






Deep Learning Model Interpretation Towards Clarity And Confidence In Artificial Intelligence

1stHumuntal Rumapea ^{1*}, 2ndDarwis Robinson Manalu ², 3rdYolanda Y.P Rumapea ³

¹Doctoral, Computer Science, University Of Methodist Indonesia, Medan, Indonesia

²Doctoral, Computer Science, University Of Methodist Indonesia, Medan, Indonesia

³Magister, Computer Science, University Of Methodist Indonesia, Medan, Indonesia

*Corresponding Author: humuntalrumampeaumi@gmail.com

Citation: Rumapea, Humuntal, Manalu, R, Darwis and Rumapea Y.P, Yolanda, "Deep Learning Model Interpretation Towards Clarity And Confidence In Artificial Intelligence," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 16, no 4, 2024. pp.1786-1794.

ARTICLE INFO

Received: 20 Sep 2024
Accepted: 25 Oct 2024

ABSTRACT

This research aims to analyze the interpretation of deep learning in reading and understanding data, so that artificial intelligence (AI) can assist humans in completing tasks and supporting decision making. The originality of this research lies in AI users, focusing on how deep learning interpretation through the Convolutional Neural Network (CNN) method can improve AI's ability to perform tasks and make decisions. This research uses a qualitative method with a quantitative descriptive approach through a systematic literature review. The results show that Deep Learning is a very powerful tool due to its ability to handle large amounts of data. The use of hidden layers in Deep Learning is proven to surpass the performance of traditional methods, especially in pattern recognition systems. Deep learning models are able to explain, read, and improve AI capabilities simultaneously, and absorb information effectively to improve quality decision making. The study also found that CNN contributes to building trust in AI by improving information accuracy, transparency, and detecting obstacles or anomalies, and can be applied in various fields. A good implementation will increase the confidence of healthcare workers in improving the quality of care, which in turn will reduce mortality, reduce the severity of illness, and improve the quality of life of patients, who should be encouraged to lead a healthy lifestyle

Keywords: deep learning model, convolutional, neural network, ai intelligibility, trustworthiness

INTRODUCTION

The development of technology makes changes to existing life, especially to life and developments in the field of computerization and the digital world. Along with the increasing changes in the digital and computerized world, it is necessary to adapt to these changes, where during the 4.0 period social media appeared which was still not as sophisticated as the 5.0 period, where this social media appeared due to the needs of young people in improving existing styles and styles, so that social media development and applications are needed in order to fulfill the desires and needs of young people on social media. (Chatzimpampas, A., 2020). Along with the development of applications in the 4.0 era through social media, there have been developments in existing computerized systems, where the development of software and hardware has developed significantly, so that an adaptation process to existing computer systems and application networks is needed, so that later people engaged in IT are able to translate and develop existing applications in order to increase understanding of application development from existing algorithms. (Kanse, Abhiraj S., 2022). The development of applications in this 4.0 era is the Internet of Things (IoT) Platforms with Examples application: Google Cloud IoT, AWS IoT Core, and Microsoft Azure which are used to connect physical devices with the internet and collect data in real-time, big data analysis is used to process and analyze large amounts of data to gain valuable business insights or information, Blockchain which is a technology that supports secure and transparent digital transactions, such as cryptocurrencies or smart contracts, robotics and automation which are industrial robots that can work automatically in the production or service process, cloud computing can be developed by providing access to computing resources, storage, and other

services in a flexible and scalable manner, smart manufacturing systems are done by integrating information technology with manufacturing processes for efficiency and precision production, predictive maintenance systems are used by analyzing machine and equipment data to predict when they need maintenance, so as to reduce downtime, and smart logistics and supply chain systems are used by optimizing the analysis of logistics and supply chain operations using data and real-time. (Karran, Alexander John, 2022). The development of these technologies and applications is very much for those who are willing to adapt and adopt such technologies, where technology adoption through the above applications is an example of technologies and solutions that support digital transformation in the Industry 4.0 era. Many companies and industries are currently investing in these technologies to improve efficiency, innovation, and power savings. (Wong, Alexander, Wang, Xiao Yu and Hryniowski, 2020).. Around 2010 until now, various developments from applications and software, as well as existing hardware have emerged because this era is the 5.0 era which is not only to improve development, where now a form of artificial intelligence (AI) application has been developed, where artificial intelligence is a branch in computer science that was born from the process of developing existing networks and devices, where a machine is needed that can carry out tasks that humans do, where this artificial intelligence is able to help problems experienced by humans by helping something that was carried out by humans before by creating an interpretation of the application development model called deep learning. (Markus, Aniek F., Kors & and Rijnbeek, 2021).. Deep Learning model interpretation is the process of understanding and explaining how the model makes predictions or decisions based on given data. While Deep Learning models such as neural networks can provide accurate results, they are often perceived as "black boxes" due to their complexity. While there are many interpretation techniques available, Deep Learning interpretation models remain an active and challenging area of research due to the high complexity and non-linearity of the models. (Kaplan, Andreas and Haenlein, 2018). The interpretation of the Deep Learning model of application development in artificial intelligence must be developed and beneficial to humans, where this interpretation must be interpreted and given a complete understanding of how to adopt a policy and decisions made so that later it does not take the wrong policy and decision in the process of handling problems, so that existing work can be easily applied in depth, and can help human work to be more effective. (Antoniadi, Anna Markella, 2021). Effective Deep Learning interpretation models are key to explaining and understanding how AI makes decisions or predictions. By understanding how the model works, we can provide better explanations to stakeholders and increase trust in AI, where the model consists of a system of tools automatically, and can easily create creativity in addressing problems by being able to solve all existing problems, and help in increasing the productivity of existing work. (Metta, Carlo, 2023). This indicates that this deep learning interpretation model can integrate and record activities experienced by humans, and record existing data to help humans not only create a system that supports work and helps human performance, but also helps in solving problems that plague the system creator or helps increase the power to complete work that cannot be carried out by humans. (Gunning, David and Aha, 2019). It must be recognized that the deep learning interpretation of artificial intelligence can create a different paradigm and help humans understand the algorithms that exist in every artificial intelligence application, where these algorithms can read and record human activities, as well as record data to create the right decision making to improve the smooth running of human work. (Raab, Dominik, Theissler, Andreas and Spiliopoulou, 2023).. The results of the interpretation of deep learning on artificial intelligence (AI) should be able to give confidence to the creators and users of the application so that later with the existing interpretation it is easier for workers to understand all forms of existing data, and be able to describe every problem and solution that exists to increase high confidence, so that the existing interpretation is very useful for human life and can help smooth human work. (Kakogeorgiou, Ioannis and Karantzalos, 2021).. The obstacle in applying deep learning interpretation of artificial intelligence that has been created through a science in computers is that there are still many who misinterpret or are mistaken in interpreting this deep learning interpretation, so that there are still many who do not understand or are mistaken in applying the work and are also late in making decisions, so they are unable to read the existing algorithms, so that the interpretation is not able to fully explain the existing tasks, so it cannot help and facilitate the existing work.

LITERATUREREVIEW

Deep Learning Interpretation

Deep Learning model interpretation is the process of understanding and explaining how the model makes predictions or decisions based on given data. Although Deep Learning models such as neural networks can provide accurate results (Luca Liehner, Gian, 2023). Here are some methods for interpreting Deep Learning models:

1. The weight and activation visualization shows how the model weighs each input feature, as well as showing how information flows through the layers of the neural network.
2. Feature importance uses techniques such as variable permutation or other techniques to find out which features have the most influence on the prediction model.
3. Shapley values use the concept of shapley values to determine the contribution of each feature to the prediction model.
4. Saliency maps can create visualizations to show the part of the input that most influences the prediction model.
5. Lime (Local Interpretable Model-agnostic Explanations) is a technique that generates local explanations for model predictions by considering the model as a "black box".
6. Grad-CAM (Gradient-weighted Class Activation Mapping) uses the gradient of the final layer to understand the part of the image that is important in class decision making.
7. Ablation Study removes one or more model features or layers to see the impact on model performance.
8. Influence analysis looks at how changes in certain data points affect the prediction model.
9. The question-and-answer method involves asking the model questions and analyzing its responses to understand the reasoning behind the predictions.
10. Interpretation dashboards can create visual dashboards to intuitively present interpretation results to stakeholders. (Vilone, Giulia and Longo, 2021).

Interpretation of deep learning models is important:

1. Trustworthiness is the process of making the model more trustworthy to users and stakeholders.
2. Ethical consideration is done by ensuring that the model does not make discriminatory or unfair decisions.
3. Debugging is used to find and fix errors or weaknesses in the model.
4. Optimization is done by understanding the parts of the model that most affect performance for further optimization. (Waa, Jasper van der, 2020).

Deep Learning Interpretation Towards Clarity in Artificial Intelligence (AI)

Effective interpretation of Deep Learning models is key to explaining and understanding how AI makes decisions or predictions. By understanding how the model works, we can provide better explanations to stakeholders and increase trust in AI. (Zhou et al., 2021). Here is how the interpretation of Deep Learning models can help explain AI:

1. Visualization and activation are done by showing how the model weights the input features. For example, in a neural network for image recognition, we can see what features the model considers important, as well as show how information flows through the network, helping in understanding how the model interprets the data.
2. Feature analysis is done by determining which features have the most influence on predictions. By knowing these features, we can explain what underlies the AI decisions.
3. The model interpretation technique is done with Shapley values by showing the contribution of each feature to the prediction. To provide insight into how each feature affects the outcome, saliency maps use the gradient of the output to determine the part of the input that has the most influence on the prediction. This is useful for models such as image recognition, and Lime (Local Interpretable Model-agnostic Explanations) is done by providing local explanations for predictions, considering the model as a "black box".
4. Model Interrogation is done by asking the model questions and analyzing its responses. This can help explain why the model made certain decisions.
5. The interpretation dashboard is done by creating a dashboard that displays the interpretation model and visualizations interactively. This makes it easier for stakeholders to understand and explain the performance model.

6. Case studies using real-life examples to explain how AI works. These can be computations or simulations to show how the model makes decisions.
7. Training sessions are implemented by conducting education sessions for stakeholders, including non-technical stakeholders, to explain the basic concepts and interpretations of deep learning models. (Kögel et al., 2019)..

Benefits of interpreting deep learning models to explain AI:

1. Transparency is done by making AI decisions more transparent and understandable.
2. Trust is done by increasing stakeholder confidence in AI.
3. Ethics is done by ensuring that AI does not reinforce bias or discrimination.

Optimization is the process of understanding the parts of the model that have the most influence on optimization and performance improvement. (Jimenez-Luna, Jose, Grisoni, Franseca and Schneider, 2020).

Deep Learning Interpretation Toward Trust in Artificial Intelligence (AI)

Deep learning interpretation models play an important role in building trust in Artificial Intelligence (AI). Trust is key to the adoption and acceptance of AI technologies by society and businesses. By providing transparent and understandable interpretations, we can increase trust in AI. (Khrais, 2020).

Here's how Deep Learning interpretations can help build trust in AI:

1. Model transparency through weight visualization by showing the weights assigned to input features, allowing stakeholders to see what the model is paying attention to in making decisions, as well as activation visualization by presenting the flow of information through the network layers, aiding in understanding how the model processes data.
2. Feature analysis is performed by identifying the features that have the most influence on predictions, providing insight into what influences the decision model.
3. Model interpretation techniques through shapley values by providing an explanation of each feature's contribution to the prediction, allowing stakeholders to understand what factors influence the outcome, as well as saliency maps by showing the parts of the input that most influence the prediction, provide a visual insight into what the model is paying attention to.
4. Model interrogation involves asking the model questions about its interpretation, then analyzing the model's interpretation and allowing stakeholders to understand the logic behind the model's decisions.
5. Dashboard interpretation is done by creating dashboards that display interpretation metrics and visualization models attractively, making it easy for stakeholders to integrate and understand the performance model.
6. Real-life examples using case studies or reflections to explain how AI works in real situations, display and relevance.
7. The training session was carried out by conducting education and training for stakeholders, explaining the basic concepts and interpretations of the Deep Learning model that will be explained to AI (Aslam et al., 2022)..

Benefits of Building Trust through deep learning interpretation:

1. Wider adoption by increasing the adoption and acceptance of AI technologies by people and businesses.
2. Better understanding that allows stakeholders to understand and explain how AI works.
3. Ethical and safety issues are used to help ensure that AI is used ethically and safely, without bias or discrimination.
4. Innovation and development are carried out by encouraging innovation and development of AI technology by obtaining feedback and better understanding from users. (Papadimitroulas et al., 2021).

Artificial Intelligence Applications

Artificial Intelligence (AI) applications have spread to various sectors and industries, offering innovative solutions to problems and increasing efficiency and productivity. (Schulz et al., 2023). Here are some examples of popular and relevant AI applications in various fields:

1. Health Services (Health)
2. AI systems can assist doctors in diagnosing diseases by analyzing medical images or patient data. AI robots or assistants can help in patient care, including monitoring and wound care. AI is used in research for genome analysis, drug development, and disease prediction models.
3. Finance
4. AI is used to analyze big data for market prediction, fraud detection, and risk management. Customer service can be improved with chatbots that can answer questions and provide information to customers. AI systems can help in investment decision-making with data analysis and prediction.
5. Automotive
6. AI technology is used in autonomous cars for route recognition, vehicle and pedestrian detection, and navigation. AI can monitor vehicle conditions and provide early warnings for maintenance or upkeep.
7. Manufacturing industry
8. AI can predict the maintenance needs of machines and equipment to reduce downtime. AI systems can identify defects or problems in the production process by visual or sensor analysis.
9. Agriculture
10. AI is used for crop condition monitoring, weather prediction, and irrigation optimization. AI systems can identify pests or diseases in crops by image analysis.
11. Education
12. AI can provide customized learning each student's needs and learning style. AI can assist teachers and administrators in analyzing student performance and predicting academic success.
13. Cybersecurity
14. AI is used to detect security threats and analyze malware to protect systems from attacks. AI can monitor network activity to detect suspicious activity or security breaches

METHODOLOGY

The research method carried out is to use a descriptive quantitative approach using the Convolutional Neural Network (CNN) method through a systematic literature review, where (Schmidt et al., 2020) descriptive qualitative research methods with Convolutional Neural Network (CNN) can help us understand the concepts and working principles of one of the most popular types of Deep Learning models in image processing and visual pattern recognition. The data collection techniques can be done using observations and documentation studies. Data analysis is carried out using the Convolutional Neural Network (CNN) approach. (Machlev et al., 2022)..

RESULTS AND DISCUSSION

Deep Learning Models Toward Clarity of Artificial Intelligence

Data analysis conducted using this *deep learning model* is carried out to read the algorithm of the resulting artificial intelligence, where to obtain clarity on the usefulness of existing *artificial intelligence*, a *deep learning* analysis model using the *Convolutional Neural Network* (CNN) approach is needed. CNN is used in applications such as object detection, face recognition, and classification of objects in images, where this deep learning model is a process to understand and explain how the model makes predictions or decisions based on the given data, where deep learning is a basic unit in neural networks that receive input, apply weights, and produce outputs based on activation functions, as well as cycleless neural networks that transfer information from input to output through one or more hidden layers. (Saleem et al., 2022). The *deep learning* model that appears in every artificial intelligence algorithm as a basis for increasing the clarity of artificial intelligence is as follows:

1. *Multilayer Perceptrons*(MLP), where MLP is one of the most basic neural network models that has been used for various tasks such as classification, regression, and multiclass classification. Understanding the structure and optimization of MLPs has helped in the development of more complex and efficient neural network models.
2. *Convolutional Neural Networks* (CNN), where CNNs are specifically designed for image processing and have revolutionized the fields of image recognition, object detection, and image segmentation. CNN's ability to recognize hierarchical features in images has been the basis for the development of other deep learning models for visual tasks.
3. *Recurrent Neural Networks* (RNN) and *Long Short-Term Memory* (LSTM), where RNN and LSTM are used for sequential data such as text, voice, and time series. The ability of RNNs and LSTMs to process and understand temporal data has led to the development of models such as Transformer for processing natural language...
4. *Generative Adversarial Networks* (GANs), where GANs are used to generate realistic new data and have contributed in fields such as learning transfer, data augmentation, and image synthesis. The ability of GANs to create new data has enhanced the creativity and adaptability of AI in understanding and generating visual and audio content.
5. *Transfer Learning* and *Pre-trained Models*, where the use of pre-trained models and transfer learning has accelerated the training process and improved model performance on specific tasks. This approach has enabled the application of AI in various domains without the need for training from scratch, thus increasing the clarity and efficiency of artificial intelligence.
6. *Autoencoders* and *Reinforcement Learning*, where Autoencoders are used for dimensionality reduction and learning representation, while Reinforcement Learning is used for interactive tasks and decision making. These models have added a new dimension to AI's capabilities in processing and understanding complex data and interacting with the environment. (Saleem et al., 2022).

The description of *deep learning* models using *Convolutional Neural Networks* (CNN) models and methods in order to improve the clarity of artificial intelligence can be seen from the following figure:

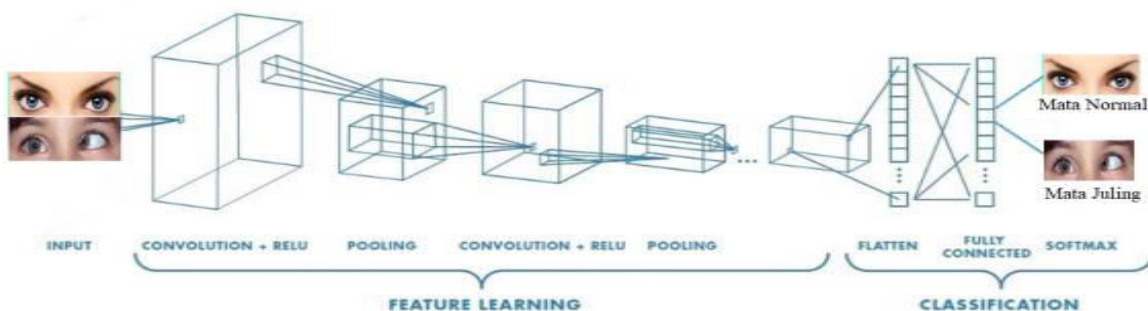


Figure 1 Deep Learning Model Using *Convolutional Neural Networks* (CNN) Method

In recent decades, *Deep Learning* has proven to be a very powerful tool due to its ability to handle large amounts of data. Even the use of hidden layers in *Deep Learning* is able to surpass the performance of traditional methods that have existed before, especially in pattern recognition systems, where *deep learning* models can explain and read AI capabilities simultaneously and improve understanding and absorb them well in order to read data into useful information and benefit users for improved quality decision making. (Ali et al., 2023).

Deep Learning Models Toward Trustworthiness of Artificial Intelligence

Convolutional Neural Networks (CNN) is one of the most successful and influential Deep Learning models in the field of image processing and visual pattern recognition. The success of CNNs has had a positive impact on the trust in Artificial Intelligence (AI) (Knapič et al., 2021).. Here is how CNNs contribute to building trust in AI:

1. High accuracy, where CNNs have demonstrated high accuracy in tasks such as object detection, image classification, and image segmentation, increases the confidence that AI can "see" and "understand" images like humans.
2. Transparency and interpretation, where visualization is the ability to visualize what has been learned by CNNs allowing us to better understand and explain how the model makes predictions, which increases transparency and trust, while interpretation is interpretation techniques such as Grad-CAM and LIME help in explaining the decision model in a way that can be understood, building trust in AI predictions.
3. Constraints and *robustness*, where CNN's ability to detect errors or anomalies can increase confidence that the model can work in a variety of different situations and environments.
4. Application in various fields, where in the medical field, the accuracy and interpretation of CNNs can improve diagnosis and treatment, building confidence in AI applications in medical diagnosis. In automobile autonomy, the accuracy and brightness of CNNs in object detection and route recognition increase confidence in autonomous technology. (Simion & Kelp, 2023).

The research results, *Deep Learning* has proven to be a very powerful tool because of its ability to handle large amounts of data. Even the use of hidden layers in *Deep Learning* is able to surpass the performance of traditional methods that have existed before, especially in pattern recognition systems, where *deep learning* models can explain and read AI capabilities simultaneously and improve understanding and absorb them well in order to read data into useful and useful information for users to improve quality decision making. This is in line with research (Buhrmester et al., 2021) which states that *deep learning* with *Convolutional Neural Networks* (CNN) can make a systematic explanation of the capabilities and algorithms for analyzing specialized data and information to explain that deep learning can generate information to improve the quality of decision making. The results state that CNN contributes to building trust in AI, making the level of information accuracy high, able to increase transparency, able to detect obstacles or anomalies and able to be applied in various fields. This is in line with research (Bao et al., 2019) which states that the level of effectiveness of deep learning will increase high trust in information users in improving decision making correctly, because the data and information presented are accurate, transparent, free of constraints or anomalies, and can be applied in various fields.

CONCLUSION

The research results, *Deep Learning* has proven to be a very powerful tool because of its ability to handle large amounts of data. Even the use of hidden layers in *Deep Learning* is able to surpass the performance of traditional methods that have existed before, especially in pattern recognition systems, where *deep learning* models can explain and read AI capabilities simultaneously and improve understanding and absorb them well in order to read data into useful and useful information for users to improve quality decision making, and the results state that CNN contributes to building trust in AI this makes the level of information accuracy high, able to increase transparency, able to detect obstacles or anomalies and able to be applied in various fields. Deep learning with CNN models will increase clarity and high confidence in the data and information presented in order to improve decisions and predict future business growth progress.

ETHICAL DECLARATION

Conflict of interest: No declaration required. **Financing:** No reporting required. **Peer review:** Double anonymous peer review.

REFERENCES

- [1] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99 (January). <https://doi.org/10.1016/j.inffus.2023.101805>
- [2] Antoniadi, Anna Markella, et al. (2021). Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Journal Applied Sciences*, 11(5088), 1-23. <https://doi.org/10.3390/app11115088>
- [3] Aslam, N., Khan, I. U., Mirza, S., Alowayed, A., Anis, F. M., Aljuaid, R. M., & Baageel, R. (2022). Interpretable Machine Learning Models for Malicious Domains Detection Using Explainable Artificial Intelligence (XAI). *Sustainability (Switzerland)*, 14(12). <https://doi.org/10.3390/su14127375>
- [4] Bao, Y., Tang, Z., Li, H., & Zhang, Y. (2019). Computer vision and deep learning-based data anomaly detection method for structural health monitoring. *Structural Health Monitoring*, 18(2), 401-421. <https://doi.org/10.1177/1475921718757405>
- [5] Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction*, 3(4), 966-989. <https://doi.org/10.3390/make3040048>
- [6] Chatzimparmpas, A., et al. (2020). The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Journal Computer Graphics Forum*, 39(3), 713-756. <https://doi.org/10.1111/cgf.14034>
- [7] Gunning, David and Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *Journal AI Magazine*, 44-58.

- [8] Jimenez-Luna, Jose, Grisoni, Franseca and Schneider, G. (2020). Drug Discovery with Explainable Artificial Intelligence. *Journal Nature Machine Intelligence*, 2, 573-584.
- [9] Kakogeorgiou, Ioannis and Karantzas, K. (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. <https://doi.org/10.1016/j.jag.2021.102520>
- [10] Kanse, Abhiraj S., et al. (2022). Cautious Artificial Intelligence Improves Outcomes and Trust by Flagging Outlier Cases. *Journal JCO Clinical Cancer Informatics*, 1-10. <https://doi.org/10.1200/cci.22.00067>
- [11] Kaplan, Andreas and Haenlein, M. (2018). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Journal Business Horizons*, 1-11. <https://doi.org/10.1016/j.bushor.2018.08.004>
- [12] Karran, Alexander John, et al. (2022). Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Journal Frontiers in Neuroscience*, 16, 1-16. <https://doi.org/10.3389/fnins.2022.883385>
- [13] Khrais, L. T. (2020). The role of artificial intelligence in shaping consumer demand in e-commerce. *Journal Future Internet*, 12(226), 1-14. <https://doi.org/10.3390/fi12120226>
- [14] Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2021). Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. *Machine Learning and Knowledge Extraction*, 3(3), 740-770. <https://doi.org/10.3390/make3030037>
- [15] Kögel, J., Schmid, J. R., Jox, R. J., & Friedrich, O. (2019). Using brain-computer interfaces: A scoping review of studies employing social research methods. *BMC Medical Ethics*, 20(1), 15-17. <https://doi.org/10.1186/s12910-019-0354-1>
- [16] Luca Liehner, Gian, et al. (2023). Perceptions, attitudes and trust toward artificial intelligence - An assessment of the public opinion. *Journal of Artificial Intelligence and Social Computing*, 72, 32-41. <https://doi.org/10.54941/ahfe100327>
- 1
- [17] Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9(May). <https://doi.org/10.1016/j.egyai.2022.100169>
- [18] Markus, Aniek F., Kors, J. A., & and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- [19] Metta, Carlo, et al. (2023). Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning. *International Journal of Data Science and Analytics*, 1-13. <https://doi.org/10.1007/s41060-023-00401-z>
- [20] Papadimitroulas, P., Brocki, L., Christopher Chung, N., Marchadour, W., Vermet, F., Gaubert, L., Eleftheriadis, V., Plachouris, D., Visvikis, D., Kagadis, G. C., & Hatt, M. (2021). Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Physica Medica*, 83(March), 108-121. <https://doi.org/10.1016/j.ejmp.2021.03.009>
- [21] Raab, Dominik, Theissler, Andreas and Spiliopoulou, M. (2023). XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Journal Neural Computing and Applications*, 35, 10051-10068. <https://doi.org/10.1007/s00521-022-07809-x>
- [22] Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., & Liu, L. (2022). Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513, 165-180. <https://doi.org/10.1016/j.neucom.2022.09.129>
- [23] Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), 260-278. <https://doi.org/10.1080/12460125.2020.1819094>
- [24] Schulz, P. J., Lwin, M. O., Kee, K. M., Goh, W. W. B., Lam, T. Y. T., & Sung, J. J. Y. (2023). Modeling the influence

- of attitudes, trust, and beliefs on endoscopists' acceptance of artificial intelligence applications in medical practice. *Frontiers in Public Health*, 11(November), 1-9. <https://doi.org/10.3389/fpubh.2023.1301563>
- [25] Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, 2(1), 447-464. <https://doi.org/10.1007/s44204-023-00063-5>
- [26] Vilone, Giulia and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Journal Information Fusion*, 76, 89-106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- [27] Waa, Jasper van der, et al. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human Computer Studies*, 144, 102493. <https://doi.org/10.1016/j.ijhcs.2020.102493>
- [28] Wong, Alexander, Wang, Xiao Yu and Hryniowski, A. (2020). How Much Can We Really Trust You? Towards Simple, Interpretable Trust Quantification Metrics for Deep Neural Networks. *Journal Research*, 1-13. <http://arxiv.org/abs/2009.05835>
- [29] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics (Switzerland)*, 10(5), 1-19. <https://doi.org/10.3390/electronics10050593>