 doi:10.20944/preprints202110.0324.v1

STUDY ON SIGNIFICANT DRIFT IN THE DOMAIN OF EXPLAINABLE ARTIFICIAL INTELLIGENCE

Hemant Palivela[1] and Tavishee Chauhan[2]

[1]Department of Artificial Intelligence, Mumbai
[2]Department of Computer Science, Savitribai Phule Pune University, Pune

Correspondence: Hemant Palivela, Department of Artificial Intelligence, Mumbai (hemant.datascience@gmail.com)

ABSTRACT

Artificial Intelligence (AI) is required since multiple resources are in need to complete depending on a daily basis. As a result, automating routine tasks is an excellent idea. This reduces the foundation's work schedules while also improving efficiency. Furthermore, the business can obtain talented personnel for the business strategy through Artificial Intelligence. Explainability in XAI derives from a combination of strategies that improve machine learning models' environmental flexibility and interpretability. When Artificial Intelligence is trained with a large number of variables to which we apply alterations, the entire processing is turned into a black box model which is in turn difficult to understand. The data for this research's quantitative analysis is gathered from the IEEE, Web of Science, and Scopus databases. This study looked at a variety of fields engaged in the (Explainable Artificial Intelligence) XAI trend, as well as the most commonly employed techniques in domain of XAI, the location from which these studies were conducted, the year-by-year publishing trend, and the most frequently occurring keywords in the abstract. Ultimately, the quantitative review reveals that employing Explainable Artificial Intelligence or XAI methodologies, there is plenty of opportunity for more research in this field.

Keywords: XAI, bibliometric analysis, black box models, artificial intelligence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

INTRODUCTION

Artificial intelligence is on track to revolutionize worldwide economy, working environments, and cultures, as well as generate immense fortune. XAI is a new and developing field that aims to improve the transparency of AI processes. The ultimate purpose of XAI is to assist people in better understanding, trusting, and managing the outcomes of AI technology. The ultimate aim of XAI is to create more explainable models while continuing to improve level of learning performance and accuracy in prediction. Through an in-depth model and data examination of your current AI system, XAI enhances the application of AI in environment. The benefits of XAI can be found across a wide range of sectors and job activities. A few domains include domains of healthcare, insurance, marketing, financial services, autonomous industry and even IT services. Major of these technologies are black boxes, which means we have no idea how they function or why do they make certain decisions.

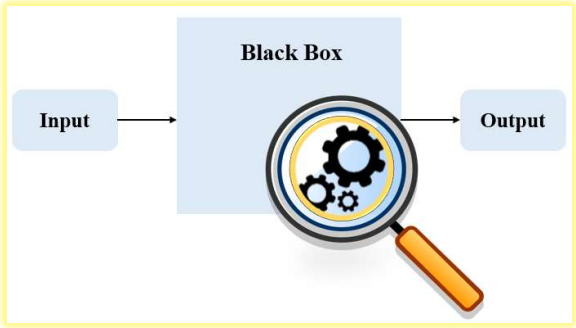

Fig. 1: Black box in Artificial Intelligence

It's risky to rely on black-box conclusions without understanding how they're made. In sectors where black-box judgments can be life-changing and have important ramifications, such as medical diagnosis, crime prevention, and autonomous-driving cars, the need for interpretability is extremely essential.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

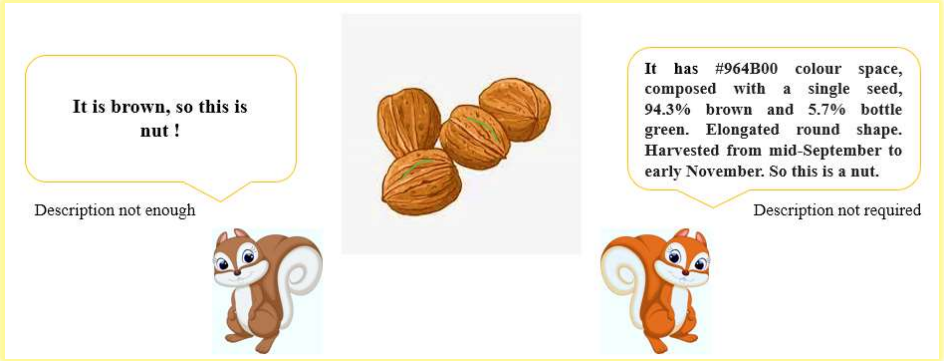

Fig. 2: Justifications should be succinct but thorough

Example as shown in Figure 2, assume an image of a nut is given to a black-box predictive model. It's not enough to simply state, "It's brown, thus it's a nut". Additionally, providing unnecessary explanations also must be avoided. When discussing a black-box decision system, it's critical to provide adequate information in a clear manner. In other words, instructions should be concise but accurate.

In this article [1], research points to the notion of Responsible Artificial Intelligence, which is a technique for large-scale AI application in real businesses that prioritises justice, model interpretability, and responsibility. This article explains how the custom code analysis to custom code transformation process can be automated in a clear and understandable manner. An explainability taxonomy is created as well as it examines the needs in terms of functions [2]. A new approach is offered for synthesizing counterfactuals that incorporates innovative concepts such as counterfactual potential and case-base explanatory scope. The novel method recycles characteristics of good counterfactuals from a case database to create related counterfactuals that can explain fresh issues and solutions [3]. New XAI techniques are frequently founded on an explicit statement of what constitutes to a successful explanation. In this study the authors [4], look at how rule and example-based explanation styles affect system behaviour, contextual relevance, and work engagement in the situation of diabetes management decision support. An overview of the history of Explainable Artificial Intelligence is given by

 doi:10.20944/preprints202110.0324.v1

authors in this article [5]. It is also mentioned about how explainability was traditionally imagined, how it is currently accepted, and how it might be recognised in the future. The authors of this article have used explainable AI (XAI), a developing subdiscipline of artificial intelligence, as a toolkit for better analysing SDMs (Species Distribution Model).

The goal of XAI is to decode the properties of different statistical and machine learning models which include neural networks, random forests, decision trees and create more accessible and meaningful predictions [6]. Considering the African elephant, the authors have done a systematic SDM analysis and demonstrate several XAI tools, like local interpretable model-agnostic explanation (LIME), to predict the model's performance [7]. Many Machine Learning relevant computing systems are opaque, hence it's difficult to understand why they do what they do or how does it work. The goal of authors [8] is to create such a framework, with special consideration paid to the diverse explanatory demands of various stakeholders. The framework differentiates among multiple questions that seek for the description and those that are expected to be asked by various stakeholders and describes the broad methods in which these questions should be addressed in order for these stakeholders to fulfil their responsibilities in the Machine Learning environment. The consistency, accuracy, and trust security features of gradient-based XAI algorithms are investigated using a unique black box attack. The authors in this [9], demonstrate that the proposed system meets the victim's goal of deceiving both the classifier and the explanatory report using three security-based data sets and models, and that only the explainability approach affects the classifier. The two threads that have emerged in the field of XAI are sometimes harmonious and sometimes contradictory. The first is concerned with the creation of practical tools for improving the transparency of automatically taught prediction models, such as deep learning or reinforcement learning. The second aims to foresee the adverse implications of opaque models, with the goal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

of regulating or restricting the repercussions of inaccurate predictions, particularly in crucial fields such as medicine and law [10].

The purpose of this paper is to describe a bibliometric analysis of scientific effort in the field of Explainable Artificial Intelligence. The ultimate purpose of XAI is to assist people in better understanding, trusting, and managing the outcomes of AI technology. The basic goal of XAI is to create more explainable models while yet keeping excellent learning performance. Explainable Artificial Intelligence needs to address the details as mentioned in Table 1.

Table 1: Core fundamental concepts of Explainable Artificial Intelligence

Conceptual principles	Requirements
Explanation	The system needs to efficient enough to provide the description
Meaningful	AI system users must be able to comprehend the description
Reliability	The description should include a comprehensive information of how the AI system generates results
Limitations	The AI system should work within the constraints for which it was created

Understanding and interpreting the outputs of AI systems is becoming increasingly vital as they become more extensively used and are applied to more significant decisions. These objectives make XAI more reliable and better version of Artificial Intelligence.

DATA COLLECTION

The availability of information on the aforementioned topic was examined in several data repositories such as Scopus, Web of Science, and IEEE Xplore. Table 2 lists the terms that were utilised to create the search.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

Table 2: Keywords utilized for generating the search of IEEE database

Base Keyword	Explainable Artificial Intelligence
Concerned keyword	XAI
Concerned keyword	Black-box models

IEEE being the widely used database, features a vast number of highly referenced and peer-reviewed publications and data analytic tools. This database contains materials from a variety of fields, such as artificial intelligence, neural networks, decision making, diseases, medical computing, etc. Whereas Web of Science provides vast number of categories to search within, which include artificial intelligence, robotics and automation, social and behavioural science, life science, business, clinical medicines and many more.

Table 3: Number of paper publications available from various database

Publisher	Number of publications
IEEE Xplore	656
Web of Science	254
Scopus	260

In this paper, we investigated and examined the findings for the study of Explainable Artificial Intelligence, from the IEEE database.

Insights derived from the search

In response to the search, the IEEE database returned 656 papers. The documents in the IEEE database include the time period from 2010 to 2021. In relevance with the various types of publication, the Table 4 mentions about their details. The publication types include Conference papers, Journals, Articles as well as Magazines. Figure 1 denotes the pie chart visualization depending on the publication type.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

Table 4: Distribution of documents with respect to type of publication

(Source: IEEE Database retrieved on 10th August 2021)

Sr. No	Publication Type	Publication count
1	Conference papers	454
2	Journal	158
3	Article	25
4	Magazine	19
	Total	656

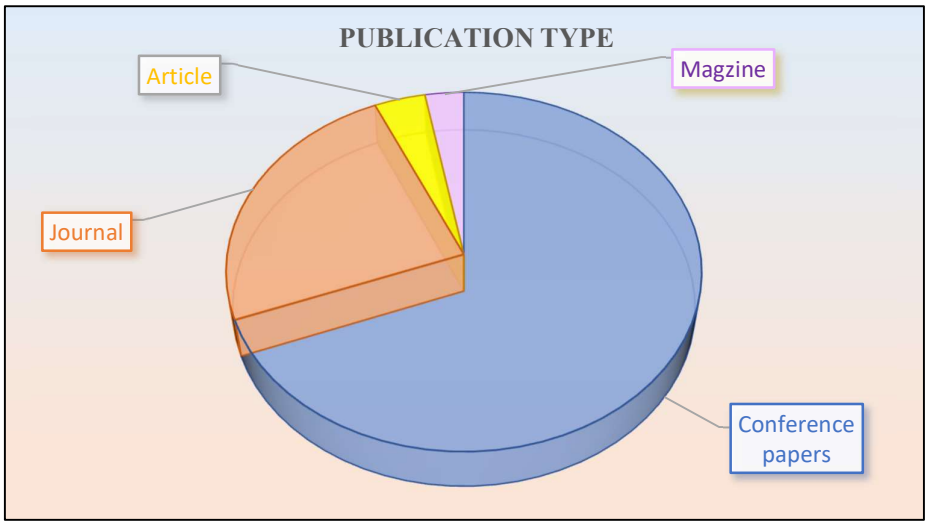

Figure 1: Pie chart distribution relevant to the type of publication

Drift in Publications

After the mid-1980s, the original conceptions of explainability in Artificial Intelligence had faded, along with that in expert systems. Recent successes in machine learning technology, for both autonomous and human interaction systems, with applications in recommender systems, and approaches to neural network learning and reasoning, have brought Explainable Artificial Intelligence back into limelight. As shown in Table 5, it is seen that from the time period of 2010 to 2021, a drastic change can be observed. With the growing trend in the field of XAI, an increase in number of publications is determined with the 297 publications made in the year 2020. It signifies that scholar working in this field are in the most promising period of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

their careers, as beneficial materials must be accessible for study, and there must be room for development in previous work. Figure 2 shows a concise and slow trend in the graph for rise in number of publications in the recent years, among the time period of 2010 to 2021.

Table 5: Drift in number of publications

Year	Number of Publications	Year	Number of Publications
2010	1	2016	1
2011	2	2017	14
2012	0	2018	47
2013	1	2019	134
2014	1	2020	297
2015	0	2021	153

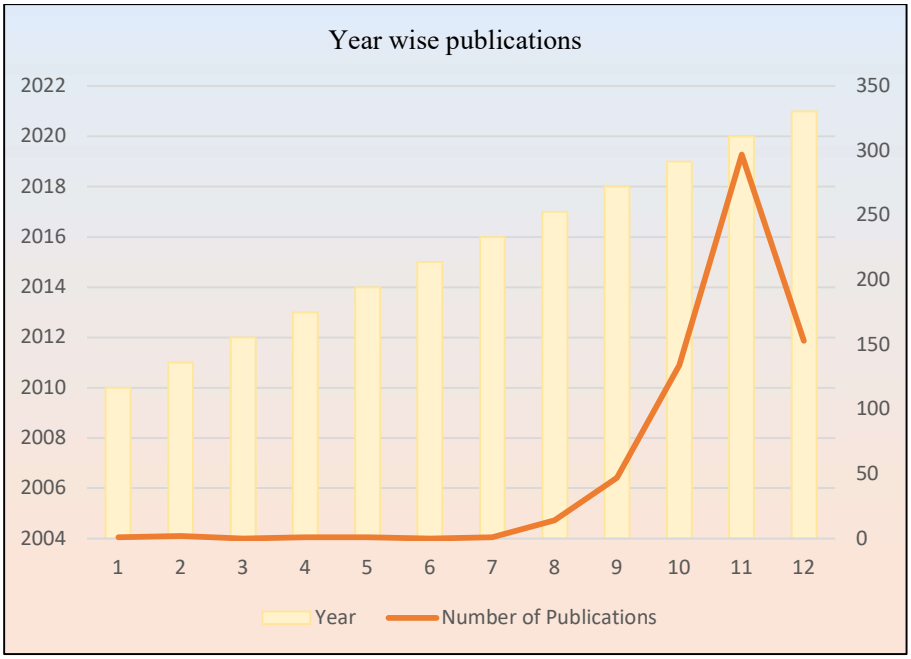

Figure 2: Sudden drift in recent years in number of publications

BIBLIOMETRIC ANALYSIS

Explainable AI (XAI) is artificial intelligence (AI) allows humans to understand the outcomes of the solution and differs with the "black box" concept in machine learning, in which

 doi:10.20944/preprints202110.0324.v1

even its creators are unable to explain why an AI made a particular decision. The authors in this article [11] review and categorise the interpretabilities given by various research studies. The different kinds depict various aspects of interpretability studies, ranging from methodologies that produce interpretable material to complicated pattern investigations. This research [12], delves into the internal dynamics of an analysis to reframe how AI model predictions can be analysed and interpreted using a variety of agnostic strategies, using a case study based on a large database of reinforced concrete (RC) beams with fiber-reinforced polymer (FRP) composite laminates. Statistical models based on time-series analysis are being used to guide policy decisions. The XGBoost technique was used to construct an artificial intelligence model based on machine learning, and additionally feature importance, selective dependence plot, and Shap Value were employed to boost the model's explanatory capabilities [13].

Recent utilization of technologies in XAI

This information is retrieved from IEEE database, the publications are distributed in correspondence with the recent technologies and methodologies. As shown in Table 6, each of the recent technologies and their number of publications is mentioned. The most emerging technologies include Artificial Intelligence, Neural Networks, Pattern classification, Feature extraction, Convolutional neural networks and Image classifications. Around 401 papers have been published considering the present times and all of them are emulsified with Artificial Intelligence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

Table 6: Publications with recent technologies

Recent technologies	Papers published
Artificial Intelligence	401
Neural networks	116
Pattern classifications	94
Feature extractions	67
Convolutional Neural Networks	60
Image classifications	60

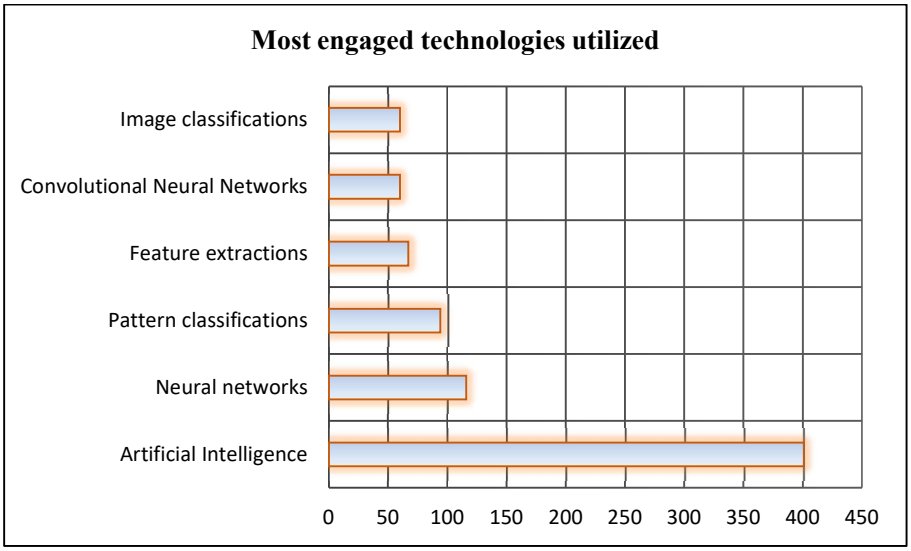

Figure 3: Publications with recent technologies

Geographical analysis

Many advanced and emerging countries have put a strong emphasis on Explainable Artificial Intelligence research. The trend in Table 7 shows that researchers in United States if America have led the global research community by being ranked first, followed by England and Netherlands. Figure 4 shows the distribution of publications in accordance with geographical location. Almost 37% of publications are contributed by USA, followed with 17%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

of England, 16% of Netherland, 15% of United Kingdom, 6% with Germany, 5% of contribution by India, and 4% by China.

Table 7: Top location in publication

Location	Total count
United States of America	99
England	45
Netherlands	44
United Kingdom	39
Germany	17
India	13
China	10

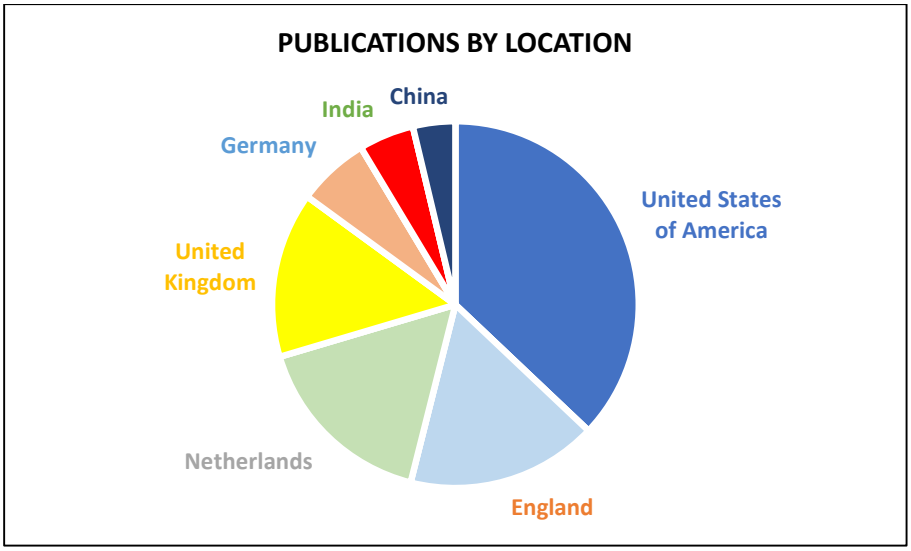

Figure 4: Top location in publication

Analysis based on keywords

The following are some of the keywords that are used in this research study: explainability, black box, transparency, knowledge, intelligence, artificial, predictive model, decision tree, classification, convolutional neural network, natural language, system, data and many more.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

VOS viewer software is used to do keyword-based analysis. In all 96 keywords matched the threshold. The frequency of co-occurrence of words is set to 15 using the VOS viewer. The keyword-based network visualisation is illustrated in Figure 5. It is observed that intelligence is the most occurring keyword.

The density mapping of keywords is shown in Figure 6. Yellow-coloured streaks represent the relatively higher density.

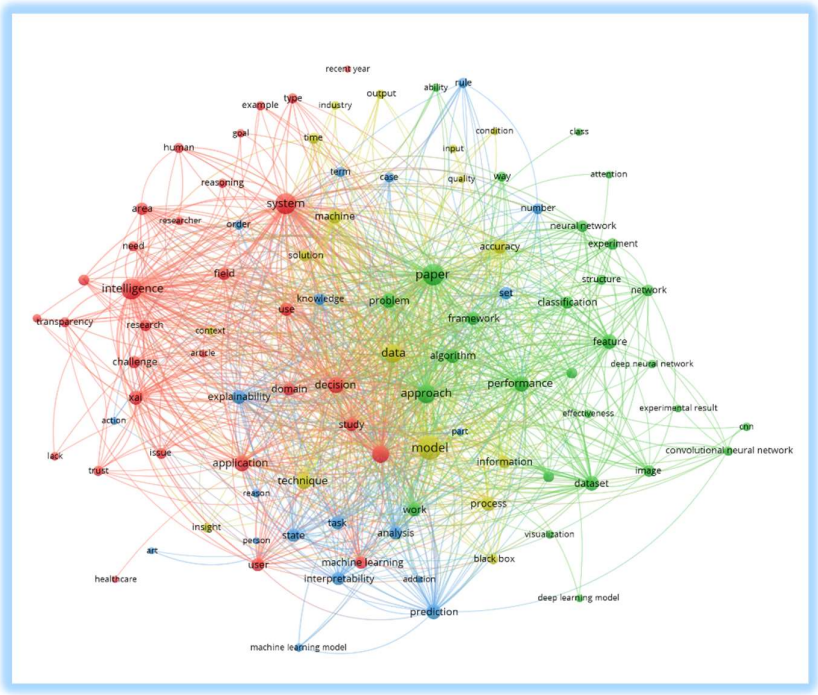

Figure 5: Network map for keyword co-occurrence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2021 doi:10.20944/preprints202110.0324.v1

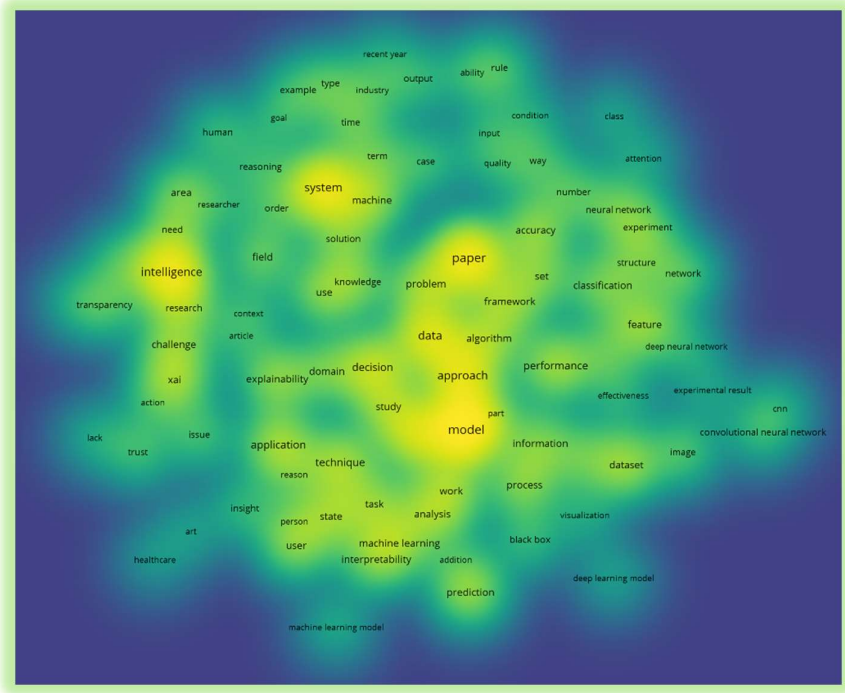

Figure 6: Density diagram for keyword co-occurrence

CONCLUSION

The challenge for Explainable AI (XAI) is to make the decision-making process transparent and rapid. To put it another way, XAI should eliminate these black boxes and explain the decision in detail. The primary goal of XAI is computational transparency. AI systems were formerly thought to be nothing more than black boxes. Despite the fact that the underlying principles of the programming are open access and made publicly accessible, the equations needed to make a choice are always restricted and hence becomes more difficult to understand. In this paper, a bibliometric analysis is represented for the research trend in the domain of Explainable Artificial Intelligence. The data needed for the quantitative analysis of this research was collected from the IEEE, Web of Science as well as Scopus database. Various categories of domains involved in the trend of XAI, highly used techniques, location from

where these researches have been done, year wise publication trend, high occurring keywords in the abstract have been analysed for this research. Additionally, importance of Artificial Intelligence and existing issues are represented which in turn enhances the necessity and need of Explainable Artificial Intelligence. These issues help in making overall model development transparent to stakeholders, and assists in diminishing the problem of black-box model. Ultimately, the quantitative review reveals that employing Explainable Artificial Intelligence or XAI methodologies, there is plenty of opportunity for more research in this field.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Höhn, S., & Faradouris, N. (2021, May). What Does It Cost to Deploy an XAI System: A Case Study in Legacy Systems. In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (pp. 173-186). Springer, Cham.

Keane, M. T., & Smyth, B. (2020, June). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In International Conference on Case-Based Reasoning (pp. 163-178). Springer, Cham.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence, 291, 103404.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), e1391.

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

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

Zednik, C. (2019). Solving the black box problem: a normative framework for explainable artificial intelligence. Philosophy & Technology, 1-24.

Kuppa, A., & Le-Khac, N. A. (2020, July). 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.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020, August). Explainable artificial intelligence: Concepts, applications, research challenges and visions. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 1-16). Springer, Cham.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems.

Naser, M. Z. (2021). An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. Automation in Construction, 129, 103821.

Lee, Y. (2021). Applying Explainable Artificial Intelligence to Develop a Model for Predicting the Supply and Demand of Teachers by Region. Journal of Education and e-Learning Research, 8(2), 198-205.