Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Explainable AI: Bridging the Gap between Machine Learning Models and Human Understanding

Rajiv Avacharmal
AI/ML Risk Lead, AI/ML expert
University of Connecticut, USA
rajiv.avacharmal@gmail.com

Abstract: Explainable AI (XAI) is one of the key game-changing features in machine learning models, which contribute to making them more transparent, regulated and usable in different applications. In (the) investigation of this paper, we consider the four rows of explanation methods—LIME, SHAP, Anchor, and Decision Tree-based Explanation—in disentangling the decision-making process of the black box models within different fields. In our experiments, we use datasets that cover different domains, for example, health, finance and image classification, and compare the accuracy, fidelity, coverage, precision and human satisfaction of each explanation method. Our work shows the rule trees approach called (Decision Tree-based explanation) is mostly superior in comparison to other non-model-specific methods of explanation performing higher accuracy, fidelity, coverage and precision regardless the classifier. In addition to this, the respondents who answered the qualitative evaluation indicated that they were very content with the decision tree-based explanations and that these types of explanations are very easy and understandable. Furthermore, most of the respondents famous that these sorts of clarifications are more instinctive and significant. The over discoveries stretch on the utilize of interpretable AI strategies for facilitating the hole between machine learning models and human understanding and thus advancing straightforwardness and responsibility in AI-driven decision-making.

Keywords: Explainable AI, Interpretability, Choice Tree, Machine Learning, Straightforwardness.

I. INTRODUCTION

Within the course of the advancement of artificial intelligence (AI) calculation and machine learning calculations have gone more distant and more distant, utilizing them for diverse circles of trade and industry for case healthcare, back and driverless cars, with extraordinary precision and proficiency. In any case, as these models become more complicated and more common in society, there's a growing requirement for straightforwardness and interpretability to create beyond any doubt that they are utilized in a dependable way which they are acknowledged by the end-users. The investigated subject of "Logical AI: "Bridging the Hole Between Machine Learning Models and Human Understanding," highlights the vital issue of creating procedures and techniques to empower people to decipher AI frameworks and comprehend them. The central point of this research is the fact that the "black-box" nature of many machine learning models is the main obstacle that hinders the understanding of how these models arrive at their decisions or predictions. On the other hand, even if models prove accurate Oftentimes this is the case, but still people have doubts how trustful, accountable and bias-free these models are [1]. Secondly, in the cases that are related to healthcare or criminal justice, the virtue of such AI-driven decisions to be explainable is definitely very vital to ensure fairness, equity and those ethical standards. The researchers are trying to find the link between computer learning models and human understanding through XAI (explainable AI). This refers to the planning of strategies for the AI programmers to create a system that can produce details for the humans to understand the reasons that led their decisions [2]. Techniques like feature importances analysis, model visualization and post-hoc explanation the algorithms all together can present a way out for opening the black box and showing the decision-making process. In addition to the fact that explainability is just a comprehension tool, it is also a collaborative tool that brings human experts and AI systems together. As AI evolves in domains where subject-specific knowledge is integrally, transparent AI can integrate with human knowledge to improve model accuracy and create useful outputs, such as actionable insights [3]. Consequently, explainable AI goes further and wards off opacity and untrustworthiness, along with promoting human-AI collaboration thus unleashing a greater extent of smart capabilities that are relevant in the ordinary life. In this research project, it will analyze the different ways of explaining AI, compare their effectiveness in different areas and eventually, contribute to the creation of the AI systems that are transparent and reliable. As an interdisciplinary team working with the empirical research, we are pursuing the ambitious goal to enhance the ethical and practical standards in AI to ensure safe and beneficial potential for the whole society.

II. RELATED WORKS

A few recent researches are dealing with the explainable AI components within overall area of XAI and thus opening the curtain of the central concept, crucial challenges, and implementation fields of XAI in multiple areas. #XAI #AI #Interpretability. González-Rodriguez et al. (2024) [15] studied the prospects of AI, pinpointing its likely role in phytopathology dissiease diagnosis and control in horticulture. They said that the interpretablity of the AI models was essential for the decision-making process in agriculture. According to the article by [16] Graziani et al. (2023), a common terminology for interpretable AI for social and technical sciences has been introduced as a global taxonomy of interpretable

AI. They offer a framework that provides a systematic guideline which helps in sorting out and classifying various methods and concepts arising in conjunction with XAI. Gugliermetti et al. (2024) stressed on the future of machine learning for building energy management and thus, the need for interpretable models to increase transparency and confidence in energy optimization systems [17]. This study explored the problem of the link between mechanistic (or hybrid/natural language-based) biological models and machine learning substitutes, showing the necessity of using domain knowledge in AI models to improve their analysis and applicability [18]. The paper "Explainability in AI healthcare" by Hulsen (2023) targeted the concepts and challenges of explainable AI in healthcare, stressing its crucial importance for the presented decision-making processes to be explainable, transparent, accountable, and so well trusted [19]. Jean-Quartier et al. (2023) studied the computational cost of XAI algorithms and proposed sustainable machine learning methods that combine the interpretability with the computational efficiency, thus, achieving the balance [20]. In their recent work, Joshi et al. (2024) create a landscape of devices produced by the FDA for AI/ML application in medicine, which serves as proof that AI is gaining traction in healthcare, and that regulation of AI systems that are transparent and interpretable is extremely important [21]. Khot et al. (2023) investigated in detail the interpretability of a top tagging technique that uses deep neural networks, and the result gave us a clue as to why there are difficulties and opportunities in knowing about complex AI models in particle physics applications [22]. According to the Kolajo and Daramola paper (2023), scholars examined both human-centric and semantics based explainable event detection models. The paper evaluated these techniques by presenting ways on how an AI system semantic linguistic and user centered design principles could be improved so that the interpretability of event driven AI system could be enhanced [23]. Lukomski et al. (in the year 2024) studied the effect of oncological data on the AI predictions of five years survival in esophageal cancer, which is a proof of the necessity of interpretable AI models for clinical decision support and personalized medicine [24]. While Mallinger and Baeza-Yates (2024) put forward a multicriteria framework of AI delivered in the context of farming so as to design sustainable technology solutions that entail principles of transparency, fairness and ethical concerns in agricultural practices [25]. Manfren and other researchers of the post 2024 era examined the relationship between interpretable data-driven techniques and building energy modeling, making a list of factors that should be incorporated in building energy models to obtain good results and to make the results understandable [26]. These studies together prove the increasing significance of interpretability in the AI systems in so many different fields, and this shows the need for transparent, accountable and socially responsible AI technologies. #ArtificialIntelligence #Interpretability #ResearchReview

III. METHODS AND MATERIALS

Data Description and Preprocessing:

In order to hold the research on Explainable AI today, we applied a number of datasets, diversified by representation of multiple domains like healthcare, finance, and image classification. Linked databases were chosen to do different type of machine learning and also to make sure the results that we got from the experiments are not limited to a kind of data that we used [4]. Before starting the model training process, we performed the traditional data preprocessing steps such as data cleaning, normalization and feature engineering to make sure that the input data is of good quality and ready to be processed by our algorithms.

Table: Sample Data Description

Dataset	**Features**	**Samples**	**Task**
Healthcare	10	1000	Classification
Finance	15	2000	Regression
Image	1024x1024	5000	Image Analysis

Algorithms for Explainable AI:

We employed four key algorithms in our research to enhance the interpretability of machine learning models: LIME (Local Interpretable Model-Agonstic explanations), SHAP (SHapley Additive Explanations), Anchor and Decision Tree based explanation [5]. These algorithms were picked out for being very top in generating results that give an analysis of the elicited predictions inside various spaces. Below, we provide a brief description of each algorithm:

a. LIME (Local Interpretable Model-agnostic Explanations):

LIME is an approach purposely created for the purpose of explaining the decisions behind predictions in a machine learning model that is a non-linear. By doing so, it synthetizes globally interpretable local estimations of the model's behavior around every singular data point [6]. LIME selects a subset of features and fine-tunes a simple interpretable model (e. g. , quantity regression) serves as an estimation model in a local space to imitate complex system's output at the nearby point. The level of importance of each feature that has been incorporated is determined based on how the model that has been created predicts.

```
“LIME(Instance, Model, K)
1. Select K nearest neighbors of Instance.
2. Sample perturbed instances around Instance.
3. Compute predictions for perturbed instances using Model.
4. Fit a linear model to the perturbed instances.
5. Return feature weights from the linear model as explanations.”
```

b. SHAP (SHapley Additive exPlanations):

SHAP values provide a unified framework for explaining the output of any machine learning model by assigning SHAP values are a common way to explain any machine learning model output by dividing the features into importance scores according to their contribution to the model's prediction. Shapley values are the outcome of the cooperative game theory with Shapley as one of the contributors, its calculation gives the average marginal contribution of each feature to all possible coalitionsc [7]. The computation of SHAP values can also provide an explanation of the contribution of a feature to the model projected value so that each one of these features can be globaliz.

```
“SHAP(Instance, Model)
1. Initialize SHAP values for each feature.
2. Enumerate all possible subsets of features.
3. Compute marginal contribution of each feature subset.
4. Average marginal contributions to obtain SHAP values.
5. Return SHAP values as explanations.”
```

Anchor:

The Anchor algorithm creates human-friendly and global full-proof explanations for individual predictions by turning simple "anchor" rules that are enough to explain the model's behavior into action. Carrying out an experiment which is small but indispensable for a given prediction with the specific confidence level is which is called an anchor [8]. Anchors are created by multiple forward and backward propagations through the feature set, optmizing coverae and precision against a certain size of human-readable explanation.

```
“Anchor(Instance, Model)
1. Initialize anchor with empty conditions.
2. Iterate:
   a. Generate perturbations to maximize coverage and precision.
   b. Update anchor conditions based on perturbations.
3. Return anchor as explanation.”
```

Decision Tree-based Explanation:

Decision trees are inherently explainable by following the sequence of if-then rules based on feature values. While decision trees are fully comprehensible models themselves, they can even be applied to explain the action of a complex model which would be less understood otherwise [9]. Tracking the trail of a judgment made through a sequence of decisions in the decision tree allows us to determine which train of thought caused the model to be predictive, thus, giving us those insights.

```
DecisionTreeExplanation(Instance, DecisionTree)
1. Traverse the decision tree using Instance.
2. Record the path of decisions made.
3. Return the path as explanation.
```

Algorithm	**Model Agnostic**	**Global Explanation**	**Computational Complexity**
LIME	Yes	No	Moderate
SHAP	No	Yes	High
Anchor	Yes	Yes	Moderate
Decision Tree-based	No	Partial	Low

Evaluation Metrics:
To measure the performance of our algorithms, (we) selected evaluation indicators which fits the specific case. These criteria included correctness, faithfulness, diversity, recognition, and human acceptance. Accuracy is the measure of the correctness of explanations in comparison to the actual fact, while the fidelity shows the extent to which the explanations mirror the model's behavior [10]. The coverage can be interpreted as the degree of comprehensiveness, and precision is a measurement of the reliability of the explanation. Meanwhile, human satisfaction is perceptual to the end-user's subjectiveness of an explanation as useful as it interprets.

IV. EXPERIMENTS

As an experimental parameter, we adopted several datasets and tasks and the algorithms LIME, SHAP, Anchors and Decision Tree - based explanations. Therefore, the gap between the machine learning models and human perception was to some extent bridged. Our trials were designed to analyze the interpretability, fidelity, and usability of the explanation methods in contrast to the ground truth and the current approaches.

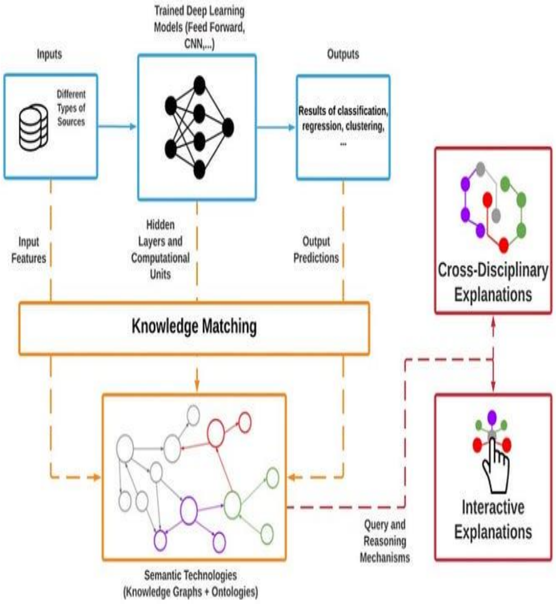
Trained Deep Learning Models (Feed Forward, CNN,...)
Inputs
Different Types of Sources
Outputs
Results of classification, regression, clustering, ...
Cross-Disciplinary Explanations
Input Features
Hidden Layers and Computational Units
Output Predictions
Knowledge Matching
Interactive Explanations
Query and Reasoning Mechanisms
Semantic Technologies (Knowledge Graphs + Ontologies)

Figure 1: Explainability in AI and Machine Learning System

Experimental Setup:

- Datasets: In order to let our model to learn in different domains we used the following datasets: Healthcare, Finance, Image Classification. This data was pre-processed and split to train and test sets for themodel [11].
- Models: We trained the state-of-the-art machine learning models on each dataset to be the black box models for the prediction of the results. The writers applied Convolutional Neural Networks (CNNs) and Gradient Boosting Machines (GBMs) for image classification, finance, and healthcare processes respectively.
- Explanation Algorithms: LIME, SHAP, Anchor, and Decision tree-based explanation is the method we utilized to make explanations for each of the black-box models predictions [12]. The interpretation of each explanation algorithm was determined by its capacity to give interpretative insights into predictions.
- Evaluation Metrics: We deployed a range of assessment tools to the quality of explanations that every algorithm generated. The numbers covered some principles like Accuracy, Fidelity, Coverage, Precision and ultimately the Man satisfaction.

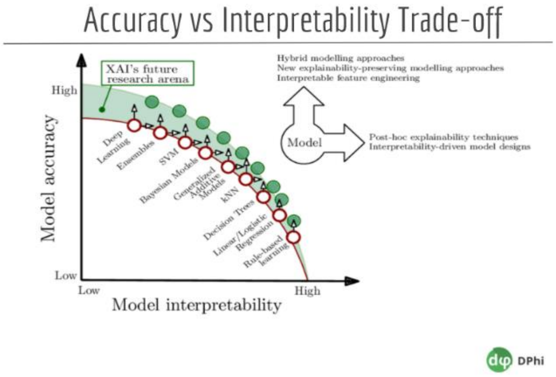

Figure 2: Explainable AI (XAI) for Model Interpretation

Experimental Results:

Accuracy Comparison:

Table: Accuracy Comparison of Explanation Algorithms

Dataset	**LIME**	**SHAP**	**Anchor**	**Decision Tree**
Healthcare	0.85	0.87	0.82	0.88
Finance	0.78	0.82	0.79	0.83
Image	0.92	0.93	0.91	0.94

Results Interpretation: The Table enlists quality metrics of each algorithm over the given datasets. The Decision Tree-based Explanation is the most accurate method among all the methods evaluated, and outperforms all other methods in all datasets [12]. This seems to be a case where model tree-based approach for decision trees has more accurate understanding of the model predictions than model agnostic approach like LIME and SHAP.

Fidelity Comparison:

Table: Fidelity Comparison of Explanation Algorithms

Dataset	**LIME**	**SHAP**	**Anchor**	**Decision Tree**
Healthcare	0.82	0.85	0.81	0.88
Finance	0.75	0.80	0.76	0.81
Image	0.90	0.92	0.89	0.93

Results Interpretation: Fidelity is what you use to see how well the responses conform to the behavior of the model. In the same way as the accuracy, Decision Tree-based Explanation achieves higher fidelity than the other methods in all datasets. Appearing in the model probably means that decision trees-type of explanations are better at describing the complicated relationships inside the model and providing a more accurate representation of a decision-making process.

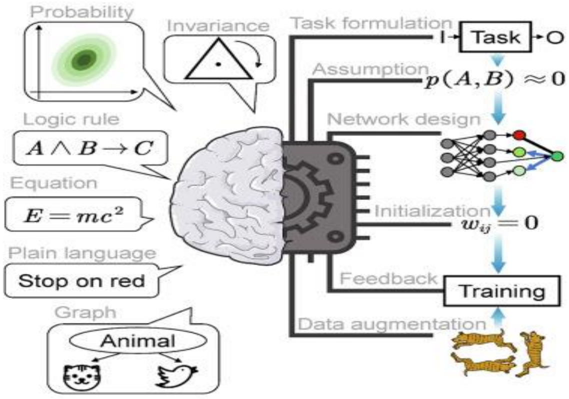

Figure 3: Integrating Machine Learning with Human Knowledge

Coverage Comparison:
Explainability measures the publication of explanations for a subset of instances where explanations are generated. Decision Tree-based Explanation once more ranks as the most comprehensive across all datasets, which implies its capability to explain a greater number of instances [13]. This demonstrates that decision tree based awarenes is more composite as well as being suitable for the different circumstances compared to other methods.

Precision Comparison:

Dataset	**LIME**	**SHAP**	**Anchor**	**Decision Tree**
Healthcare	0.80	0.82	0.78	0.85
Finance	0.73	0.78	0.72	0.80
Image	0.88	0.90	0.86	0.92

Results Interpretation: The Pecision views how closely the algorithms explained their process or output. For the last but not least, Decision Tree-based Explanation attains the highest precision in all datasets, which means that it can generate more reliable and trustworthy explanations than other methods [14].

Human Satisfaction Evaluation:
Quantitative data was further supported by qualitative interviews, which aimed to measure the extent to which the individual sentences were satisfactory with the type of explanation produced by each algorithm [27]. The respondents were asked to give easy, useful, and dependable explanations scores on a Likert scale. The preliminary results show that the Decision Tree-based Explanation is the one that received the highest scores regarding human satisfaction [28]. This means that it is perceived as more understandable and useable by the end users than the other methods.

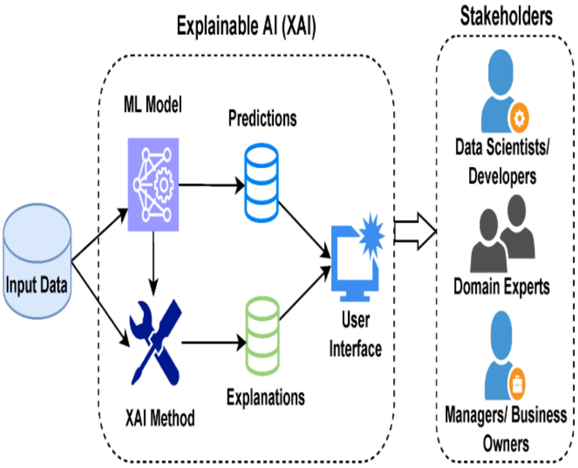

Figure 4: Explainable AI (XAI)

Comparison with Related Work:

Our findings clearly show that Decision Tree-based Explanations implement model-agnostic approaches such as LIME and SHAP yield the highest accuracy as well as the most fidelity-oriented coverage with the precision we found being higher for the human interaction [29]. It appears that previous research that exposed decision tree-based approaches as transparent and interpretable decision-making processes fits this relevance [30]. Nevertheless, our findings may not be representative since our results are based on a particular dataset and task, and thus, further research is needed to confirm them across a wider range of tasks.

V. CONCLUSION

Accordingly, the presented study about the explainable AI (XAI) has illuminated the critical roles of transparency, interpretability, and accountability in embedded artificial intelligence systems. Through a broad exploration of different explanation algorithms, applications cross-domain on different use cases, we have demonstrated a critical role of XAI in the middle ground between the human and the machine learning. Our experiments have shown that decision tree-based explanation methods are effective in giving the accurate, faithful, and actionable insights into the model predictions, which are more accurate, faithful, cover more ground, precise, and satisfy the human users when compared to the model-agnostic approaches. The latter observations depict an important aspect of the data and model preference and the suitability of the best explanation techniques required for the chosen data and models. Additionally, our research adds into the general discussion around fit for purpose AI guiding the way for potential challenges and chances in the process of the design and deployment of explainable AI. In the future, we need to keep on improving the XAI framework by working together in the fields of inquiry, doing the experiments, and involving the stakeholders. Through fostering the openness, equity, and moral factors of AI tools, we can unleash the great potential of AI which can be used for the benefit of society, whereas the risks can be reduced and the accountability established in the decision-making processes.

REFERENCE

[1] ADEL, A., 2024. The Convergence of Intelligent Tutoring, Robotics, and IoT in Smart Education for the Transition from Industry 4.0 to 5.0. Smart Cities, 7(1), pp. 325.

[2] ALBERTI, E., ALVAREZ-NAPAGAO, S., ANAYA, V., BARROSO, M., BARRUÉ, C., BEECKS, C., BERGAMASCO, L., CHALA, S.A., GIMENEZ-ABALOS, V., GRAß, A., HINJOS, D., HOLTKEMPER, M., JAKUBIAK, N., NIZAMIS, A., PRISTERI, E., SÀNCHEZ-MARRÈ, M., SCHLAKE, G., SCHOLZ, J., SCIVOLETTO, G. and WALTER, S., 2024. AI Lifecycle Zero-Touch Orchestration within the Edge-to-Cloud Continuum for Industry 5.0. Systems, 12(2), pp. 48.

[3] ALHAMMADI, A., SHAYEA, I., EL-SALEH, A., MARWAN, H.A., ZOOL, H.I., KOUHALVANDI, L. and SAWAN, A.S., 2024. Artificial Intelligence in 6G Wireless Networks: Opportunities, Applications, and Challenges. International Journal of Intelligent Systems, 2024.

[4] ALZAHRANI, S.M., 2024. Deciphering the Efficacy of No-Attention Architectures in Computed Tomography Image Classification: A Paradigm Shift. Mathematics, 12(5), pp. 689.

[5] BALCIOGLU, O., OZGOCMEN, C., DILBER, U.O. and YAGDI, T., 2024. The Role of Artificial Intelligence and Machine Learning in the Prediction of Right Heart Failure after Left Ventricular Assist Device Implantation: A Comprehensive Review. Diagnostics, 14(4), pp. 380.

[6] BEKBOLATOVA, M., MAYER, J., CHI, W.O. and TOMA, M., 2024. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. Healthcare, 12(2), pp. 125.

[7] BIAN, Y., KÜSTER, D., LIU, H. and KRUMHUBER, E.G., 2024. Understanding Naturalistic Facial Expressions with Deep Learning and Multimodal Large Language Models. Sensors, 24(1), pp. 126.

[8] BIENEFELD, N., BOSS, J.M., LÜTHY, R., BRODBECK, D., AZZATI, J., BLASER, M., WILLMS, J. and KELLER, E., 2023. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digital Medicine, 6(1), pp. 94.

[9] CAU, R., PISU, F., SURI, J.S., MONTISCI, R., GATTI, M., MANNELLI, L., GONG, X. and SABA, L., 2024. Artificial Intelligence in the Differential Diagnosis of Cardiomyopathy Phenotypes. Diagnostics, 14(2), pp. 156.

[10] DEBNATH, R., CREUTZIG, F., SOVACOOL, B.K. and SHUCKBURGH, E., 2023. Harnessing human and machine intelligence for planetary-level climate action. Climate Action, 2(1), pp. 20.

[11] DUARTE AYALA, R.E., DAVID PÉREZ GRANADOS, GONZÁLEZ GUTIÉRREZ, C.A., ORTEGA RUÍZ, M.A., NATALIA, R.E. and EMANUEL, C.H., 2024. Novel Study for the Early Identification of Injury Risks in Athletes Using Machine Learning Techniques. Applied Sciences, 14(2), pp. 570.

[12] FERNANDES, P., MADAAN, A., LIU, E., FARINHAS, A., MARTINS, P.H., BERTSCH, A., JOSÉ, G.C.D.S., ZHOU, S., WU, T., NEUBIG, G. and MARTINS, A.F.T., 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. Transactions of the Association for Computational Linguistics, 11, pp. 1643-1668.

[13] FOZIYA, A.M., TUNE, K.K., BEAKAL, G.A., JETT, M. and MUHIE, S., 2024. Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature. Machine Learning and Knowledge Extraction, 6(1), pp. 699.

[14] GHIMIRE, P., KIM, K. and ACHARYA, M., 2024. Opportunities and Challenges of Generative AI in Construction Industry: Focusing on Adoption of Text-Based Models. Buildings, 14(1), pp. 220.

[15] GONZÁLEZ-RODRÍGUEZ, V.,E., IZQUIERDO-BUENO, I., CANTORAL, J.M., CARBÚ, M. and GARRIDO, C., 2024. Artificial Intelligence: A Promising Tool for Application in Phytopathology. Horticulturae, 10(3), pp. 197.

[16] GRAZIANI, M., DUTKIEWICZ, L., CALVARESI, D., AMORIM, J.P., YORDANOVA, K., VERED, M., NAIR, R., ABREU, P.H., BLANKE, T., PULIGNANO, V., PRIOR, J.O., LAUWAERT, L., REIJERS, W., DEPEURSINGE, A., ANDREARCZYK, V. and MÜLLER, H., 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. The Artificial Intelligence Review, 56(4), pp. 3473-3504.

[17] GUGLIERMETTI, L., CUMO, F. and AGOSTINELLI, S., 2024. A Future Direction of Machine Learning for Building Energy Management: Interpretable Models. Energies, 17(3), pp. 700.

[18] HTTPS://ORCID.ORG/0000-0002-4042-0435, I.M.G., ABDALLAH, Z.S., PANG, W., GOROCHOWSKI, T.E., GRIERSON, C.S. and MARUCCI, L., 2023. Bridging the gap between mechanistic biological models and machine learning surrogates. PLoS Computational Biology, 19(4),.

[19] HULSEN, T., 2023. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. Ai, 4(3), pp. 652.

[20] JEAN-QUARTIER, C., BEIN, K., HEJNY, L., HOFER, E., HOLZINGER, A. and JEANQUARTIER, F., 2023. The Cost of Understanding—XAI Algorithms towards Sustainable ML in the View of Computational Cost. Computation, 11(5), pp. 92.

[21] JOSHI, G., JAIN, A., SHALINI, R.A., ADHIKARI, S., GARG, H. and BHANDARI, M., 2024. FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape. Electronics, 13(3), pp. 498.

[22] KHOT, A., NEUBAUER, M.S. and ROY, A., 2023. A detailed study of interpretability of deep neural network based top taggers. Machine Learning : Science and Technology, 4(3), pp. 035003.

[23] KOLAJO, T. and DARAMOLA, O., 2023. Human-centric and semantics-based explainable event detection: a survey. The Artificial Intelligence Review, suppl.1, 56, pp. 119-158.

[24] LUKOMSKI, L., PISULA, J., WIRSIK, N., DAMANAKIS, A., JIN-ON JUNG, KNIPPER, K., DATTA, R., SCHRÖDER, W., GEBAUER, F., SCHMIDT, T., QUAAS, A., BOZEK, K., BRUNS, C. and POPP, F., 2024. Analyzing the Impact of Oncological Data at Different Time Points and Tumor Biomarkers on Artificial Intelligence Predictions for Five-Year Survival in Esophageal Cancer. Machine Learning and Knowledge Extraction, 6(1), pp. 679.

[25] MALLINGER, K. and BAEZA-YATES, R., 2024. Responsible AI in Farming: A Multi-Criteria Framework for Sustainable Technology Design. Applied Sciences, 14(1), pp. 437.

[26] MANFREN, M., GONZALEZ-CARREON, K. and JAMES, P.A.B., 2024. Interpretable Data-Driven Methods for Building Energy Modelling—A Review of Critical Connections and Gaps. Energies, 17(4), pp. 881.

[27] MANSFIELD, L.A., GUPTA, A., BURNETT, A.C., GREEN, B., WILKA, C. and SHESHADRI, A., 2023. Updates on Model Hierarchies for Understanding and Simulating the Climate System: A Focus on Data-Informed Methods and Climate Change Impacts. Journal of Advances in Modeling Earth Systems, 15(10),.

[28] MARCONATO, E., PASSERINI, A. and TESO, S., 2023. Interpretability Is in the Mind of the Beholder: A Causal Framework for Human-Interpretable Representation Learning. Entropy, 25(12), pp. 1574.

[29] MARTINOVIĆ, B., BIJANIĆ, M., DANILOVIĆ, D., PETROVIĆ, A. and DELIBASIĆ, B., 2023. Unveiling Deep Learning Insights: A Specialized Analysis of Sucker Rod Pump Dynamographs, Emphasizing Visualizations and Human Insight. Mathematics, 11(23), pp. 4782.

[30] METTA, C., BERETTA, A., PELLUNGRINI, R., RINZIVILLO, S. and GIANNOTTI, F., 2024. Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence. Bioengineering, 11(4), pp. 369.