Navigating the Trade-Off Between Explainability and Privacy

Johanna Schmidt[1,*][a], Verena Pietsch[2][b], Martin Nocker[3][c], Michael Rader[4][d] and Alessio Montuoro[5][e]

[1] *VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Vienna, Austria*
[2] *FILL Gesellschaft m.b.H, Gurten, Austria*
[3] *MCI The Entrepreneurial School, Innsbruck, Austria*
[4] *Fraunhofer Austria Research GmbH, Wattens, Austria*
[5] *Software Competence Center Hagenberg GmbH, Hagenberg, Austria*

Keywords: Explainable AI, Privacy, Explainability, Data Visualization, Homomorphic Encryption.

Abstract: Understanding the rationale behind complex AI decisions becomes increasingly vital as AI evolves. Explainable AI technologies are pivotal in demystifying these decisions, offering methods and tools to interpret and communicate the reasoning behind AI-driven outcomes. However, the rise of Explainable AI is juxtaposed with the imperative to protect sensitive data, leading to the integration of encryption techniques in AI development. This paper explores the intricate coexistence of explainability and encryption in AI, presenting a dilemma where the quest for transparency clashes with the imperative to secure sensitive information. The contradiction is particularly evident in methods like homomorphic encryption, which, while ensuring data security, complicates the provision of clear and interpretable explanations for AI decisions. The discussion delves into the conflicting goals of these approaches, surveying the use of privacy-preserving methods in Explainable AI and identifying potential directions for future research. Contributions include a comprehensive survey of privacy considerations in current Explainable AI approaches, an exemplary use case demonstrating visualization techniques for explainability in secure environments, and identifying avenues for future work.

1 INTRODUCTION

Artificial Intelligence (AI) has witnessed remarkable advancements within the last years, revolutionizing industries and shaping our daily lives. People widely adopt AI for its unparalleled ability to automate tasks, derive insights, and offering enhanced efficiency and innovative solutions across diverse industries and applications. As such, AI algorithms have become a transformative force across various domains (Saghiri et al., 2022). For example, AI aids in diagnostics and personalized healthcare treatments. The financial sector benefits from AI for fraud detection and risk assessment. Mobility applications leverage AI for traffic management. In manufacturing, AI optimizes processes through predictive maintenance and quality control. As AI continues to evolve and AI systems become increasingly sophisticated and integrated into critical decision-making processes, there is a growing demand for transparency and accountability, as decisions are made based on intricate patterns and correlations within vast datasets (Zhang and Lu, 2021).

While AI models demonstrate high accuracy in various tasks, understanding the rationale behind their choices becomes increasingly crucial. Explainability in AI refers to the ability to comprehend and interpret the decisions and actions taken by AI systems (Barredo Arrieta et al., 2020). Explainable AI (XAI) technologies are pivotal in demystifying the decision-making processes of complex algorithms, addressing the challenges of the “black box” nature of AI, and mitigating biases. XAI provides methods and tools to interpret and communicate the rationale behind AI-driven decisions, making them more understandable for users and stakeholders. Techniques range from integrating inherently interpretable models (Wang and Tuerhong, 2022) to model-agnostic approaches (Choo and Liu, 2018). Such model-agnostic

[a] https://orcid.org/0000-0002-9638-6344
[b] https://orcid.org/0000-0002-9146-4649
[c] https://orcid.org/0000-0002-6967-8800
[d] https://orcid.org/0009-0000-1819-5246
[e] https://orcid.org/0000-0002-2605-9081
[*] Corresponding author

Schmidt, J., Pietsch, V., Nocker, M., Rader, M. and Montuoro, A.
Navigating the Trade-Off Between Explainability and Privacy.
DOI: 10.5220/0012472200003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 1: GRAPP, HUCAPP and IVAPP*, pages 726-733
ISBN: 978-989-758-679-8; ISSN: 2184-4321

approaches increasingly use interactive data visualization (Ward et al., 2015) approaches, as these technologies provide intuitive and transparent means for understanding complex model decisions, identifying patterns, and fostering trust among users and stakeholders (Alicioglu and Sun, 2022).

With increasing interest in explainability, protecting sensitive data became paramount, prompting the integration of privacy-preserving technologies in AI development to ensure privacy and security (Timan et al., 2021). Cryptography protects data from unauthorized access, maintaining confidentiality for those possessing the secret key. While traditional encryption techniques protect data only during storage or transmission, Homomorphic Encryption (HE) allows for computations on encrypted data without prior decryption (Yi et al., 2014). HE has already been successfully applied for enabling secure collaboration in scenarios where data confidentiality is critical, like Machine-Learning-as-a-Service (Nocker et al., 2023). HE encrypts model updates that are transferred and collected at the aggregation server. Federated Learning (Esterle, 2022) is commonly utilized to collaboratively train an AI model across multiple distributed data owners while preserving each party's privacy.

The coexistence of explainability and privacy in AI poses a complex dilemma (Ogrezeanu et al., 2022). On the one hand, explainability seeks to demystify the decision-making processes of AI algorithms, enhancing transparency and fostering trust, involving making the inner workings of models interpretable for users and stakeholders. On the other hand, privacy-preserving techniques aim to secure sensitive data and models, rendering them indecipherable to unauthorized entities. The contradiction arises when such methods make providing a clear and interpretable explanation of AI decisions challenging.

In this paper, we discuss the conflicting goals of both approaches. We surveyed the use of privacy-preserving methods in Explainable AI, an underexplored topic, and identified possible directions for future work. Our contributions can be listed as follows:

- **A survey** on how current XAI approaches have considered privacy and security.
- Exemplary use case of using **visualization techniques for explainability** in secure environments.
- Identification of **directions for future work**.

Keeping the balance is crucial for ensuring that AI systems safeguard sensitive information and maintain the transparency necessary for user understanding and acceptance. Addressing this paradox involves exploring innovative hybrid solutions that combine elements of encryption and explainability.

2 RELATED WORK

Working on Explainable and Privacy-Preserving AI touches several scientific directions. A review of different AI technologies and applications (Sheikh et al., 2023) would be beyond the scope of this paper. Instead, we focus on predictive AI, as our use case in this paper is centered around predictive maintenance, on model-agnostic XAI with a focus on using visualization techniques, and methods for privacy-preserving AI.

Predictive AI. AI models for prediction specifically aim at making predictions or forecasts based on data analysis. Predictive AI involves identifying patterns, trends, and relationships within datasets, allowing systems to predict future events or outcomes. Several types of machine learning algorithms are commonly used in predictive AI, each with its strengths and suitable applications. These algorithms include approaches from supervised learning like linear and logistic regression (Lederer, 2021), where temporal data is used to predict future outcomes. Other supervised learning approaches comprise decision trees to build tree-like models out of data (Song and Lu, 2015) and support vector machines (SVM) to classify data points into different categories (Sapankevych and Sankar, 2009). Random forest, an ensemble learning approach (Schonlau and Zou, 2020), combines multiple decision trees to make more informed predictions. Predictive deep learning approaches (Emmert-Streib et al., 2020) include recurrent neural networks (RNN) and long short-term memory (LSTM), which handle sequential data for time series prediction. The choice of algorithm depends on the data's specific characteristics and the prediction task's nature. In practice, a combination of these algorithms, known as ensemble methods (Maimon and Rokach, 2006), is often employed to enhance predictive performance. Predictive AI encompasses a large number of possible applications. For example, predictive AI plays an important role in Industry 4.0 for predictive maintenance applications (Nunes et al., 2023), in finance for calculating forecasts (Broby, 2022), and in energy production to predict energy use (Wang and Srinivasan, 2017). **In this paper we focus on a predictive maintenance use case.**

Explainable AI. Systems and approaches helping users understand the workings of an AI system can be summarized as approaches toward XAI. Explainability covers several understandings (Meske et al., 2022). Explainability can help users understand the system's behavior to detect unknown vulnerabilities and flaws, bias, and avoid phenomena related to spurious correlations – to evaluate the AI model. Especially from a developer's design perspective, understanding the inner workings of AI and consequent outcomes can be vital to increasing the system's accuracy and value with a focus on improvement. From the end user's perspective, the application of XAI can have a positive effect on user trust in the system's decisions (Preece, 2018). Besides inherently interpretable models (Wang and Tuerhong, 2022), which we will not cover on this paper, model-agnostic approaches (Choo and Liu, 2018) provide explanations for the behaviour of existing AI models. Explanations (Yang et al., 2023) can either be source-oriented (i.e., focusing on input data), representation-oriented (i.e., connecting input and output), or logic-oriented (i.e., explaining the inner structure of the model). All three types of explainability approaches (source-oriented, representation-oriented, and logic-oriented) greatly benefit from the use of data visualization to empower humans to explore and understand the results (Samek et al., 2019). Examples for representation-oriented approaches employing visualization include the *Calibrate* framework (Xenopoulos et al., 2023) for the analysis of the output of probabilistic models and approaches for comparing the output of different, similar models (Wang et al., 2023). As examples for a logic-oriented visualization, the *ActiVis* framework (Kahng et al., 2018) aims at visually displaying neuron activity for model input and outputs and the *LSTMVis* (Strobelt et al., 2018) approach to visualize hidden state dynamics in recurrent neural networks. Approaches for evaluating and improving model performances also employ visualization to make the differences understandable (Jin et al., 2023). **We employed source-oriented visualization techniques in our use case.**

Secure AI. Enhancing data security in AI can be achieved through robust encryption strategies. File-level encryption, database encryption, and transparent data encryption (Gasser and Aad, 2023) work at the data level and locally secure the data used by the AI model. End-to-end encryption (Sukhodolskiy and Zapechnikov, 2020) ensures data protection throughout its lifecycle. While such encryption techniques protect data while being transmitted or stored, homomorphic encryption (HE) (Yi et al., 2014) is a technique that enables computations to be performed on encrypted data without prior decryption.

HE for AI facilitates secure collaborative processing and can be used to secure inference inputs and the model (Acar et al., 2018). Developing XAI algorithms for models or input data that are homomorphically encrypted presents significant challenges due to the inherent complexities of this cryptographic technique. Non-linear operations are not natively supported by HE, limiting the types of computations that can be performed in explainability algorithms. Furthermore, HE adds a substantial computational overhead which makes XAI on encrypted data a resource-intensive task. As a result, designing and implementing explainability algorithms is a challenging task. Addressing these challenges is crucial for unlocking the potential for XAI in scenarios where data confidentiality is paramount.

Federated learning (Zhang et al., 2021) is a machine learning approach that allows a model to be trained across decentralized devices or servers holding local data samples without exchanging them. In this collaborative learning paradigm, the model is trained locally on individual devices using local data, and only model updates (not raw data) are shared with a central server or aggregator. The central server aggregates these updates to improve the global model, which is then sent back to the devices. Federated learning enables the development of machine learning models without centralizing sensitive data. HE is utilized in federated learning frameworks to hide model updates from the aggregator or any other party (Phong et al., 2017). **The input data for prediction computations of the use case were protected using HE.**

3 SURVEY: EXPLAINABILITY VS. SECURITY

We conducted a literature survey to determine whether existing XAI approaches also address the dilemma of secure access to data and models and the need for explainability primarily for end users. We concentrated on studies, surveys, state-of-the-art reports, reviews, and books describing general overviews of XAI approaches in public libraries like IEEExplore, Google Scholar, Elsevier, and ACM. Afterward, we studied the publications and checked whether they addressed the topics *security*, *privacy*, *data protection*, or *encryption*. A list and analysis of the reviewed references can be found in Table 1. All publications are from recent years, not because we limited our search but because the XAI topic is still relatively new in research interest.

Six of the 17 reviewed publications address the topic of data and model security. Adadi and Berrada (Adadi and Berrada, 2018) discuss applications in the finance industry where data security and fair lending are to be addressed. Hohman et al. (Hohman et al., 2019) mention the risk of AI

Table 1: Reviewed surveys, state-of-the-art reports, and books on Explainable AI (sorted alphabetically). The first column (*Reference*) shows the reference to the specific paper. In the second column (*Type*) we define the type of the publication (S = survey, R = state-of-the-art report or review, B = book). The third column (*Sec.*) shows whether the topic of data security and the interplay with explainability has been addressed.

Reference	Type	Sec.
(Adadi and Berrada, 2018)	S	X
(Alicioglu and Sun, 2022)	S	–
(Angelov et al., 2021)	R	–
(Barredo Arrieta et al., 2020)	S	X
(Brasse et al., 2023)	R	–
(Chatzimparmpas et al., 2020)	R	X
(Choo and Liu, 2018)	S	–
(Dağlarli, 2020)	S	–
(Danilevsky et al., 2020)	S	–
(Hohman et al., 2019)	S	X
(Holzinger et al., 2020)	S	–
(Liang et al., 2021)	S	–
(Miller, 2019)	S	–
(Saeed and Omlin, 2023)	S	X
(Samek et al., 2019)	B	X
(Xu et al., 2019)	S	–
(Yang et al., 2023)	S	–

models being fooled by inserting wrong data (e.g., images) into the training dataset. Saeed and Omlin (Saeed and Omlin, 2023) note that model creators might regard a model's synthesized knowledge as confidential and that having only model inputs and outputs available might be a compromise for still employing explainability. Together with Barredo Arrieta et al. (Barredo Arrieta et al., 2020), they recommend further research on XAI tools that explain model decisions while maintaining a model's confidentiality (Hohman et al., 2019). In the survey by Samek et al. (Samek et al., 2019), the authors discuss the possibility of security breaches when exposing internal details of model structures to outsiders. They also discuss possible legal implications due to intellectual property rights since internal information about models may be proprietary and a fundamental property right of companies, and the training data may be privacy sensitive.

A notable void exists concerning the comprehensive integration of robust data security measures in the realm of XAI. We noticed that there has yet to be a survey concentrating on the interplay between explainability, security, and privacy. While XAI solutions aim to enhance transparency and interpretability in AI decision-making, the focus has often been directed towards unveiling the intricacies of algorithms rather than fortifying the security of the underlying data. This oversight raises critical concerns as sensitive information becomes increasingly vulnerable to potential breaches or unauthorized access. Addressing this gap becomes imperative for developing responsible and trustworthy AI systems.

4 EXPLAINABILITY USE CASE

Embarking on a practical application of predictive AI and XAI, we describe a practical use case in predictive maintenance and reflect on our learnings.

4.1 Use Case Setup

Our investigation focuses on systematically utilizing HE (Yi et al., 2014) for input privacy in a predictive analytics use case.

Data and Model. In this use case, a machine learning model was used to predict the wear of tools in machine tools to optimize both the service life of tools and product quality: The inference data and the model needed to be secured not to reveal intellectual property rights, both from the model owner (i.e., the machine builder) regarding the model and training data as well as the data owner (i.e., the machine operator) regarding the inference data and underlying process know-how. The inference data is encrypted using not the original data but features computed from the raw data for training the model (referred to as *training features*). This way, it was also possible to let external stakeholders use the model by sending only features (not the raw data) and receiving predictions from the model. The model remains on the owner's side and doesn't require encryption. Encrypted inputs are received by the model owner, who then computes the output in an oblivious manner.

User Task. The model predicts the wear of tools based on previous measurements. Process managers can use the prediction output of the model to determine the current wear condition of the tool and estimate when the tool should be replaced and whether the tool life can be extended or should be shortened.

Explainability. Explainability was especially relevant in this use case because, naturally, high economic interests are behind it. Replacing parts too early only partially utilizes the service life of components, some of which are quite expensive. Replacing components too late can impair the component's quality and lead to unnecessary rejects or damage to other machine components. End users, therefore, need to know how much they can trust the prediction to make the right decisions according to their priorities. We wanted to give users of the prediction model an impression of how confident they can be in the prediction of the model. Since the model cannot provide any information about uncertainties, we decided to focus on the training features for the analysis. We agreed that users should be able to compare their input features with the training features that the model already *knows* to see whether the input features roughly correspond to the situations that the model has already learned. According to Yang et al. (Yang et al., 2023), this describes a source-oriented approach for explainability.

4.2 Explainability Visualization

Similar to existing approaches (Chatzimparmpas et al., 2020), we employed data visualization (Ward et al., 2015) techniques to foster explainability and communicate model outputs to the end users. We evaluated three visualization techniques that we in the following refer to as *low*, *medium*, and *high explainability*. *Low explainability* refers to providing as little information about the underlying data and model as possible, but still providing background information for users to increase explainability. In the *medium explainability* visualization we included more information about the data and the model in the visualization. The *high explainability* visualization provides as much information about the underlying data and model as possible. A higher level of explainability comes with a lower security level. We visualized the input features and training data in an *explainability plot*. In all cases, we used a summary bar to the left of the explainability plot to aggregate the results (i.e., how many input data points match the training data).

Low Explainability. To reach low explainability, we used the training features' total value range and displayed this in the background (Figure 1). When using the low explainability visualization, users could judge how many input features were outside the training features (i.e., the known model space). Based on the colored background, end users could not reveal the distribution of features in the training dataset.

Medium Explainability. For medium explainability, we employed *Mean Square Error* (Dodge, 2008) as a similarity measure and mapped the calculated similarity values to color (Figure 2). We used a categorical color map with three color bins. This way, users can compare the similarities of their input features with the training data features. End users could disclose the distribution of the training features from this visualization, though, when using the visualization for sampling with different input features. After some iterations, end users could get an impression of the training feature distribution. Model owners could prevent this by employing technical barriers, e.g., allowing only a certain number of requests per minute/hour.

High Explainability. The visualization for high explainability employs density (Figure 3). A density plot shows all training data points; the input data points are plotted as dots in the foreground. Users can use this representation to evaluate the distribution of the training features and input features together. The best comparison between input and training features is possible when end users see the actual distribution of training features. In this distribution visualization, outliers and areas of high density are visible. However, this visualization reveals the values of the underlying training features.

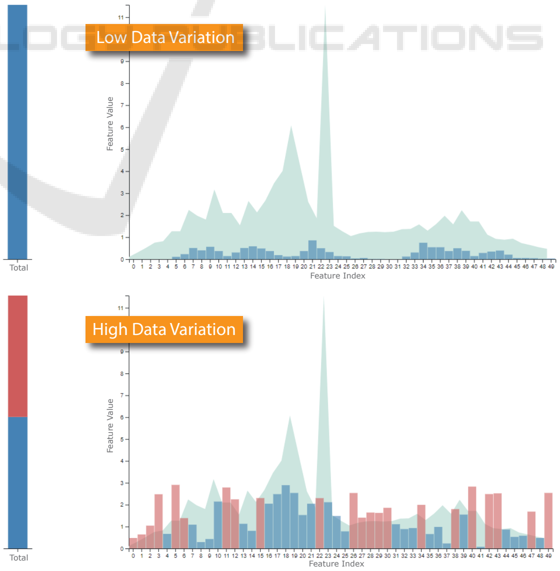

Figure 1: Low Explainability visualization. The training features' total value range and displayed this in the background (light blue). In the foreground, the input features are shown as bars, either dark blue if inside the training feature area or red if outside. The bars' height indicates the input feature's value at this position.

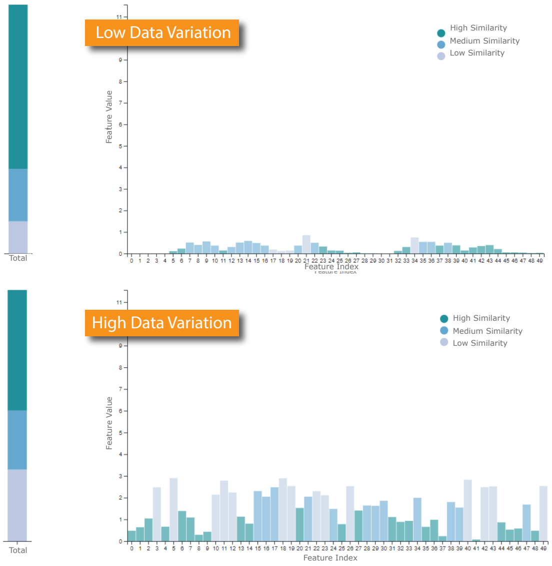

Figure 2: Medium Explainability visualization. We printed the similarity value of each feature onto the bars. This way users can analyze the similarity distribution.

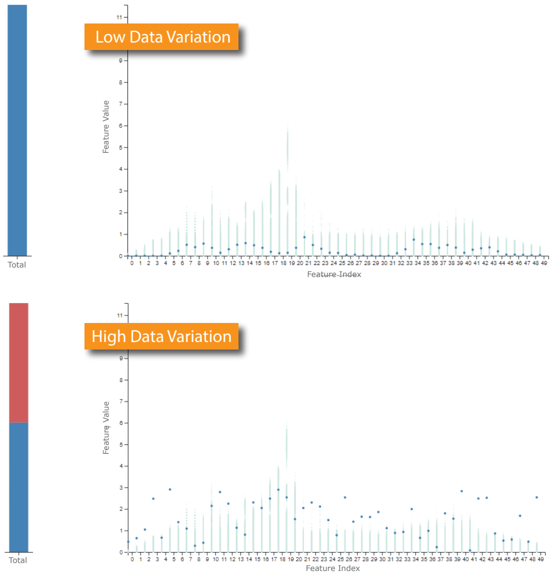

Figure 3: High Explainability visualization. A density plot shows the training data features and the input features are printed as dots in the foreground.

4.3 Learnings

In our use case, we could show that discrepancies between explainability and security still exist. We could also show that the choice of trust measure impacts the results (Papenmeier et al., 2022). When analyzing the visualization results in our use case, we could see that low explainability correlates with reduced security risks. In contrast, high explainability increases security vulnerability and privacy leakage, confirming the known interplay between explainability, security, and privacy in AI (Ogrezeanu et al., 2022).

For our use case, we clearly identified **medium explainability** as the visualization technique that best strives the compromise between explainability and security risk. The similarity measure provides a good overview of how well the input features match the knowledge of the model. Securing the model can be achieved by employing technical barriers.

5 FUTURE WORK

Researchers can contribute to advancing the field of AI, ensuring that future systems are secure, transparent, interpretable, and aligned with ethical principles, and continuing future research and development avenues. Directions for future work may include:

- **Contextual Explainability:** Considering contextual explainability, where the level of interpretability is adapted based on the context and sensitivity of the data involved, would help to protect data while still providing a high level of security. Such approaches could dynamically adjust the trade-off between privacy and interpretability, depending on the specific requirements of different applications.
- **User-Controllable Explainability:** Investigation of systems that allow users to control the level of explainability, e.g., low, medium, or high explainability in our scenario. Users might have varying preferences for transparency, thus, providing a set of explainability options can help balance explainability with privacy.
- **Education and Awareness:** From a user perspective, fostering education and awareness regarding the trade-offs between explainability and privacy would improve the understanding of secure environments. Teaching may involve training AI practitioners, policymakers, and the general public to understand the challenges and potential solutions in navigating the delicate balance between these objectives.
- **Privacy-Preserving Explainability Techniques:** The development and optimization of explainability methods while using privacy-preserving technologies, e.g., HE, multi-party computation, or differential privacy. When input data are encrypted, the predictive AI computation must be performed homomorphically encrypted, therefore also the explainability method, which has not been addressed enough so far.

6 CONCLUSION

This position paper addresses the interplay between explainability, security, and privacy in AI environments. In conclusion, our use case of an explainability visualization dashboard for a predictive maintenance application has illuminated the efficacy of medium explainability, striking a meaningful compromise between providing users with interpretable insights and fortifying the model. These learnings contribute to the ongoing discourse surrounding the delicate balance required in navigating the realms of explainability and privacy in AI, providing practical implications for future implementations and emphasizing the need for tailored approaches in different contexts.

ACKNOWLEDGEMENTS

This work was conducted within the project SMiLe (No. 4109009) funded by the 'ICT of the Future' program by the Federal Ministry of the Republic of Austria for Climate Action, Environment, Energy, Mobility, Innovation and Technology. VRVis is funded by BMK, BMDW, Styria, SFG, Tyrol, and Vienna Business Agency in the scope of COMET (No. 879730), which is managed by FFG.

REFERENCES

Acar, A., Aksu, H., Uluagac, A. S., and Conti, M. (2018). A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys*, 51(4).

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Alicioglu, G. and Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics*, 102:502–520.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5):e1424.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Brasse, J., Broder, H. R., Förster, M., Klier, M., and Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33(1):26.

Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science*, 8:145–161.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., and Kerren, A. (2020). The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, 39(3):713–756.

Choo, J. and Liu, S. (2018). Visual Analytics for Explainable Deep Learning. *IEEE Computer Graphics and Applications*, 38(4):84–92.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, AACL-IJCNLP '20, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Dağlarli, E. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In Aceves-Fernandez, M. A., editor, *Advances and Applications in Deep Learning*, chapter 5. IntechOpen.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. Springer.

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, 3.

Esterle, L. (2022). Deep Learning in Multiagent Systems. In Iosifidis, A. and Tefas, A., editors, *Deep Learning for Robot Perception and Cognition*, pages 435–460. Academic Press.

Gasser, L. and Aad, I. (2023). *Disk, File and Database Encryption*, pages 201–207. Springer Nature Switzerland.

Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. (2019). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Trans. on Visualization and Computer Graphics*, 25(8):2674–2693.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2020). *Explainable AI Methods - A Brief Overview*, pages 13–38. ICML '20. Springer, Vienna, Austria.

Jin, S., Lee, H., Park, C., Chu, H., Tae, Y., Choo, J., and Ko, S. (2023). A Visual Analytics System for Improving Attention-based Traffic Forecasting Models. *IEEE Trans. on Visualization and Computer Graphics*, 29(1):1102–1112.

Kahng, M., Andrews, P. Y., Kalro, A., and Chau, D. H. (2018). ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):88–97.

Lederer, J. (2021). *Fundamentals of High-Dimensional Statistics*. Springer.

Liang, Y., Li, S., Yan, C., Li, M., and Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419:168–182.

Maimon, O. and Rokach, L. (2006). *Data Mining and Knowledge Discovery Handbook*. Springer.

Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1):53–63.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Nocker, M., Drexel, D., Rader, M., Montuoro, A., and Schöttle, P. (2023). HE-MAN – Homomorphically Encrypted MAchine Learning with ONnx Models. In *Proceedings of the 8th International Conference on Machine Learning Technologies*, ICMLT '23, page 35–45, Stockholm, Sweden. ACM.

Nunes, P., Santos, J. P., and Rocha, E. M. (2023). Challenges in predictive maintenance – A review. *CIRP Journal of Manufacturing Science and Technology*, 40:53–67.

Ogrezeanu, I., Vizitiu, A., Ciușdel, C., Puiu, A., Coman, S., Boldișor, C., Itu, A., Demeter, R., Moldoveanu, F., Suciu, C., and Itu, L. (2022). Privacy-Preserving and Explainable AI in Industrial Applications. *Applied Sciences*, 12(13):6395.

Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4):35.

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. (2017). Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345.

Preece, A. (2018). Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72.

Saeed, W. and Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., and Forestiero, A. (2022). A Survey of Artificial Intelligence Challenges: Analyzing the Definitions, Relationships, and Evolutions. *Applied Sciences*, 12(8):4054.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer.

Sapankevych, N. I. and Sankar, R. (2009). Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38.

Schonlau, M. and Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal*, 20(1):3–29.

Sheikh, H., Prins, C., and Schrijvers, E. (2023). *Mission AI - The New System Technology*. Springer Cham.

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2):130–135.

Strobelt, H., Gehrmann, S., Pfister, H., and Rush, A. M. (2018). LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):667–676.

Sukhodolskiy, I. and Zapechnikov, S. (2020). Analysis of Secure Protocols and Authentication Methods for Messaging. *Procedia Computer Science*, 169:407–411.

Timan, T., Mann, Z., Curry, E., Metzger, A., Zillner, S., Pazzaglia, J.-C., and Robles, A. G. (2021). *Data Protection in the Era of Artificial Intelligence: Trends, Existing Solutions and Recommendations for Privacy-Preserving Technologies*, pages 153–175. Springer Int Publishing.

Wang, J., Wang, L., Zheng, Y., Yeh, C.-C. M., Jain, S., and Zhang, W. (2023). Learning-From-Disagreement: A Model Comparison and Visual Analytics Framework. *IEEE Trans. on Visualization and Computer Graphics*, 29(9):3809–3825.

Wang, Y. and Tuerhong, G. (2022). A Survey of Interpretable Machine Learning Methods. In *Proceedings of the International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence*, VRHCIAI '22, pages 232–237, Changsha, China.

Wang, Z. and Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75:796–808.

Ward, M., Grinstein, G. G., and Keim, D. (2015). *Interactive data visualization: Foundations, Techniques, and Applications*. AK Peters.

Xenopoulos, P., Rulff, J., Nonato, L. G., Barr, B., and Silva, C. (2023). Calibrate: Interactive Analysis of Probabilistic Model Output. *IEEE Trans. on Visualization and Computer Graphics*, 29(1):853–863.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, NLPCC '19, pages 563–574, Dunhuang, China. Springer Int Publishing.

Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., and Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, 3:161–188.

Yi, X., Paulet, R., and Bertino, E. (2014). *Homomorphic Encryption*, pages 27–46. Springer Int Publishing.

Zhang, C. and Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23:100224.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. (2021). A Survey on Federated Learning. *Knowledge-Based Systems*, 216:106775.