International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v10i4.5538
***Article Received**: 24 February 2022 **Revised**: 12 March 2022 **Accepted**: 15 March 2022 **Publication**: 30 April 2022*

Contemplation of Explainable Artificial Intelligence Techniques

Model Interpretation using Explainable AI

Tavishee Chauhan
Department of Computer Engineering
SCTR's Pune Institute of Computer Technology
Pune, India
chauhantavi@gmail.com

Sheetal Sonawane
Department of Computer Engineering
SCTR's Pune Institute of Computer Technology
Pune, India
sssonawane@pict.edu

Abstract—Machine intelligence and data science are two disciplines that are attempting to develop Artificial Intelligence. Explainable AI is one of the disciplines being investigated, with the goal of improving the transparency of black-box systems. This article aims to help people comprehend the necessity for Explainable AI, as well as the various methodologies used in various areas, all in one place. This study clarified how model interpretability and Explainable AI work together. This paper aims to investigate the Explainable artificial intelligence approaches their applications in multiple domains. In specific, it focuses on various model interpretability methods with respect to Explainable AI techniques. It emphasizes on Explainable Artificial Intelligence (XAI) approaches that have been developed and can be used to solve the challenges corresponding to various businesses. This article creates a scenario of significance of explainable artificial intelligence in vast number of disciplines.

Keywords-Explainable AI, Artificial intelligence, Black box systems, Transparent systems, Model Interpretation

I. Introduction

It's a privilege to live in such a beautiful, fast-paced technical world. This technological planet has proven to be useful in almost every way possible, from saving people's lives to diagnosing diseases. When people are unable to communicate due to the spread of the coronavirus, digital transformation has proven to be essential. Indeed, no other frontier technology has a longer history than artificial intelligence. Even though AI has an academic heritage hooking up back to previous eras, and these features raise expectations that the intellectual ability of these systems is improved, recent advances in computing have grown exponentially the technology's potential. The potential loss of jobs as intelligent computers takes over more and more tasks is worth remembering. Predicting the exact impact of technology across the entire domain of applications is terribly challenging.

Artificial intelligence carries many complications and challenges, but its significance is tremendous. Global progress will necessitate not only the valuable information provided by AI, but also the creativeness and thoughtfulness that only humans connected in a global community can provide. The two types of machine learning algorithms used in AI are white-box and black-box machine learning algorithms [1]. Explainable Artificial Intelligence stands in stark contrast to the "black box" concept in machine learning (XAI). It differs in that black-box models are those in which even the creators are unable to explain why an AI made a specific decision. Transparency, interpretability, and explainability are three concepts that XAI algorithms embrace as shown in Figure 1.

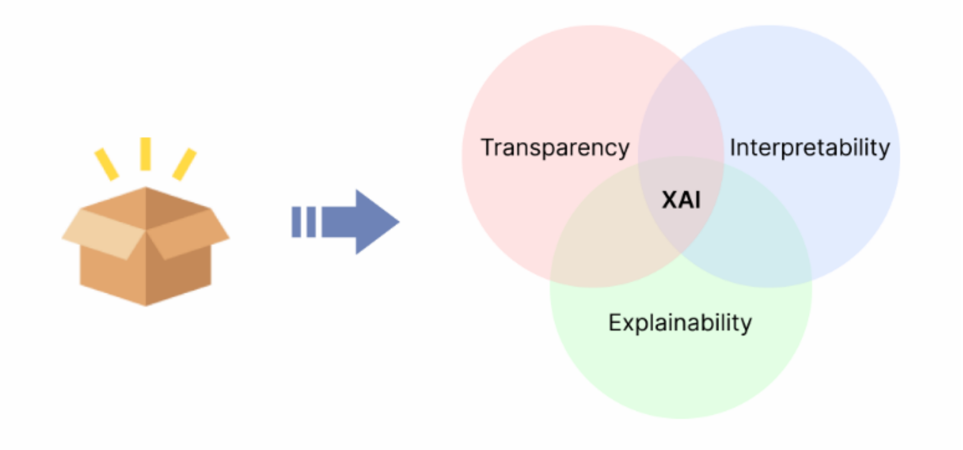


Figure 1. The black-box explained by the major XAI concepts: transparency, interpretability and explainabilit

The XAI research will implement innovative explanation methodologies based on computer results to create more explainable models and outputs. Computational efficiency, design elements, and a variety of other approaches are used to investigate and develop interpretable models of AI machines. Explainable AI has a wide range of applications, including banking, financial services and insurance (BFSI), healthcare, automobiles, and judicial systems [2]. An AI system may make mistakes in the justice or health care systems. For example, it can completely destroy a human life, lowering the user's trust in these systems and limiting their use. That is why explainable AI is so important in healthcare. Both doctors and patients must understand the rationale for major AI recommendations, such as orthopedic surgeries or hospitalizations. XAI provides interpretable explanations in simple language or other simple formats, allowing doctors, clients, and other participants to better understand the reasoning behind a decision.

The emphasis of this research is on Explainable AI in the domain of image processing. Image processing is a technique for modifying images or extracting useful information from them.

The Figure 2 below depicts how explainable AI works when a simple black-box model is used. Explainable AI provides three main insights to reach the desired questions relevant to the model output:

- Justifies the reasons in possible regions of enhancements
- Explain specific reasons for behaviour of model
- Human insights along with the justification of explainable AI helps to reach an accurate decision

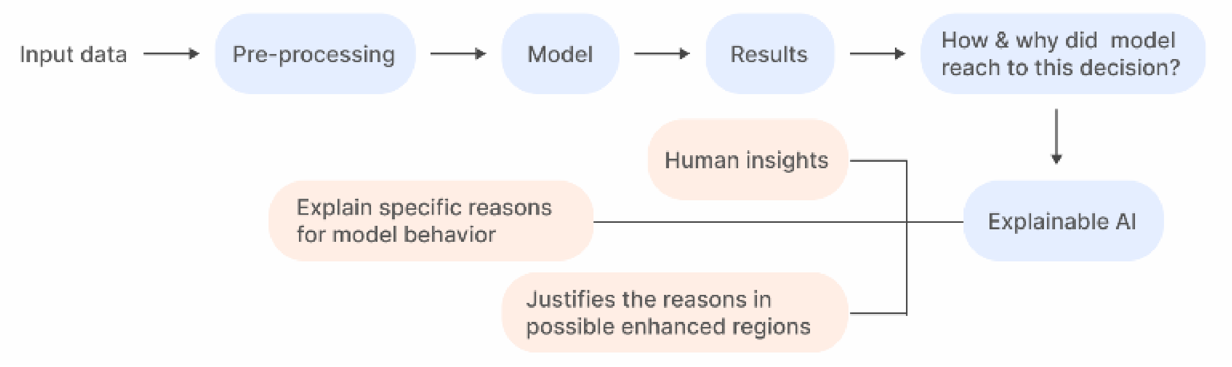

Input data
Pre-processing
Model
Results
How & why did model reach to this decision?
Human insights
Explain specific reasons for model behavior
Explainable AI
Justifies the reasons in possible enhanced regions

Figure 2. General flow of a model that utilizes explainable AI

The remaining elements of the paper is structured as follows: The following section, 2, describes the work of various authors. It describes the authors' approaches and techniques for Explainable AI in various domains. Section 3 also discusses the significance of explainable AI and existing explainability techniques. Section 3 discusses a variety of applications that have incorporated Explainable AI techniques into their systems. Section 5 delves into the challenges of explainable AI. Finally, the sixth section describes the survey's findings.

II. Utilization of Explainable AI in various scenarios

The goal of XAI is to decode the properties of various statistical and machine learning models, such as neural networks, random forests, and decision trees, to make more accessible and meaningful predictions [3]. The authors conducted a systematic SDM analysis on the African elephant and demonstrated several XAI tools, such as local interpretable model-agnostic explanation (LIME), to predict the model's performance [4].

Deep learning models were trained with cutting-edge capabilities on the benchmark BigEarthNet and SEN12MS datasets. Ten XAI approaches were used to comprehend and interpret the models' predictions, as well as quantitative indicators to compare and evaluate their performance. Several tests were carried out to assess the performance of XAI algorithms in scenarios involving simple prediction, competing multiple labels, and misclassification. According to the authors' findings, Occlusion, Grad-CAM, and Lime were the most interpretable and trustworthy XAI approaches [5]. For the automatic detection of COVID-19, a comprehensible deep neural networks (DNN)-based method is used. Explanations are offered using class activation maps (Grad-CAM++) and layer-wise relevance propagation (LRP). CXR images are thoroughly preprocessed in this paper, then augmented and classed using the ensemble technique, and finally the class activation map is implemented using Grad-CAM++ and layer-wise relevance propagation to highlight the class-discriminating regions (LRP) [6].

The XAI tool is used in this paper to help interpret the SDM (Species Distribution Model) model's local-scale behavior. Explainable AI (XAI) is a developing artificial intelligence technology that is used as a toolbox for deeper understanding SDMs. For instance, on the African elephant, a predictable SDM analysis is performed, and several XAI tools, such as local interpretable model-agnostic explanation (LIME), are demonstrated to aid in the interpretation of the model's small-scale behavior. Finally, the advantages and disadvantages of these strategies are debated, with the conclusion that they should be used to improve the interpretability of machine learning SDMs [4]. The XGBoost Model is used to forecast infection severity. The images were used to examine the patterns, distribution, and CT scores of lung disorders. Six machine learning models were developed to predict the severity of COVID-19. Shapley Additive Explanations (SHAP) were used to output the critical elements from the best model [7].

COVID-19 X-ray images were classified using a custom CNN. The baseline threat evaluation outcome is combined with an advanced deep learning-based analysis of chest X-ray scans to provide an accurate prediction of COVID-19 infection risk. The authors provided the LIME explanation for the prediction made by the custom CNN model to further examine and explain the quality of the prediction [8]. The authors proposed a new deep learning-based method that uses chest X-rays to aid in COVID-19 patient repair and recovery. To classify chest X-rays into three categories, the proposed method employs image enhancement, image segmentation, and a customized stacked ensemble model comprised of four CNN ground learners with Naive Bayes as a meta-learner to identify COVID-19, pneumonia, and normal. Grad-CAM visualization was used to add explainability to the paper in order to increase trust in the medical AI system [9].

This study used eight different deep learning algorithms to identify patients with COVID-19 symptoms. Among the various deep learning approaches, NasNetMobile outperformed the others in terms of accuracy. In addition, as an explainability technique, LIME (Local Interpretable Model-agnostic Explanations) was used [10]. Among the deep neural network designs tested for the COVID CT scan picture classification task, the VGG16 network design was found to be the most influenced by spurious artefact techniques. Explainable AI was used for the classification overview. LIME, RISE, Squaregrid, and direct gradient approaches (Vanilla, Smooth, Integrated) were used, as well as Grad-CAM explainability techniques [11]. According to research, the XAI improved classifier model not only produces interesting and consistent classification results, but it also provides a convincing explanation for the classification results. The results were then compared to the most used grads, CAM, LIME, and SHAP [12].

The various methodologies, clinical deployment problems, and topics requiring future research are presented from the perspective of a deep learning expert developing a platform for healthcare purposes. It provides an overview of the current applications of explainable deep learning for medical imaging tasks such as brain imaging, retinal imaging, breast imaging, CT imaging, X-ray imaging, and skin imaging [13]. The authors first attempt to indicate the existence of pneumonia in a chest X-ray, then reveal the presence of pneumonia in a chest X-ray, and finally reveal the presence of pneumonia in a chest X-ray. Using the GRAD-CAM activation maps technique, the areas in the X-ray that are indicative of the presence of COVID-19 are located using the explainable AI for identifying the difference between COVID-19 and pneumonia [14].

A distinctive transfer-trained dual-domain DCNN (Deep-learning convolutional neural networks) architecture built from the AlexNet model trained on ImageNet data was used to process the aspects of each DCE-MRI ROI. The authors of this article demonstrated DCNN learning using XAI tools such as the Integrated Gradients attribution method and the SmoothGrad sound absorption algorithm [15]. The authors suggested a B5G strategy that includes improved detection of COVID-19 using chest X-ray or CT scan images, as well as a massive surveillance framework that tracks social distancing, mask use, and body temperature. The proposed framework ends up looking into three deep learning models: Deep Tree, ResNet50, and Inception v3. In addition, an explainable AI technique called GRAD-CAM activation map is used to explain the images [16].

Explainable AI (XAI) is a field that develops strategies to help us understand AI system predictions. The authors investigate XAI as a technique for using AI-based systems to examine and make a diagnosis health data, as well as a proposed solution for enhancing transparency, availability, and model refinement in the medical space in this paper [17].

The Table 1 below shows literature survey analytical table, representing comparative approaches incorporated by various authors and their specific techniques.

TABLE I. LITERATURE SURVEY ANALYTICAL TABLE

Sr.no.	Title	Approach	Techniques
1	Kakogeorgiou, et al. & 2021 [5]	LIME, Grad-CAM, DeepLift, Saliency and Occlusion approach was used by authors	XAI methodologies
2	Karim, et al. & 2020 [6]	The method for automatic detection of COVID-19 is based on explainable deep neural networks (DNN). Explanations are provided using class activation mappings (Grad-CAM++) and layer-wise relevance propagation (LRP)	XAI – Grad-CAM++ and LRP
3	Ryo, et al. & 2021 [4]	The XAI tool is used to interpret the SDM (Species Distribution Model) model's local-scale behaviour	Local interpretable model-agnostic explanation (LIME)
4	Zheng, et al. & 2021 [7]	For the prediction of covid-19, six machine learning models were examined, with XGBoost outperforming the others	SHAP
5	Sharma, et al. & 2021 [8]	The classification of covid-19 X-ray images was done with a custom CNN	LIME explanation
6	Singh, et al. & 2021 [9]	The chest X-ray is classified into three groups using pruned ensemble learning: normal, pneumonia, and covid-19	Grad-CAM visualizations
7	Ahsan, et al. & 2020 [10]	MobileNetV2 and NasNetMobile	LIME explanation
8	Palatnik de Sousa, et al. 2021 [11]	The categorization summary was done using Explainable AI. Image classification was accomplished using VGG and DenseNet	The techniques used included GradCAM, LIME, RISE, Squaregrid, and direct Gradient approaches
9	Ye, et al. & 2021 [12]	The XAI enhanced classifier model produced intriguing and consistent classification findings, as well as a convincing explanation of the result	Grad-CAM, LIME and SHAP
10	Singh, et al. & 2020 [13]	It provides an overview of the current applications of explainable deep learning in medical imaging	The viewpoint of explainable deep learning is discussed

III. MODEL INTERPRETABILITY AS A WHOLE AND EXPLAINABLE AI APPROACHES

Suppose a physician employs artificial intelligence to predict the onset of COVID-19 sickness. According to the AI, one of the patients was diagnosed with a COVID-19 positive report. When the doctor informs the patient of their positive results, he or she is terrified. After that, the patient is expected to ask two crucial questions:

DOI: https://doi.org/10.17762/ijritcc.v10i4.5538
***Article Received**: 24 February 2022 **Revised**: 12 March 2022 **Accepted**: 15 March 2022 **Publication**: 30 April 2022*

- Can you explain why positive report was predicted?
- Next, is there anything that can be done to prevent it?

To answer these questions, explainable AI techniques will be necessary. Similar situations can arise in a variety of industries, including health care, banking, automobiles, and many others. Model interpretability methods are shown in the diagram below.

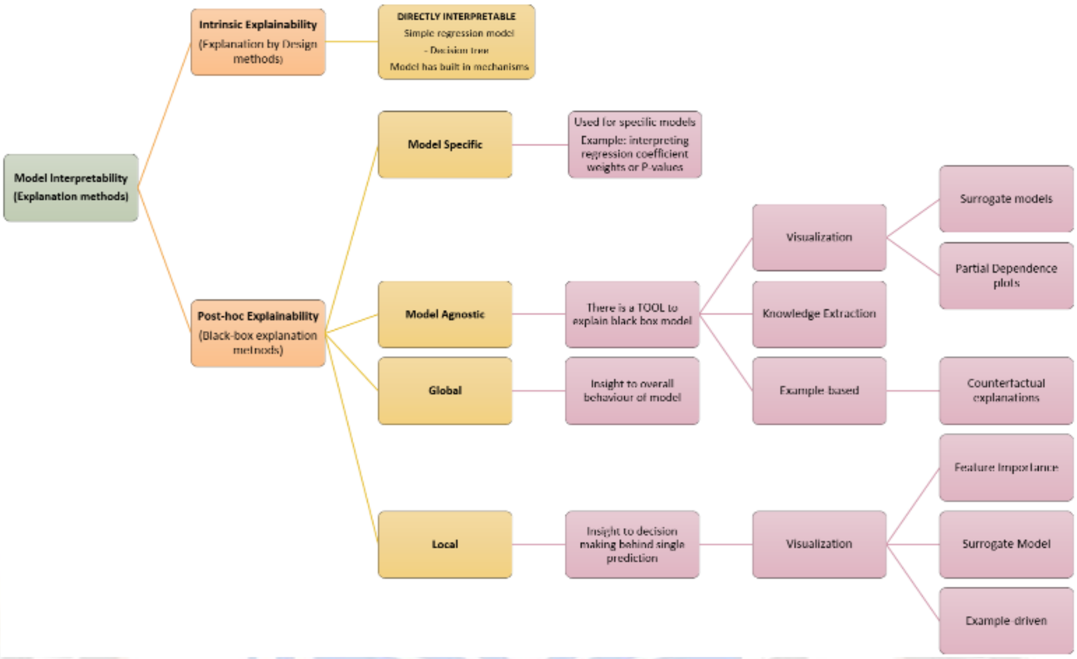
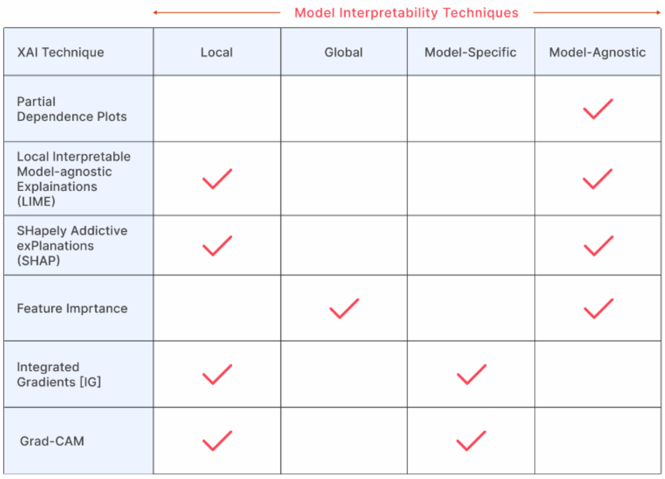

Model Interpretability (Explanation methods)
Intrinsic Explainability (Explanation by Design methods)
DIRECTLY INTERPRETABLE
Simple regression model
- Decision tree
Model has built in mechanisms
Post-hoc Explainability (Black-box explanation methods)
Model Specific
Used for specific models
Example: interpreting regression coefficient weights or P-values
Model Agnostic
There is a TOOL to explain black box model
Visualization
Surrogate models
Partial Dependence plots
Knowledge Extraction
Example based
Counterfactual explanations
Global
Insight to overall behaviour of model
Local
Insight to decision making behind single prediction
Visualization
Feature Importance
Surrogate Model
Example driven

Figure 3. Methodologies currently acquired by explainable AI

Model interpretability divides explanation techniques into two groups as shown in Figure 3: intrinsic explainability and post-hoc explainability. Intrinsic explainability involves procedures that are directly interpretable, as shown in Figure 3, methodologies currently obtained by explainable AI. Simple regression models and decision trees, for example, belong within this group. Post-hoc explainability, on the other hand, entails explaining black-box models. As a result, depending on other sub-categories, explainable AI strategies can be designed within this area. Model specific, model agnostic, global, and local post-hoc methods are the most common types. Various explainability strategies are included in each of these sub-categories.

The Figure 4 represents Explainable AI (XAI) techniques classified with respect to model interpretability, which describes explainable AI techniques for various interpretability methods.



Model Interpretability Techniques

XAI Technique	Local	Global	Model-Specific	Model-Agnostic
Partial Dependence Plots				✓
Local Interpretable Model-agnostic Explainations (LIME)	✓			✓
SHapely Addictive exPlanations (SHAP)	✓			✓
Feature Imprtance		✓		✓
Integrated Gradients [IG]	✓		✓	
Grad-CAM	✓		✓	

Figure 4. Explainable AI (XAI) techniques classified with respect to model interpretability

IV. Technology Utilization in Range of Domains

Machine learning's meteoric rise has ushered in a new era of AI applications in a variety of fields, including transportation, defense, healthcare, banking and finance, and military, that offer enormous benefits yet are unable to explain their decisions and actions to humans. Explainable artificial intelligence (XAI) is a DARPA effort aimed at producing AI technology with simple learning models that consumers could understand and trust [18]. Most domains, including healthcare, finance, education, and cars, use explainable AI.

Figure 5 shows applications utilizing explainable AI, describes most essential domains which utilize explainable AI.

It's critical to think about how the AI system is built and interacts with the vehicle. It has the potential to save lives. What actions should a self-driving automobile take if it finds itself in an inevitable accident situation? Should driver safety take precedence over passenger safety, or should pedestrian safety take precedence over passenger safety? These issues are difficult to answer and are constrained by time constraints. The AI black box model does not work in this scenario. Whether you're a passenger or a pedestrian, you'll need to be ready to explain each decision you make.

Explainable Intelligence solutions could save doctors a lot of time by allowing them to focus on the explanatory work of medicine rather than a repetitive task. Rather than spending time providing requirements to a single patient, more patients may receive attention. Explainable AI allows a computer to analyze data and draw conclusions while also presenting decision line data to a medical expert so they can understand how a specific result was reached. If some scenarios lead to a different conclusion, however, human interpretation may be required.

AI can be used to seek credit ratings, assess insurance payments, and appraise financial assets, among other things. However, biased findings from programs might damage a company's reputation and result in serious action.

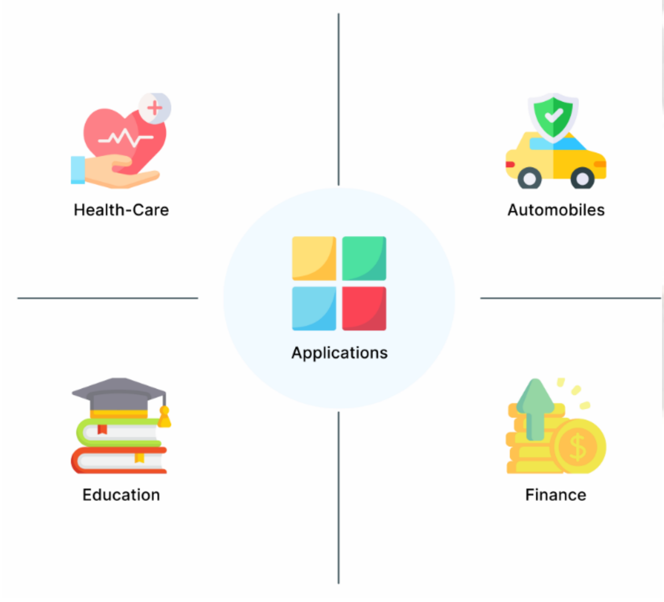


Figure 5. Applications utilizing explainable AI

V. Challenges

Getting biasness in the research is a small concern that may be handled with the use of sound research methods. However, the intrinsic black-box nature, which leads to higher generalization ability, might become a problem in slightly elevated decision making if appropriate explanations for the decisions are not really given. It's critical to understand when a model would function effectively and when it will collapse, as well as why the model is generating accurate predictions and how trustworthy the model is. When these crucial questions are answered, AI technologies in vital decision-making applications will gain more trust.

An XAI model will begin with the creation of a black box model, followed by the creation of a post hoc justification for how the model came to be in the current condition. XAI aids in examining the underlying fundamentals of a black box to understand the value of individual functionalities and the judgments it can drive into, whereas interpretability aids people in understanding the source and consequence of a model prediction.

VI. Conclusion

This research investigates numerous model interpretability methods and the necessity of implementing these methods into systems. The purpose of this research is to investigate into explainable artificial intelligence systems and their potential applications in various fields. It focuses on diverse model interpretability strategies, particularly in relation to Explainable AI techniques. It focuses on Explainable AI (XAI) approaches that have been developed and can be applied to a range of businesses to solve challenges. This research is beneficial to anyone interested in learning more about Explainable AI, its methodology, the need for it, and how it has been applied to diverse fields.

References

[1] L. Gonzalez, Octavio, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view.", IEEE Access 7 (2019): 154096-154113.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] A. Adadi, & M. Berrada. (2018), “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”, *IEEE access*, *6*, 52138-52160.

[3] F. Emmert-Streib, O. Yli-Harja & M. Dehmer (2020), “Explainable artificial intelligence and machine learning: A reality rooted perspective”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), e1368.

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

[5] I. Kakogeorgiou, & K. Karantzalos, (2021), “Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote

sensing", *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102520.

[6] Md. R. Karim, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann & S. Decker. (2020, December), "Deepcovidexplainer: explainable COVID-19 diagnosis from chest X-ray images", In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1034-1037), IEEE.

[7] B. Zheng, Y. Cai, F. Zeng, M. Lin, J. Zheng, W. Chen, G. Qin, and Y. Guo, "An interpretable model-based prediction of severity and crucial factors in patients with COVID-19.", *BioMed Research International* 2021 (2021).

[8] V. Sharma, Piyush, S. Chhatwal & B. Singh. (2021), "An explainable artificial intelligence based prospective framework for COVID-19 risk prediction", *medRxiv*.

[9] R. K. Singh, R. Pandey & R.N. Babu, (2021), "COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays", *Neural Computing and Applications*, *33*(14), 8871-8892.

[10] M. M. Ahsan, K. D. Gupta, M. M. Islam, S. Sen, M. L. Rahman & M. S. Hossain, (2020), "Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities", *Machine Learning and Knowledge Extraction*, *2*(4), 490-504.

[11] I. Palatnik de Sousa, M. M. B. R. Vellasco & E. Costa da Silva, (2021), "Explainable artificial intelligence for bias detection in covid ct-scan classifiers", *Sensors*, *21*(16), 5657.

[12] Q. Ye, J. Xia & G. Yang, (2021, June), "Explainable AI for COVID-19 CT classifiers: an initial comparison study", In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 521-526), IEEE.

[13] A. Singh, S. Sengupta & V. Lakshminarayanan, (2020), "Explainable deep learning models in medical image analysis", *Journal of Imaging*, *6*(6), 52.

[14] L. Brunese, F. Mercaldo, A. Reginelli & A. Santone, (2020), "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays", *Computer Methods and Programs in Biomedicine*, *196*, 105608.

[15] Z. Papanastasopoulos, R. K. Samala, H. P. Chan, L. Hadjiiski, C. Paramagul, M. A. Helvie & C. H. Neal, (2020, March), "Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI", In *Medical imaging 2020: Computer-aided diagnosis* (Vol. 11314, p. 113140Z), International Society for Optics and Photonics.

[16] M. S. Hossain, G. Muhammad & N. Guizani, (2020), "Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics", *IEEE Network*, *34*(4), 126-132.

[17] U. Pawar, D. OâShea, S. Rea and R. OâReilly, "Explainable ai in healthcare", In 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pages 1–2, 2020.

[18] D. Gunning & D. Aha, (2019), "DARPA's explainable artificial intelligence (XAI) program", *AI magazine*, *40*(2), 44-58.