial identification of (potential) user groups can be based on the assessment of domain experts.

Step 2: Collection of Information about Users This step has four distinct goals, of which the first two are also found in traditional persona development, and goals 3 and 4 are specific to personas for AI.

The first goal is to get to know (potential) users personally, discover their goals and motivations, and learn about their frustrations and hopes, their skills, education, knowledge, and personality traits. The second is to get to know the users' tasks and discover the context in which they would use the AI solution.

The third goal is specific to application cases and domains in which AI is perceived as new and innovative: find out the users' attitudes toward working with new technologies and innovations. Finally, the fourth goal is to find out the users' attitudes towards machine decisions, under which conditions they would trust a decision of an AI application, under which conditions they would follow the decision of an AI application, and whether or not they would be willing to accept support by an AI application.

Ideally, this collection of information about the users is done through ethnographic interviews or contextual inquiries (Cooper and Reimann 2003; Cooper and Saffo 1999; Pruitt and Grudin 2003). In cases, where such on-site interviews are not feasible, remote interviews should be conducted. In addition, also questionnaires or (internet) research can be utilized to complete the information.

Step 3: Consolidation and Analysis of the Collected Information

The goals of this step in the development process of personas for AI are threefold: First, to gain an overview of the collected information; second, to filter out the important findings, and third, to decide, based on these findings, which personas to develop.

The first task is to gather all collected information in one place. Depending on the kind of collected information, this central information storage can be a database or a simple spreadsheet document. It is important to take care that for each piece of information the connection to the origin is preserved throughout the whole process of organizing, structuring, splitting, and condensing the collected information.

For consolidation and analysis of the collected information various methods can be applied: visualization diagrams (such as, for example, bar charts or scatter plots) support consolidation and analysis of structured categorical or numerical information, 'affinity diagramming' helps with consolidation and analysis of unstructured information; for example, information obtained through open-ended questions in an interview or questionnaire.

These visualization and affinity diagrams demonstrate stratification within user groups, i.e., concerning features such as education, working style, personality traits, etc. It is important that each such cluster, which is related to an aspect that might influence the usage of the product, forms the basis for a persona. Clusters pertaining to user attitudes toward AI or new technologies should always be regarded as important, and should be represented accordingly in the resulting personas, as they strongly influence usage of the AI application.

Step 4: Creating the Foundation for Personas

The aim of this step in the development process of personas for AI is to create for each persona a so-called *foundation document*. This tabulates all information about a specific persona in a structured way. Various structures and templates for foundation documents of traditional personas are described in the literature (Castro and Acuña 2012; Pruitt and Grudin 2003; Pruitt and Adlin 2006). When developing a persona for AI, it is important to include in the foundation document, a specific section regarding the attitude of the persona toward AI, and toward new technology in general (potentially with notice given to whether the former category falls into the latter in the context of the application domain in question).

The purpose of the foundation document is twofold: First, the structured presentation of the collected information for a persona highlights gaps in the data where additional research may be necessary. Second, the foundation document is the basis for any usage of the persona, for example when creating a visualization of the persona or developing their use cases and scenarios.

Step 5: Visualizing Personas

The final step in this process is to transform the fictive persona into a tangible, relatable character. To bring the persona to life, it is visualized in an aesthetically appealing 1-page layout, the so-called *persona sheet*. This visualization shows the persona's name, picture, and a story conveying the persona's interests, values, lifestyle, attitudes, and behavioral patterns.

Although most of these elements are based on the information from the persona's foundation document, some fictional information may be included (e.g., regarding family or hobbies), to bring depth to the character. These fictional elements must be chosen carefully and deliberately, with the aim of supporting the communication of the persona's characteristics, whilst taking care to avoid the reinforcement of stereotypes. All important aspects of the persona described in their foundation document should be represented, in particular, regarding the persona's attitude towards AI and new technologies. Finally, to validate the visualization of a persona, it is recommended to obtain feedback from domain experts, or to show the persona sheet to

people from the respective user group and ask whether they feel plausibly and fairly represented (Marsden and Proebster 2019).

11.4.2 User-Testing for AI

As described in the previous sections, stakeholder analysis and personas are helpful methods for becoming familiar with the (potential) users and use context of AI applications, such that designers and developers may better empathize with their needs throughout the development process. However, to realize a product that is usable in practice, focusing on fictitious users is not sufficient. It is also necessary to involve real users.

User tests, e.g., Thinking Aloud methods, are proven tools and well-known from HCI and development of conventional software products (Alhadreti and Mayhew 2018; Nielsen 1993). Usually, these methods test the 'usability' of a product (Behringer et al. 2007), where this is defined by the International Organization for Standardization (ISO) as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (International Organization for Standardization (ISO) 2022). ISO defines 'effectiveness' as "*the accuracy and completeness with which users achieve specified goals*", 'efficiency' as "*resources used in relation to the results achieved"* and 'satisfaction' as "*freedom from discomfort and positive attitudes towards the use of the product*" (International Organization for Standardization (ISO) 2022).

However, for trustworthy AI solutions not only usability is important, but *causability* is equally crucial. Causability is defined by Holzinger et al. as *"the extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use"* (Holzinger et al. 2019). Thus, user tests of AI solutions should focus on not only how effectively, efficiently, and comfortably a user can achieve a specific goal using the AI solution, but also on these same criteria applied to explanations provided by the AI solution – and additionally, how *satisfied* they are with these explanations.

Aside from qualitative surveys and questionnaires (Zhou et al. 2021), causability may be quantified using the System Causability Scale (SCS) (Holzinger et al. 2020). The SCS helps determine to what extent the explanation (including process and presentation) of a result delivered by an AI solution fits the intended purpose and needs of the recipient user (Holzinger et al. 2020). When measuring causability with the SCS, the user is asked to score on a five-point scale ranging from *1=strongly agree* to *5=strongly disagree*, in response to the following ten Likert statements:

1. I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. I understood the explanations within the context of my work.
3. I could change the level of detail on demand.

4. I did not need support to understand the explanations.
5. I found the explanations helped me to understand causality.
6. I was able to use the explanations with my knowledge base.
7. I did not find inconsistencies between explanations.
8. I think that most people would learn to understand the explanations very quickly.
9. I did not need more references as medical guidelines and regulations.
10. I received the explanations in a timely and efficient manner.

As with the System Usability Scale (SUS) (Lewis 2018), the final SCS Likert score is calculated as the sum of all ratings of the ten statements divided by 50 (Holzinger et al. 2020). In addition to self-reported SCS results, human mental involvement, level of understanding, emotional arousal, and stress in response to human-AI interfaces may also be measured. This analysis can be performed on eye-tracking data (Pivec et al. 2004; Preis and Müller 2003) and additional sensors such as facial expression analysis, and electrodermal activity.

11.5 An Example Use Case: Computational Pathology

Medicine is an application field in which AI solutions may bring about great benefits for individual patients, as well as public health. However, it is also a domain where stakes are high and AI solutions may introduce a high risk of harm. Therefore, as specifically mentioned in the Artificial Intelligence Act (European Commission 2021), the health sector is one of the application fields for which the trustworthiness of the implemented AI solutions is of utmost importance. Thus, we have chosen computational pathology, a subdomain of the medical imaging field, as a case study of how stakeholder analysis and the personas method may be applied to develop human-AI interfaces supporting trustworthiness.

11.5.1 AI in Computational Pathology

In histopathology, human tissue samples are investigated for cellular and/or molecular indications of diseases. In preparation, formalin-fixed-paraffin-embedded (FFPE) tissue samples are cut into ultra-thin slices, mounted on glass-slides, and pre-processed to make cellular structures and bio-markers visible under microscopy. Traditionally, these glass slides are examined by pathologists under a light microscope (Golob-Schwarzl et al. 2019; Kargl et al. 2020). In digital pathology, scanned representations of these glass slides, so-called Whole Slide Images (WSI), are examined by pathologists on a monitor (Jahn et al. 2020).

Computational pathology adds computational steps to support pathologists in their analysis of WSIs (Holzinger et al. 2017). AI for histopathological image analysis is a dynamic and growing research field (Srinidhi et al. 2021; Wulczyn et al. 2021, 2020;

Yi et al. 2019), and various AI solutions are in development to support pathologists with challenging tasks including detection of micrometastasis deposits in lymph nodes, detection and grading of prostate cancer, and immunohistochemistry-based prognostics for breast cancer (Regitnig et al. 2020).

Expectations of AI solutions in computational pathology include time savings, increasing accuracy and quality, and extraction of new medical knowledge. However, since the results of AI solutions in this area can have a tremendous impact on therapy decisions for patients, human oversight is indispensable. This renders explainability and accountability major challenges to overcome for the safe application of AI to histopathology (Holzinger et al. 2020).

11.5.2 Stakeholder Analysis for Computational Pathology

Stakeholder analysis for identifying the potential user groups of AI solutions in computational pathology is the first step in user personas development. This analysis is grounded on the question: "Who will need to understand the rationale behind the results provided by an AI solution for computational pathology and thus will need explanations for the results provided by this AI solution"?

These stakeholder groups can be identified for solutions that are in the development phase, solutions dedicated for research use only, and/or solutions approved for use in clinical work. As depicted in Fig. 11.2, these stakeholder groups include staff of the AI solution provider (software developer, quality manager, sales, customer support), staff of organizations assessing market conformity of medical software solutions (for example auditors at notified bodies designated under the EU In-Vitro Devices Regulation (IVDR) The European Parliament, The Council of the European Union 2017), staff of the pathology laboratory (pathologists, AI laboratory technician, quality manager), and researchers in medicine or molecular biology as well as researchers in data science or AI.

The need of these stakeholder groups to understand the result of an AI solution for computational pathology is based on different underlying objectives, such as debugging or improving an AI system, ensuring compliance with standards and regulations, understanding how to incorporate the AI results into further actions, and justifying or explaining actions influenced by the AI results (Suresh et al. 2021). There is therefore no one-size-fits-all solution with regards to explaining the results of AI applications in computational pathology. For example, while software developers' explanatory requirements will probably include technical details of the inner workings of the model, sales and customer support staff will usually require less technical details of the underlying algorithms but will need to understand the limits of use of the AI application and the expected accuracy of the results.

Differing levels of expected computer literacy and medical domain knowledge between stakeholder groups are illustrated in Fig. 11.3. These are important aspects to be taken into account when designing a human-AI interface or developing an explanation component for an AI solution in computational pathology.

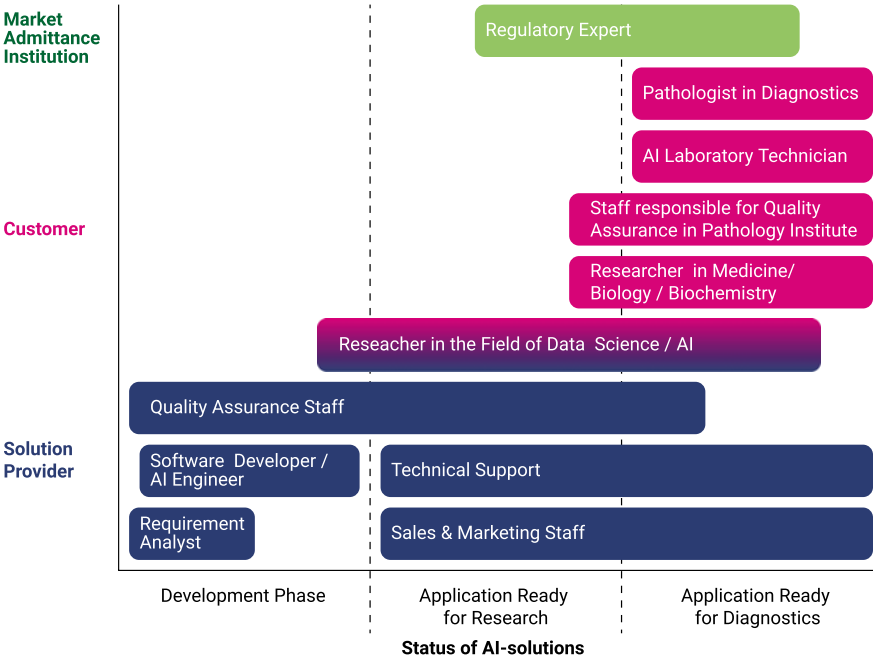


Fig. 11.2 Relevant stakeholders in different states of an AI solution for computational pathology, and their level of expertise in medicine/molecular biology

11.5.2.1 Relevant Stakeholder Groups for Computational Pathology

Groups of stakeholders who need to understand and interpret the results of AI solutions in computational pathology include staff of the solution provider, pathology institutes, market admittance institutions for medical devices, and scientists in the field of medicine and molecular biology, as well as in the field of data science and AI. All these stakeholder groups are briefly introduced in the following paragraphs.

Staff of AI Solution Providers
Roles with varying degrees of technical expertise and explainability requirements can be identified among the staff of an AI solution provider. These include requirement analysts, AI engineers and software developers, staff working in sales and marketing and technical customer support departments, as well as quality assurance managers and persons responsible for regulatory compliance.

Requirement Analysts identify the needs and demands of (potential) customers and define requirements for the to-be-developed software based on the intersection between these and the company's policies (Regitnig et al. 2020). Expertise in economics and a good understanding of the medical field in question is necessary to accomplish these tasks.

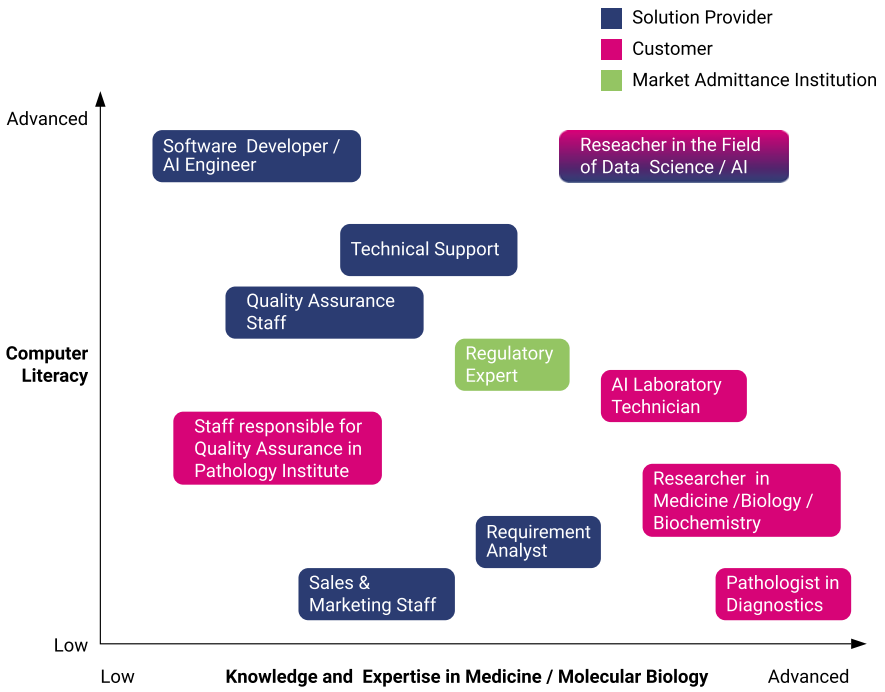


Fig. 11.3 Schematic overview of the expertise of stakeholders in computational pathology

AI Engineers and Software Developers design and implement software solutions based on the requirements collected by the requirement analyst. Typically, an education in informatics or software engineering is required, with expertise in Information Technology (IT) and computer science. They should be well aware of DICOM standards in digital pathology (Herrmann et al. 2018) and biobanking standards such as the ISO 20387 and the MIABIS ontology (Eklund et al. 2020). These stakeholders do not necessarily have extensive knowledge of medicine or molecular biology.

Sales and Marketing Staff bring the AI solution to customers. Usually, this personnel does not have extensive IT knowledge and only limited medical knowledge related to the software's application domain. However, they have a marketing and sales background and can convincingly present the solution to a (potential) customer.

Technical Support Staff are in direct contact with users and solve problems that arise during usage of the AI solution. Technical support staff often have extensive IT and/or computer science expertise, albeit without any corresponding requirement for medical domain knowledge.

Quality Assurance Manager and Person Responsible for Regulatory Compliance must have insight into the development processes and high awareness of quality standards (O'Sullivan 2019). The Person Responsible for Regulatory Compliance (PRRC), who establishes, documents, implements, and maintains a quality management system ensuring compliance with the EU In-vitro Diagnostics Regulation (IVDR), must have a degree in a relevant scientific discipline (law, medicine, pharmacy, or engineering) or four years of experience in regulatory affairs or quality management systems relating to medical devices (The European Parliament, The Council of the European Union 2017).

Staff of Pathology Institutes
Stakeholder roles with differing technical skills and needs for explainability can be identified amongst the staff of a pathology institute. The most obvious stakeholders are pathologists. However, technicians and quality managers at a pathology institute are also among those who must understand the results of AI solutions for computational pathology.

Pathologists are medical doctors who examine human tissues, cells, and body fluids in order to diagnose and monitor diseases, or predict, indicate, and monitor the outcome of therapies. Besides the findings from the microscopic examination of the specimen, a pathologist takes into account the case history and the results of other laboratory tests for these purposes. They have completed a comprehensive general medical education and (to differing degrees) highly specialized training in histopathology, and have got a strong understanding of laboratory medicine (including management, safety, and quality issues for the laboratory), excellent skills in interpreting complex patterns of test results, and knowledge regarding further tests needed for correct diagnoses.

AI Laboratory Technicians prepare AI results for pathologists. This task requires intermediate knowledge in both the IT and the medical domains. The AI laboratory technician digitizes histopathological glass slides to generate Whole Slide Images (WSIs), potentially applying pre-configured AI solutions to the resulting data. The tasks of the AI laboratory technician include the initial evaluation of AI results and adjustment of the system parameters where necessary. Furthermore, the AI laboratory technician transfers the AI results from a technical language into a format that is better suited for further analysis by pathologists.

Quality Manager at a Pathology Institute ensures the establishment, implementation, and maintenance of a quality management system compliant with standards and norms, as well as legal and regulatory requirements. The tasks of the quality manager comprise the development and monitoring of key quality indicators, key performance indicators, audit schedules, and contingency plans. Furthermore, the quality manager is responsible for risk assessment and mitigation. Typically, the quality manager at a pathology institute has a medical background and experience in quality management and economics.

Staff of Market Admittance Institutions
Every AI solution to be applied in medical diagnostics must be approved by the respective market admittance institution; i.e., the Food and Drug Administration (FDA) in the USA, or a notified body according to the Medical Device Regulation (MDR) or In-vitro Diagnostics Regulation (IVDR) in the European Union.

Auditors at a Market Admittance Institution are responsible for assessing the regulatory compliance of AI applications in medical diagnostics. Auditors at a market admittance institution usually work in interdisciplinary teams including computer experts and medical experts, such that the range of expertise necessary for the assessment of an AI solution for computational pathology is covered. While these rely predominantly on documents provided by the AI vendor to assess regulatory compliance of the product, they will also use the explanatory component of the AI solution in the evaluation of scientific validity, analytical performance, and clinical performance.

Scientific Staff
There are two groups of researchers, who may need to understand the results of computational pathology AI solutions. On the one hand, researchers in medicine or molecular biology use AI solutions to analyze human, plant, or animal specimens for their scientific work. On the other hand, there are researchers in data science or AI, who either are involved in the design and implementation of computational pathology AI solutions, or utilize (as customers) such solutions for their scientific work.

Researchers in Medicine or Molecular Biology are usually employed at a (medical) university, at the research department of a company (for example, in the pharmaceutical industry), or at a public institution (for example, the World Health organization). Typically, these researchers are not trained IT experts, but rather possess a university degree in medicine, biology, biochemistry, or pharmacy. They need AI solutions that are easily available and affordable, whilst remaining adaptable and extendable to their specific research question.

Researchers in Data Science or AI are involved in the development of AI solutions for computational pathology, trained as they are for translating real-world problems into machine learning approaches. They are skilled in designing, configuring, and adjusting complete AI solutions for discovering patterns in data. Usually, researchers in data science or AI have got an education in mathematics, computer science, or software engineering.

Aside from, or in addition to, being involved in AI development, they may use AI solutions as tools for their scientific work, and therefore may also be found on the consumer side of the AI solution life cycle.

11.5.2.2 Personas for Computational Pathology

As described in Sect. 11.4, user personas are a proven method to help understand (future) users, user interaction, and the context of use of a product, therefore supporting designers and developers in focusing on the needs and goals of their products' users.

We have developed user personas for all stakeholders of AI in computational pathology - a detailed description of this process can be found in (Holzinger et al. 2022). In the following paragraphs, the personas developed for the user group of pathologists are provided as an example.

In step 1 of the 5-step process for persona development: *Identification of (Potential) User Groups*, pathologists were identified as one of the user groups who must understand and interpret results delivered by AI solutions in computational pathology.

In step 2: *Collection of Information about the Users*, personal and work-related information was collected via contextual interviews with pathologists, an online survey among pathologists, and internet research. The collected work-related information included tasks, workflows, work context, education, experience, skills, and knowledge. Personal information comprised of goals, motivations, frustrations, personal traits, values, learning style, as well as attitudes about new technologies and AI.

In step 3: *Analysis of the Collected Information*, we recognized clusters in the answers of the pathologists with regards to two aspects that may influence pathologists' usage of AI solutions in computational pathology:

(a) some pathologists work only in diagnostics, while others (particularly in research hospitals connected to a medical university) work also in research, and
(b) some pathologists like to work with new technologies and are open-minded towards the usage of AI, while others are no technophiles and not so fond of working with new technologies.

Both of these aspects are important for the design and development of AI solutions for computational pathology: The finding (a) that some pathologists work only in diagnostics while others work also in research is relevant with respect to how an AI solution would be used by these two groups, since research work is different from routine diagnostics:

First, research work is usually multi-disciplinary and typically includes experts from various research fields such as medicine, pharmacy, biochemistry, biology, and bioinformatics. Second, research work introduces new methods and approaches and can be more exploratory and experimental than routine diagnostics, as it aims to go beyond the state-of-the-art and generate new scientific findings. Third, research work is usually conducted under less strict time constraints than routine diagnostics, therefore allowing more resources to be expended on a single medical case.

The finding (b) that some pathologists are technophiles and some are not is relevant with respect to their requirements for AI solutions: Users who are technophiles and well versed in exploring and trying out new technologies, are likely to be open to

applications that offer more options for customization, while users who are not so fond of spending time getting familiar with new technologies may require more guidance and rely on applications that offer easy-to-use default procedures.

For step 4, we created foundation documents for 3 different 'pathologist' personas based on these findings, which are shown in Table 11.1.

Finally, in the last step of the 5-step process for persona development, each of these 3 'pathologist' personas was visualized in a persona-sheet, rendering these fictional characters in a tangible and realistic way. These persona-sheets can be found in the appendix (Sect. 11.7).

11.5.3 Human-AI Interface in Computational Pathology

As so far demonstrated, it is essential to consider the users and the usage context when designing and developing human-AI interfaces. Thus, in defining the needed functionalities of a user interface (UI) for computational pathology AI solutions that aim to support pathologists with WSI analysis, it is important to take into account the workflows in pathology (Kargl et al. 2020) and the process of how pathologists develop a diagnosis (Pohn et al. 2019a, b).

With regards to functionality, the UI in computational pathology can be split into two components: (1) The basic digital pathology UI, enabling the user to manage and view WSIs, and (2) the AI-specific component, which is a human-AI interface that enables the user to interact with a specific computational pathology AI solution.

This AI-specific component can be further split into two components, based on functionality: (2a) the primary AI interface, which enables the user to request, view, and interact with the primary result of the AI solution, and (2b) the explanation component, which enables the user to request and view an explanation as required for the primary result of the AI solution, as depicted in Fig. 11.4. Ideally, these components should be integrated, with the basic UI providing the possibility to dynamically add (plugin) user interaction functionalities for specific application contexts.

11.5.3.1 Basic UI Components: Case and Slide Viewer

The *Case Viewer* and the *Slide Viewer* are at the core of the UI in computational pathology. They can be implemented as combined software or as stand-alone products.

The Case Viewer enables the user to access and view a set of WSIs, as shown in Fig. 11.5. In diagnostics, where a single medical case can comprise up to 100 WSIs, the task of the case viewer is to present all these WSIs of a medical case to the pathologist in a structured way. In research, the case viewer helps to manage and organize WSIs for projects. Frequently, the case viewer is a pathologist's entry point to digital pathology, as it forms the link between the slide viewer and the Laboratory Information System (LIMS) in a pathology institute.

Table 11.1 Foundation documents for 'pathologist' personas

	Persona 1 Pathologist in diagnostics type 1	Persona 2 Pathologist in diagnostics type 2	Persona 3 Pathologist in research
Tasks	Diagnostics macroscopic examination of specimens microscopic examination of specimens examination of frozen sections participate in tumor-boards autopsy administration	Diagnostics macroscopic examination of specimens microscopic examination of specimens examination of frozen sections participate in tumor-boards autopsy administration	Examination of WSIs for research scientific publications participate in scientific projects applying for research funding administration
Personal traits	Likes working accurately and precisely obsessed with details conscientious curios, spirit of research not very spontaneous not technophile rather reluctant, resistant to changes	Likes working accurately and precisely obsessed with details conscientious curios, spirit of research not very spontaneous likes working with new technologies open-minded to changes	Likes working accurately and precisely obsessed with detail curios, spirit of research likes working with new technologies open-minded to changes
Motivations	Enjoys pathology work finding the right diagnosis for patients solving challenging and difficult cases recognition of my work among peers	Enjoys pathology work finding the right diagnosis for patients solving challenging and difficult cases recognition of my work among peers bring new scientific findings to practice	Enjoys pathology work new scientific insights recognition of my work in scientific community
Frustrations	Feuding among colleagues unnecessary delays and waiting times administrative stuff inaccuracy and imprecision insufficient clinical information too much work disturbances at work (phone calls, people dropping in, training sessions, slow IT services, computer error...) lack of recognition of pathologists' work by the general public and by clinicians	Feuding among colleagues unnecessary delays and waiting times administrative stuff inaccuracy and imprecision insufficient clinical information too much work to be pressed for time no time for reading specialist literature	Feuding among colleagues unnecessary delays and waiting times administrative hurdles too much work lack of finance
Ideal working environment	Just let me work—no disturbances minimise administrative work amount of work just right flexible working hours, home office calm surroundings (no other people in the room) reliable infrastructure ergonomic workplace	Minimise administrative work being innovative support of employer for my visions good colleagues relaxed atmosphere at the workspace amount of work just right up-to-date infrastructure and tools flexible working hours, home office ergonomic workplace	Just let me work—no disturbances minimise administrative work being innovative support of employer for my visions good colleagues relaxed atmosphere at the workspace amount of work just right up-to-date infrastructure and tools flexible working hours, home office

The case viewer provides functionality related to the organization of cases and WSIs (e.g., grouping of WSIs in cases and blocks, reordering of WSIs, prioritization of cases, and user access management to cases), and functionality related to the enrichment of WSIs with metadata (e.g., gathering case-relevant metadata from LIMS, and triggering AI solutions to get specific WSI analysis results). Furthermore,

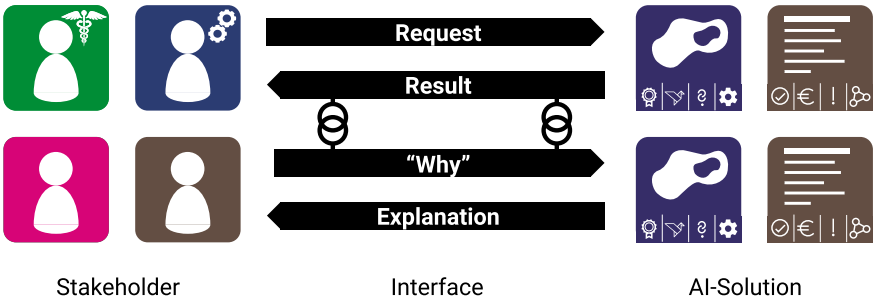


Fig. 11.4 The human-AI interface in computational pathology

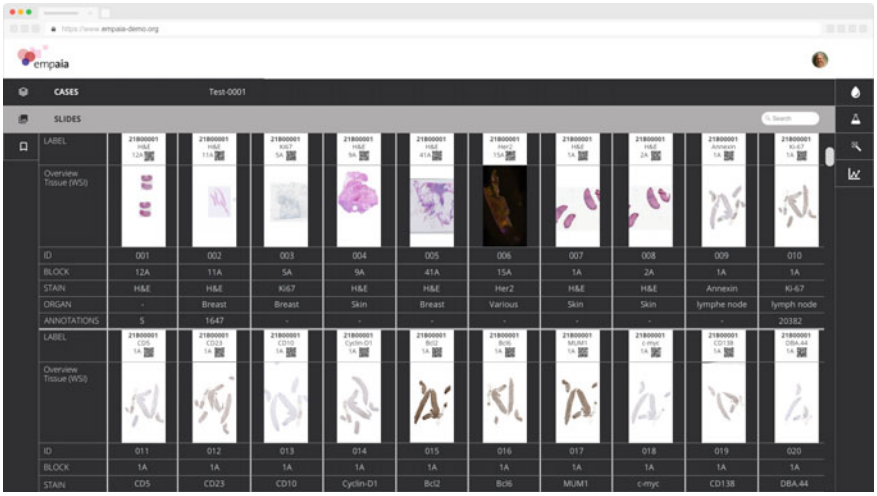


Fig. 11.5 Example of a Case Viewer, which enables the user to access, view, organize, and manage a set of whole slide images (WSIs)

the case viewer offers plugin integration, which can be used to add for, example, WSI pre-processing functionality to the case viewer.

The Slide Viewer enables the user to view a WSI and related annotations on the screen, and provides appropriate interaction techniques for such gigapixel images, as shown in Fig. 11.6. Functionalities provided by the slide viewer include: visualizing the WSI at different magnifications, creating user annotations (e.g., text annotations or marking a region of interest (ROI) with a circle, rectangle or polygon), visualizing annotations created by the user or by an AI solution, scrolling through different focus-layers of a WSI (so-called z-stacking), automated alignment of WSIs, performing color-correction and channel-adjustment for WSI, as well as tracking the user's viewing activities for a WSI and marking already-viewed areas. Furthermore, the slide viewer offers plugin integration, which can be used to add third-party applications and extend the slide viewer's functionality, for example, with a quantification tool.

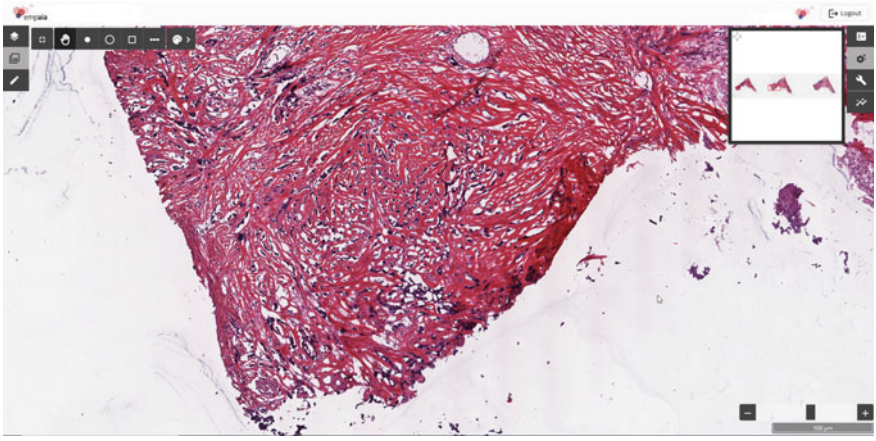


Fig. 11.6 Example of a Slide Viewer, which enables the user to view a whole slide image (WSIs) on the screen, navigate through the WSI and make annotations

11.5.3.2 AI-Specific UI Components

The AI-specific UI component comprises two functional parts: the primary AI interface providing the interaction and visualization functionalities, supporting the primary/main purpose of the AI solution, and the explanation component supporting the user's 'Why' questions (Miller 2019). Figure 11.7 shows an example of such an AI-specific UI in computational pathology.

The Primary AI Interface must, on the one hand, support communication of a specific user request to the AI solution, and, on the other hand, it must support communication of the AI solution's result to the user. To enable the user to formulate a specific request for the AI solution, the interface shall, for example, provide annotation functionality enabling the user to easily define ROIs and select specific parts of the WSI as input parameters for the AI solution.

Since pathologists are accustomed to looking at images and trained in interpreting image information, the results provided by an AI solution in computational pathology should, whenever possible, be communicated to the pathologists visually, as an overlay on the WSI, for example as pixel-based or ROI-based annotations. Ideally, different AI solutions used in a domain such as computational pathology, should apply a consistent visual vocabulary of symbols and color codes when visualizing results and/or explanations.

The Explanation Component should convey explanatory information to the user. A large variety of explanation approaches for AI exists, and a comprehensive overview of those state-of-the-art XAI methods, which are applicable to computational pathology, is given by Pocevičiūtė (Pocevičiūtė et al. 2020). Possible modalities include visual overlays on the WSI (such as shapes, heatmaps, or saliency maps), figures and

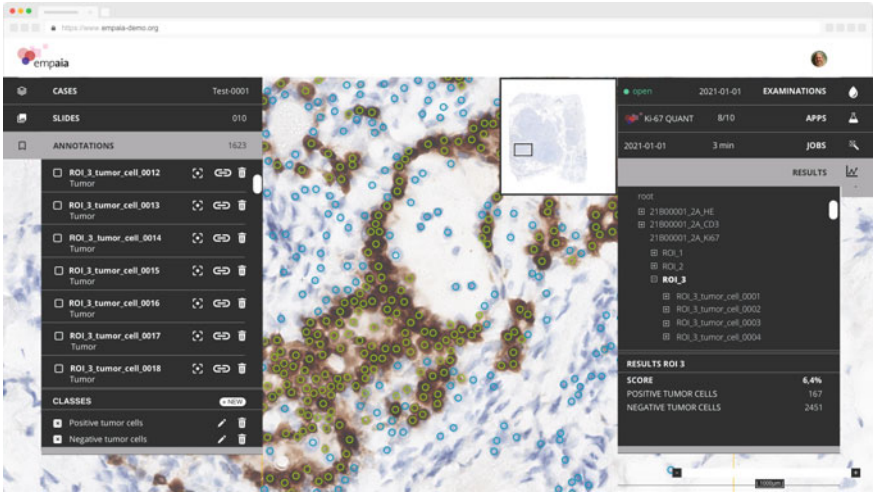


Fig. 11.7 Example of a UI displaying results and explanatory annotations generated by an AI solution. Here, the result is the overall positivity score (bottom right), whereas an intermediate result, consisting of individual cell annotations, is included as an explanatory element that helps the user to understand this outcome

charts, text labels, example images (e.g., counterfactuals or prototypes), interactive dialogue systems, or even simply intermediate results of the AI solution.

When choosing an explanation method, it must be considered that a good explanation should support the context, task, and goal of the recipient of the explanation. Thus, to determine which kind of explanation is appropriate and how the explanatory information should be provided, it is necessary to analyze the specific task and context in which a specific user requires an explanation. As many different stakeholder groups have been identified to be relevant for AI solutions in computational pathology, it should be considered to implement explanation components optimized for different user groups. When the same explanation method is used for different stakeholder groups, it is necessary to thoroughly adapt it to the knowledge and needs of the respective members. For example, since a pathologist's common task of examining WSIs to derive a diagnosis is a visual task, pathologists prefer also visual modalities for explanations of AI solution's results (Evans et al. 2022). In addition, also the simplicity of the explanation method is an important criterion for pathologists, as they usually face a high workload and a tight schedule and thus consider time savings as the most valuable benefit of AI assistance (Evans et al. 2022).

It is important that the explanation considers the user's background knowledge and uses concepts, which are familiar to the user. Moreover, an explanation must also match the user's need for information. For example, if a pathologist needs to verify the correctness of a cell ratio calculated by an AI solution, an explanation highlighting the cells recognized by the AI solution and the respective ROI may be easily understandable, since it uses concepts from the medical domain (cells) and from common knowledge (how to calculate a ratio) that are familiar to the pathologist.

However, an explanation approach displaying a saliency map of the most relevant parts of the image for a given AI output, may not help the pathologist to assess the correctness of the ratio calculated by the AI solution, and may therefore be inappropriate in this context. Such a saliency map could, however, be a helpful explanation for a software developer searching for clues to a 'Clever-Hans' problem (Pfungst 1911) in the AI model.

The explanation component of the human-AI interface is a crucial element for the trustworthiness of AI solutions in computational pathology. However, as shown by Evans et al. (2022), ambiguous explanations may also pose a significant risk of introducing inappropriate trust in AI solutions. This risk is particularly pronounced when the explanation seems to imply a causal decision-making process of the AI model similar to the user's own. For example, saliency maps pointing to diagnostically relevant regions, whilst obscuring or omitting the important features represented within these regions, or synthetically created counterfactuals with several features changing simultaneously leaving it unclear which of these were truly relevant (Evans et al. 2022).

Therefore, especially in high-risk domains such as medicine, it is important that human-AI interfaces are carefully designed and thoroughly tested with users from the target group before they are applied in practice.

11.6 Conclusion

In summary, both the design and the evaluation of the human-AI interface play a central and increasingly important role in achieving and verifying the trustworthiness of an AI solution. This is particularly relevant in application domains that directly impact human life and livelihood. In reviewing ethical criteria for AI applications in biomedical research and biobanking (Kargl et al. 2022), there is a clear need for further basic research that will then lead to concrete practical guidelines.

In the context of legislation and regulatory practice, e.g., the European In-Vitro Diagnostics Device Regulation (IVDR) / Medical Device Regulation (MDR) and U.S. Food and Drug Administration (FDA) activities, aspects of explainability and causability of the human-AI interface will be an indispensable component for the validation of critical AI solutions (Müller et al. 2022). In order to meet the stringent requirements of both extant and upcoming regulatory frameworks, further fundamental research is imperative.

Robust and explainable human-AI interfaces will be the central component for building truth and trustworthiness in AI systems, and therefore constitute an important and exciting area of future research. In particular, further work is required to make the explainability and causability of Human-AI interfaces in critical domains measurable and quantifiable. To this effect, it is necessary to develop strategies for the collection of ground truth relating to these metrics. Future dialog systems and benchmark datasets such as the Kandinsky Patterns (Müller and Holzinger 2021) can provide valuable assistance to the international research community toward this goal.

11.7 List of Abbreviations

AI	Artificial Intelligence
DICOM	Digital Imaging and Communications in Medicine (standard)
EC	European Commission
EU	European Union
FDA	United States Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
HAII	Human-AI Interaction
HCI	Human-Computer Interaction
ISO	International Organization for Standardization
IT	Information Technology
IVDR	European In-Vitro Devices Regulation
I/O	Input-Output
LIMS	Laboratory Information System
MDR	European Medical Devices Regulation
MIABIS	Minimum Information About Biobank data Sharing (standard)
PRRC	Person Responsible for Regulatory Compliance
ROI	Region of Interest
SCS	System Causability Scale
SUS	System Usability Scale
UI	User Interface
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence

Acknowledgements Parts of this work have received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No. 857122 (CY-Biobank), No. 824087 (EOSC-Life), and No. 874662 (HEAP). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. Parts of this work have received funding from the Austrian Research Promotion Agency (FFG) under grant agreement No. 879881 (EMPAIA) and by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence. Part of this work was done in the context of EMPAIA, a project funded by the German Federal Ministry for Economic Affairs and Climate Action, with funding codes 01MK20002A, 01MK20002C, and 01MK20002E. We are very grateful for the proofreading of the manuscript by Bettina Kipperer.

Appendix

Three sample personas for AI applications in digital pathology, as described in Sect. 11.5.2.2 (Figs. 11.8, 11.9 and 11.10).

Fig. 11.8 Visualization of the persona representing a pathologist in research

Fig. 11.9 Visualization of the persona representing a 'technophile' pathologist in diagnostics

Fig. 11.10 Visualization of the persona representing a pathologist in diagnostics, who is not so fond of new technologies and changes

References

Alhadreti, O., Mayhew, P.: Rethinking thinking aloud: A comparison of three think-aloud protocols. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, p. 1-12. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173618

Behringer, R., Christian, J., Holzinger, A., Wilkinson, S.: Some usability issues of augmented and mixed reality for e-health applications in the medical domain. In: HCI and Usability for Medicine and Health Care. Lecture Notes in Computer Science (LNCS 4799), pp. 255–266. Springer (2007). https://doi.org/10.1007/978-3-540-76805-0_21

Calero Valdez, A., Schaar, A., Ziefle, M., Holzinger, A., Jeschke, S., Brecher, C.: Using mixed node publication network graphs for analyzing success in interdisciplinary teams. In: R. Huang, A.A. Ghorbani, G. Pasi, T. Yamaguchi, N.Y. Yen, B. Jin (eds.) Active Media Technology, Lecture Notes in Computer Science LNCS 7669, pp. 606–617. Springer, Heidelberg, Berlin (2012). https://doi.org/10.1007/978-3-642-35236-2_61

Castro, J.W., Acuña, S.T.: Extension of personas technique for the requirements stage. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7545 LNCS**, 94–103 (2012). https://doi.org/10.1007/978-3-642-33760-4_8

Cooper, A., Reimann, R.: About Face 2.0 - The Essentials of Interaction Design. John Wiley & Sons (2003). https://flylib.com/books/en/2.153.1/

Cooper, A., Saffo, P.: The Inmates Are Running the Asylum. Macmillan Publishing Co., Inc, USA (1999)

Crisan, A., Correll, M.: User ex machina: Simulation as a design probe in human-in-the-loop text analytics. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2021). https://doi.org/10.1145/3411764.3445425

Dix, A., Finlay, J., Abowd, G.D., Beale, R.: Human-computer interaction. Prentice Hall, Harlow (1993)

Ebner, M., Holzinger, A., Maurer, H.: Web 2.0 technology: Future interfaces for technology enhanced learning? In: C. Stephanidis (ed.) Universal Access to Applications and Services, Lecture Notes in Computer Science (LNCS 4556), pp. 559–568. Springer (2007). https://doi.org/10.1007/978-3-540-73283-9_62

Eklund, N., Andrianarisoa, N.H., van Enckevort, E., Anton, G., Debucquoy, A., Müller, H., Zaharenko, L., Engels, C., Ebert, L., Neumann, M., et al.: Extending the minimum information about biobank data sharing terminology to describe samples, sample donors, and events. Biopreservation and biobanking **18**(3), 155–164 (2020)

European Commission: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (com(2021 206) (2021). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

European Commission, Directorate-General for Communications Networks, Content and Technology: Ethics guidelines for trustworthy AI. European Commission Publications Office (2019). http://orcid.org/doi/10.2759/177365

European Parliament, the Council and the Commission: Charter of fundamental rights of the european union (2012). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

Evans, T., Retzlaff, C.O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.R., Zerbe, N., Holzinger, A.: The explainability paradox: Challenges for xAI in digital pathology. Future Generation Computer Systems **133**, 281–296 (2022). https://doi.org/10.1016/j.future.2022.03.009

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: Ai4people-an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Minds and Machines **28**, 689–707 (2018)

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89 (2018). https://doi.org/10.1109/DSAA.2018.00018

Golob-Schwarzl, N., Bettermann, K., Mehta, A.K., Kessler, S.M., Unterluggauer, J., Krassnig, S., Kojima, K., Chen, X., Hoshida, Y., Bardeesy, N.M., Müller, H., Svendova, V., Schimek, M.G., Diwoky, C., Lipfert, A., Mahajan, V., Stumptner, C., Thüringer, A., Fröhlich, L.F., Stojakovic, T., Nilsson, K., Kolbe, T., Rülicke, T., Magin, T.M., Strnad, P., Kiemer, A.K., Moriggl, R., Haybaeck, J.: High keratin 8/18 ratio predicts aggressive hepatocellular cancer phenotype. Translational Oncology **12**(2), 256–268 (2019). https://doi.org/10.1016/j.tranon.2018.10.010

Harper, E.R., Rodden, T., Rogers, Y., Sellen, A.: Being Human: Human-Computer Interaction in the year 2020. Microsoft Research, Cambridge (UK) (2008)

Herrmann, M.D., Clunie, D.A., Fedorov, A., Doyle, S.W., Pieper, S., Klepeis, V., Le, L.P., Mutter, G.L., Milstone, D.S., Schultz, T.J., Kikinis, R., Kotecha, G.K., Hwang, D.H., Andriole, K.P., lafrate, A.J., Brink, J.A., Boland, G.W., Dreyer, K.J., Michalski, M., Golden, J.A., Louis, D.N., Lennerz, J.K.: Implementing the DICOM standard for digital pathology. Journal of Pathology Informatics **9**(1), 37 (2018). https://doi.org/10.4103/jpi.jpi_42_18

Holzinger, A.: Rapid prototyping for a virtual medical campus interface. IEEE Software **21**(1), 92–99 (2004). https://doi.org/10.1109/MS.2004.1259241

Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016). https://doi.org/10.1007/s40708-016-0042-6

Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv:1712.09923 (2017)

Holzinger, A., Carrington, A., Müller, H.: Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z

Holzinger, A., Errath, M., Searle, G., Thurnher, B., Slany, W.: From extreme programming and usability engineering to extreme usability in software engineering education. In: 29th International Annual IEEE Computer Software and Applications Conference (IEEE COMPSAC 2005), pp. 169–172. IEEE (2005). https://doi.org/10.1109/COMPSAC.2005.80

Holzinger, A., Goebel, R., Mengel, M., Müller, H.: Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50402-1

Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., Muller, H.: Personas for Artificial Intelligence (AI) an Open Source Toolbox. IEEE Access **10**, 23732–23747 (2022). https://doi.org/10.1109/ACCESS.2022.3154776

Holzinger, A., Kieseberg, P., Weippl, E., Tjoa, A.M.: Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable ai. In: Springer Lecture Notes in Computer Science LNCS 11015, pp. 1–8. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_1

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(4), 1–13 (2019). https://doi.org/10.1002/widm.1312

Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R., Zatloukal, K.: Machine learning and knowledge extraction in digital pathology needs an integrative approach. In: Towards Integrative Machine Learning and Knowledge Extraction, Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344, pp. 13–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69775-8-2

International Organization for Standardization (ISO): ISO 9241-11:2018(en), Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (2018). https://www.iso.org/obp/ui/iso:std:iso:9241:-11:ed-2:v1:en

Jahn, S.W., Plass, M., Moinfar, F.: Digital pathology: Advantages, limitations and emerging perspectives. Journal of Clinical Medicine **9**, 3697 (2020). https://doi.org/10.3390/jcm9113697
Kargl, M., Plass, M., Müller, H.: A literature review on ethics for AI in biomedical research and biobanking. Yearbook of Medical Informatics p. to appear (2022)
Kargl, M., Regitnig, P., Müller, H., Holzinger, A.: Towards a better understanding of the workflows: Modeling pathology processes in view of future AI integration. In: Artificial Intelligence and Machine Learning for Digital Pathology. Lecture Notes in Computer Science, vol. 12090, pp. 102–117. Springer International Publishing, Cham (2020)
Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J.: Understanding, scoping and defining user experience: a survey approach. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 719–728 (2009). https://doi.org/10.1145/1518701.1518813
LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
Lewis, J.R.: The system usability scale: past, present, and future. International Journal of Human-Computer Interaction **34**(7), 577–590 (2018). https://doi.org/10.1080/10447318.2018.1455307
Lipton, Z.C.: The mythos of model interpretability. Queue **16**, 31–57 (2018). https://doi.org/10.1145/3236386.3241340
Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: Concepts, applications, research challenges and visions. In: Machine Learning and Knowledge Extraction, vol. 12279 LNCS, pp. 1–16. SpringerInternational Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
Louie, R., Coenen, A., Huang, C.Z., Terry, M., Cai, C.J.: Novice-AI music co-creation via AI-steering tools for deep generative models. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020). https://doi.org/10.1145/3313831.3376739
Marsden, N., Proebster, M.: Personas and identity: Looking at multiple identities to inform the construction of personas. In: Conference on Human Factors in Computing Systems CHI 2019 - Proceedings, pp. 1–14. Association for Computing Machinery (2019). https://doi.org/10.1145/3290605.3300565
Miaskiewicz, T., Kozar, K.A.: Personas and user-centered design: How can personas benefit product design processes? Design Studies **32**(5), 417–430 (2011)
Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., Zatloukal, K.: Explainability and causability for artificial intelligence-supported medical image analysis in the context of the european in vitro diagnostic regulation. New Biotechnology p. to appear (2022)
Müller, H., Holzinger, A.: Kandinsky patterns. Artificial Intelligence **300**, 103,546 (2021)
Müller, H., Kargl, M., Plass, M., Kipperer, B., Brcic, L., Regitnig, P., Geißler, C., Küster, T., Zerbe, N., Holzinger, A.: Towards a Taxonomy for Explainable AI in Computational Pathology, pp. 311–330. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-72188-6_15
Müller, H., Mayerhofer, M.T., Veen, E.B.V., Holzinger, A.: The ten commandments of ethical medical AI. IEEE COMPUTER **54**(7), 119–123 (2021). https://doi.org/10.1109/MC.2021.3074263
Nielsen, J.: Usability Engineering. Academic Press (1993)
Nielsen, L.: Design personas - new ways, new contexts. Persona Studies **4**, 1–4 (2018). https://doi.org/10.21153/psj2018vol4no2art799
Nielsen, L.: Personas - User Focused Design. Springer, London (2019). https://doi.org/10.1007/978-1-4471-7427-1
Norman, D.: The design of everyday things: Revised and expanded edition. Basic books (2013)
Ortigosa, A., Carro, R.M., Quiroga, J.I.: Predicting user personality by mining social interactions in facebook. Journal of computer and System Sciences **80**(1), 57–71 (2014). http://orcid.org/0.1016/j.jcss.2013.03.008

O'Sullivan, S., et al.: Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (ai) and autonomous robotic surgery. The International Journal of Medical Robotics and Computer Assisted Surgery **15**(1), e1968 (2019)

Pfungst, O.: Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology. Holt, Rinehart and Winston, London (1911)

Pivec, M., Preis, A., Garcia-Barrios, V., Gütl, C., Müller, H., Trummer, C., Mödritscher, F.: Adaptive knowledge transfer in e-learning settings on the basis of eye tracking and dynamic background library. In: Proceedings of EDEN 2004 Annual Conference, pp. 295–301 (2004)

Pocevičiūtė, M., Eilertsen, G., Lundström, C.: Survey of xai in digital pathology. In: A. Holzinger, R. Goebel, M. Mengel, H. Müller (eds.) Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges, pp. 56–88. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-50402-1_4

Pohn, B., Kargl, M., Reihs, R., Holzinger, A., Zatloukal, K., Müller, H.: Towards a deeper understanding of how a pathologist makes a diagnosis: Visualization of the diagnostic process in histopathology. In: 2019 IEEE Symposium on Computers and Communications (ISCC), pp. 1081–1086. IEEE (2019)

Pohn, B., Mayer, M.C., Reihs, R., Holzinger, A., Zatloukal, K., Müller, H.: Visualization of histopathological decision making using a roadbook metaphor. In: 2019 23rd International Conference Information Visualisation (IV), pp. 392–397. IEEE (2019)

Preis, A., Müller, H.: Eyetracking in usability research & consulting: What do the eyes reveal about websites & their users. In: European Congress of Psychology: Psychology in Dialogue with Related Disciplines, Austria, vol. 164 (2003)

Pruitt, J., Grudin, J.: Personas: practice and theory. DUX 03: Proceedings of the 2003 conference on Designing for user experiences pp. 1–15 (2003). https://doi.org/10.1145/997078.997089

Pruitt, J.S., Adlin, T.: The Persona Lifecycle - Keeping People in Mind Throughout Product Design. Elsevier Inc., USA (2006). https://doi.org/10.1016/B978-0-12-566251-2.X5000-X

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., Mazurek, G.: In bot we trust: A new methodology of chatbot performance measures. Business Horizons **62**(6), 785–797 (2019). https://doi.org/10.1016/j.bushor.2019.08.005

Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: AI in health and medicine. Nature Medicine **28**, 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0

Regitnig, P., Müller, H., Holzinger, A.: Expectations of artificial intelligence in pathology. In: Springer Lecture Notes in Artificial Intelligence LNAI 12090, pp. 1–15. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50402-1-1

Robert, S., Büttner, S., Röcker, C., Holzinger, A.: Reasoning under uncertainty: Towards collaborative interactive machine learning. In: A. Holzinger (ed.) Machine Learning for Health Informatics: State-of-the-Art and Future Challenges, pp. 357–376. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-50478-0_18

Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67** (2021). https://doi.org/10.1016/j.media.2020.101813

Suresh, H., Gomez, S.R., Nam, K.K., Satyanarayan, A.: Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445088

The European Parliament, The Council of the European Union: Regulation (EU) 2017/ 746 of the European Parliament and of the Council - of 5 April 2017 - on in vitro diagnostic medical devices. Official Journal of the European Union **L117**, 176–332 (2017). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0746&rid=6

Vredenburg, K., Mao, J.Y., Smith, P.W., Carey, T.: A survey of user-centered design practice. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 471–478 (2002). https://doi.org/10.1145/503376.503460

Wulczyn, E., Nagpal, K., Symonds, M., Moran, M., Plass, M., Reihs, R., Nader, F., Tan, F., Cai, Y., Brown, T., et al.: Predicting prostate cancer specific-mortality with artificial intelligence-based gleason grading. Communications Medicine **1**(1), 1–8 (2021)

Wulczyn, E., Steiner, D.F., Moran, M., Plass, M., Reihs, R., Müller, H., Sadhwani, A., Cai, Y., Flament, I., Chen, P.H.C., et al.: A deep learning system to predict disease-specific survival in stage ii and stage iii colorectal cancer (2020)

Xu, W., Dainoff, M.J., Ge, L., Gao, Z.: From human-computer interaction to human-ai interaction: New challenges and opportunities for enabling human-centered ai. arXiv:2105.05424 [cs.HC] (2021). http://arxiv.org/abs/2105.05424

Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis **58** (2019). https://doi.org/10.1016/j.media.2020.101552

Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021). https://doi.org/10.3390/electronics10050593

Ziefle, M., Klack, L., Wilkowska, W., Holzinger, A.: Acceptance of telemedical treatments - a medical professional point of view. In: S. Yamamoto (ed.) Human Interface and the Management of Information. Lecture Notes in Computer Science LNCS 8017, pp. 325–334. Springer (2013). https://doi.org/10.1007/978-3-642-39215-3-39