Explainable Federated Learning: Taxonomy, Evaluation Frameworks, and Emerging Challenges

Rishika Singh
Dept. of Artificial Intelligence & Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra

Swati Joshi
Dept. of Artificial Intelligence & Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra

ABSTRACT
Solutions that guarantee data privacy and model transparency are required due to the quick integration of AI into delicate industries like cybersecurity, healthcare, and finance. Federated Learning (FL) is a promising paradigm that allows for cooperative model training across decentralized datasets while maintaining privacy by avoiding the sharing of raw data. Simultaneously, Explainable AI (XAI) makes otherwise opaque models interpretable, promoting stakeholder trust and assisting with regulatory compliance. Using techniques like SHAP, LIME, Grad-CAM, fuzzy logic, and rule-based systems, recent research has investigated the nexus between FL and XAI in tasks like intrusion detection, fraud detection, and medical diagnosis. Despite the impressive performance of these efforts, there are still unresolved issues with scalability, non- IID data, privacy–interpretability trade-offs, standardized evaluation metrics, and resilience to adversarial manipulation. The present state of research is compiled in this review, which also identifies important gaps, emphasizes methodological trends, and suggests future directions. These issues could be resolved by integrating FL and XAI, which could lead to reliable, private, and interpretable AI systems in high-stakes situations where security and explainability are crucial.

General Terms
Artificial Intelligence, Machine Learning, Data Privacy, Model Interpretability, Trustworthy AI

Keywords
Federated Learning, Explainable AI, Privacy-preserving AI, Post-hoc Explanations, Intrinsic Explainability, Healthcare, FinTech, Cybersecurity.

1. INTRODUCTION
1.1 Motivation
Artificial intelligence (AI) and machine learning (ML) solutions for tasks like fraud detection, credit scoring, risk management, and medical-financial applications have propelled the financial technology (FinTech) industry's recent explosive growth [6,25,29]. However, lack of interpretability and data privacy are two major obstacles that prevent AI from being widely used in delicate industries like healthcare and finance. Federated Learning (FL) protects privacy and complies with regulations like the GDPR by facilitating cooperative model training across several institutions without the need for data sharing [12,16]. In the meantime, Explainable AI (XAI) offers transparency into black-box models, which is crucial for maintaining user confidence in decision-making, regulatory trust, and fairness [14,26].

1.2 Contributions of the Review
Artificial intelligence (AI) and machine learning (ML) solutions for tasks like fraud detection, credit scoring, risk management, and medical-financial applications have propelled the financial technology (FinTech) industry's recent explosive growth [6,25,29]. However, the widespread application of AI in delicate fields like healthcare and finance is constrained by two basic issues: With an emphasis on their applications in FinTech and related fields, this review offers a thorough summary of the body of research on the nexus between explainable AI and federated learning. The following are the primary contributions:

1. Survey of FL+XAI frameworks: Current methodologies in cybersecurity, healthcare, and finance, emphasizing their approaches, aggregation strategies, and explanation tactics are examined [1–5,25,29,33].
2. Finding research gaps: Unresolved issues like the lack of standardized evaluation metrics for interpretability [26,31], the need for intrinsic explainability mechanisms [9,10], and the limited empirical validation of explanations in FL are talked about [7,8].
3. Applying findings from fields like healthcare [4,5,8], imaging [14,15], and fraud detection [6,25,29] to the FinTech landscape, the improvement in regulatory compliance, credit scoring, insurance modeling, and fraud detection using FL+XAI is demonstrated.
4. Future research directions: Areas like standardized evaluation techniques [26,32], reliable FL+XAI frameworks for high-stakes decisions [11,24,31], and counterfactual explanations in federated settings are suggested [18,19].

2. BACKGROUND
2.1 Federated Learning
Finding research gaps: Unresolved issues like the lack of standardized evaluation metrics for interpretability [26,31], the need for intrinsic explainability mechanisms [9,10], and the limited empirical validation of explanations in FL [7,8] are talked about. Google was the first to introduce Federated Learning (FL), a decentralized training paradigm [16]. FL enables institutions to work together to create a global model while retaining raw data on local devices, in contrast to traditional centralized approaches that require data from various sources to be gathered in one place. Only the learned parameters or gradients are transmitted to a central server for aggregation after each client trains a local model using its own private dataset [12,16]. This approach preserves privacy and reduces the risks of data leakage, which is particularly important for domains handling sensitive information such as healthcare, finance, and cybersecurity [5,28,32].

To enhance global model performance in a range of scenarios, several aggregation techniques have been put forth. Federated Averaging (FedAvg), which averages model parameters across clients, is the most popular [16]. While hybrid schemes and rule-based aggregation methods try to handle heterogeneous or non-IID data distributions [9,12], other methods, like weighted

averaging, take client data sizes into account [1,2,5]. Despite its potential, FL still has to deal with real-world issues like effective communication, model convergence for diverse clientele, and striking a balance between participant fairness and global performance [7,12].

2.2 Explainable AI (XAI)

2.2.1 Post-hoc vs intrinsic methods

Intrinsic approaches in Explainable AI (XAI) entail creating models that are transparent by nature, such as decision trees or linear regression, so that their decision-making procedure can be directly comprehended. Post-hoc approaches, on the other hand, use tools such as SHAP or feature importance to analyze a pre-trained "black-box" model after it has been created in order to produce explanations; however, they only offer approximations of the model's actual behavior [21,25].

2.2.2 Model-agnostic vs model-specific.

In machine learning, explainability techniques fall into one of two general categories: model-specific or model-agnostic. Any kind of machine learning model, regardless of its underlying architecture, can be used with model-agnostic techniques. By examining input-output relationships or by approximating the decision boundary, they produce explanations for the model, which they treat as a "black box." LIME and SHAP are two examples that can be used with ensemble models, neural networks, or linear classifiers [21,25]. Although they are widely applicable due to their flexibility, they may only offer approximations and frequently come with additional computational costs.

Conversely, model-specific approaches are customized for specific model classes and use their internal organization to generate explanations. Grad-CAM, for instance, is made especially for convolutional neural networks (CNNs) and highlights significant areas of images using gradient information [17]. In a similar vein, transformer attention mechanisms offer inherent justifications connected to the model's structure [23]. These methods lack the broad applicability of model-agnostic techniques, but they usually provide more accurate and computationally efficient explanations. In conclusion, the choice depends on the use case and model type; model-specific approaches emphasize faithfulness and efficiency, whereas model-agnostic approaches emphasize flexibility.

2.2.3 Types of explanations

Conversely, model-specific approaches are designed to Borys et al. [15] performed a PubMed analysis based on manual classification of all methods into visual and non-visual categories in order to comprehend the current trends in the application of XAI methods in medical imaging. In order to shed light on a model's decision-making process, explainability techniques that rely on visual explanations are widely employed [14].

2.2.3.1 Quantitative explanations

These provide quantifiable indicators of feature relevance. For instance, techniques like feature importance scores or SHAP values give input features weights that show how much each feature influences the model's prediction. For structured or tabular data, where interpretability frequently hinges on knowing the relative importance of features, these explanations are especially helpful. For instance, LIME quantifies the influence of each feature on the model's output and builds interpretable surrogate models around local instances to approximate the model's behavior [21].

2.2.3.2 Visual explanations

These use visual aids to draw attention to specific areas or trends in the input that influence the model's judgment. For example, methods such as Grad-CAM and heatmaps highlight and identify important regions in an image that have the greatest impact on classification [17]. These techniques are particularly helpful in computer vision tasks where interpretability depends heavily on spatial patterns and visual cues.

2.2.3.3 Symbolic explanations

These explain decision-making processes in terms that are easy for humans to understand by using interpretable structures like rules, decision trees, or logic-based models [8,9,19]. Users can follow and analyze the steps that result in a prediction thanks to these explanations, which offer a clear mapping between inputs and outputs

Table 1: Taxonomy of XAI methods

Methods	Post-hoc/ Intrinsic	Type of explanation	Form of explanation	Advantages	Disadvantages
SHAP (SHapley Additive exPlanations)	Post-hoc, model-agnostic	Feature attribution	Quantitative (feature importance scores)	Theoretically grounded, consistent feature attributions, works across models	Computationally expensive, not scalable to very large models, may leak sensitive info in FL
LIME (Local Interpretable Model-agnostic Explanations)	Post-hoc, model-agnostic	Local surrogate models	Quantitative + Symbolic (linear models, rules)	Simple, intuitive, works with any model, provides local explanations	Unstable (different runs may give different results), approximations may be misleading
Grad-CAM (Gradient-weighted Class Activation Mapping)	Post-hoc, CNN-specific	Saliency/heatmap visualization	Visual	Highlights critical image regions, good for CNN interpretability	Limited to CNNs, low resolution heatmaps, not faithful to exact decision logic

NAM (Neural Additive Models)	Intrinsic, interpretable	Neural-based additive feature contributions	Quantitative + Symbolic	More flexible than GAMs, scalable, interpretable feature functions	Still limited compared to black-box deep models, requires careful training
Attention Models (e.g., Transformers, Attention-based FL)	Intrinsic, model-specific	Attention weights as explanations	Quantitative + Visual (attention maps)	Naturally interpretable, integrates into deep models, scalable	Attention ≠ explanation (weights may not always reflect true reasoning), can be misinterprete d

3. NEED FOR EXPLAINABILITY IN AI

Concerns regarding the opacity of complex models have increased as artificial intelligence is being used more widely in delicate fields. Although deep learning has demonstrated cutting-edge performance in tasks ranging from fraud detection to medical diagnosis, its opaque nature raises questions about accountability, trust, and regulatory compliance [25,26,30]. This lack of interpretability restricts the use of AI in high-stakes decision-making, where justifications for predictions are crucial, in addition to impeding user acceptance [11,24].

When implementing AI systems in federated environments, explainability is especially crucial. FL works with distributed, non-IID datasets, in contrast to centralized models, which can result in a variety of local behaviors and intricate global dynamics [7,12]. It becomes challenging to assess how client heterogeneity affects model decisions or to guarantee participant fairness in the absence of strong explanations [8,10]. Additionally, federated models frequently call for the aggregation of local contributions, necessitating the explanation of both local forecasts and the global decision- making process [9,32].

From the standpoint of a review paper, explainability is necessary due to the absence of standardized frameworks as well as technical difficulties. Although post-hoc techniques like Grad-CAM and SHAP are frequently used [17,21,25], little is known about how reliable they are in federated settings. In a similar vein, intrinsic interpretable models exhibit potential [8,9,19], but a thorough assessment of their accuracy, scalability, and privacy trade-offs is necessary. These gaps show how important it is to compile the body of research in order to give a thorough picture of the state of the art, its shortcomings, and its prospects.

4. LITERATURE REVIEW

The application of explainable AI (XAI) and federated learning (FL) in fields like cybersecurity, finance, and healthcare has been the subject of recent research. Numerous studies show how FL can maintain high predictive accuracy in the healthcare industry while protecting privacy. For instance, Briola et al. [1] used SHAP to highlight feature importance in their federated explainable model for breast cancer classification, which achieved over 97% accuracy. Likewise, in their investigation of red blood cell abnormality detection, Dipto et al. [2] found that VGG16 achieved 96% accuracy in centralized training and 94–95% in federated settings, with Grad-CAM offering post- hoc explanations. Other studies extended FL+XAI to time- series data: Mastoi et al. [5] used FL with GoogLeNet for brain tumor classification, improving interpretability through Grad- CAM and saliency maps, while Raza et al. [4] created an ECG monitoring framework that combined federated transfer learning with a modified Grad-CAM to produce interpretable heatmaps.

Some studies have investigated intrinsic explainability in federated environments, going beyond traditional post-hoc methods. Fuzzy rules and SHAP were combined by Ducange et al. [8] to classify Parkinson's disease, yielding results that were easy to understand without appreciably compromising accuracy. Although scalability and robustness are still issues, Bárcena et al. [9] presented LR-XFL, a logic-driven framework that incorporates reasoning rules straight into the FL process. These pieces show the promise of intrinsic approaches, but they also highlight performance and generalization trade-offs. Embedding explainability into federated settings is still experimental and necessitates a careful balancing act between interpretability, accuracy, and privacy, according to other surveys like López-Blanco et al. [10].

Emerging research focuses on FinTech and cybersecurity applications in addition to healthcare. In their federated

framework for malware classification and intrusion detection, Timofte et al. [32] integrated SHAP to interpret predictions and integrated secure communication and differential privacy.

Similar initiatives have been documented in the field of network intrusion detection, where key features that contribute to classification outcomes were highlighted with the aid of SHAP-based explanations [7]. Aljunaid et al. [28] used SHAP and LIME to detect banking fraud in the financial sector with nearly perfect accuracy (~99.9%). Similarly, Sharma et al. [6] used autoencoders and deep learning to detect fraud in credit card transactions, highlighting the importance of decision- making transparency. These studies show that in high-risk financial applications, FL in conjunction with XAI can successfully strike a balance between interpretability and predictive power.

The state of the field is also mapped by a number of recent surveys. While López-Ramos et al. [7] specifically review the intersection of FL and XAI, highlighting issues with privacy, fairness, and heterogeneity, Adadi and Berrada [25] give a general overview of XAI techniques. By classifying techniques into visual and non-visual explanations, Van der Velden et al.

[14] and Borys et al. [15] draw attention to trends in medical imaging. Grad-CAM and heatmaps are the most popular image-based applications. These surveys highlight the variety of approaches, but they also highlight the lack of standard criteria for assessing interpretability across tasks.

When combined, these studies show how flexible FL+XAI is in a variety of high-stakes situations. Due to their ease of integration, post-hoc methods currently dominate

implementations [17,21,25], but intrinsic approaches are becoming more popular [8,9,19]. Furthermore, despite significant advancements, the field still lacks standardized frameworks for evaluation, which makes cross-domain comparison challenging. Another problem is that the majority of current research is proof-of-concept, meaning it has only been tested on small datasets or simulated environments instead of large-scale, real-world deployments.

In conclusion, FL+XAI techniques are still in their infancy even though they have demonstrated promise. Scalability issues, explanation consistency across diverse client distributions, and the establishment of uniform interpretability standards must all be resolved before they can be used in the real world. Table 2 provides a consolidated overview of representative works, their aggregation strategies, XAI methods, performance, and limitations in order to compare contributions in a methodical manner.

Table 2 Literature review

Title	Domain	FL Aggregation	XAI Method	Performance	Gaps / Limitations
A Federated Explainable AI Model for Breast Cancer Classification [1]	Healthcare	FedAvg (Flower)	SHAP	Acc. 97.6%, F1 98.4%	Limited datasets; no scalability analysis
Red Blood Cell Abnormality Detection in Federated Environment [2]	Healthcare	Vanilla & Weighted Avg	Grad-CAM	Acc. 94–96%	Non-IID data impact not addressed
ECG Monitoring with Federated Transfer Learning and XAI [4]	Healthcare	Weighted Avg	Modified Grad-CAM	Acc. 94.5% (noisy), 98.9% (clean data)	Limited to MIT-BIH dataset; scalability
Interpretable FL Model for Brain Tumor Classification [5]	Healthcare	Weighted Avg	Grad-CAM, Saliency Maps	Acc. 94%	No tests on heterogeneous clients
Pediatric Echocardiography with XAI and FL [3]	Healthcare	Not specified	SHAP, Grad-CAM	Case-based study	Early-stage; clinical validation pending
Federated XAI for Parkinson's Disease [8]	Healthcare	FedAvg	Fuzzy Rules + SHAP	High accuracy on case study	Trade-off: interpretability vs. accuracy
LR-XFL: Logical Reasoning-based FL [9]	Cross-domain	Logic-driven rules	Rule-based, intrinsic XAI	Conceptual framework	Scalability, robustness not tested
Federated XAI Review (FED-XAI) [10]	General	Multiple	SHAP, rule-based, hybrid	Survey	Lack of benchmarks, evaluation standards
Federated Learning for Cybersecurity [32]	Cybersecurity	FedAvg + DP	SHAP	Acc. >90%, Privacy loss <5%	Limited to benchmark datasets
Credit Card Fraud Detection using DL + FL [6,28]	Finance	FedAvg	SHAP, LIME	Acc. ~99.9%	Dataset imbalance, limited explainability

While Table 2 provides a comprehensive overview of existing federated learning and explainable AI studies, it remains difficult to discern how specific XAI techniques are distributed across application domains. To provide a clearer comparative perspective, a domain-wise visualization was constructed using only the empirical studies cited in this review. This heatmap aggregates the occurrence of each XAI method—such as SHAP, LIME, Grad-CAM, Saliency Maps, Fuzzy or Logic-based rules, and intrinsic interpretable models—across key domains including healthcare, finance, and cybersecurity. The visualization highlights the concentration of post-hoc techniques in healthcare applications and the limited exploration of intrinsic interpretability in real-world federated settings.

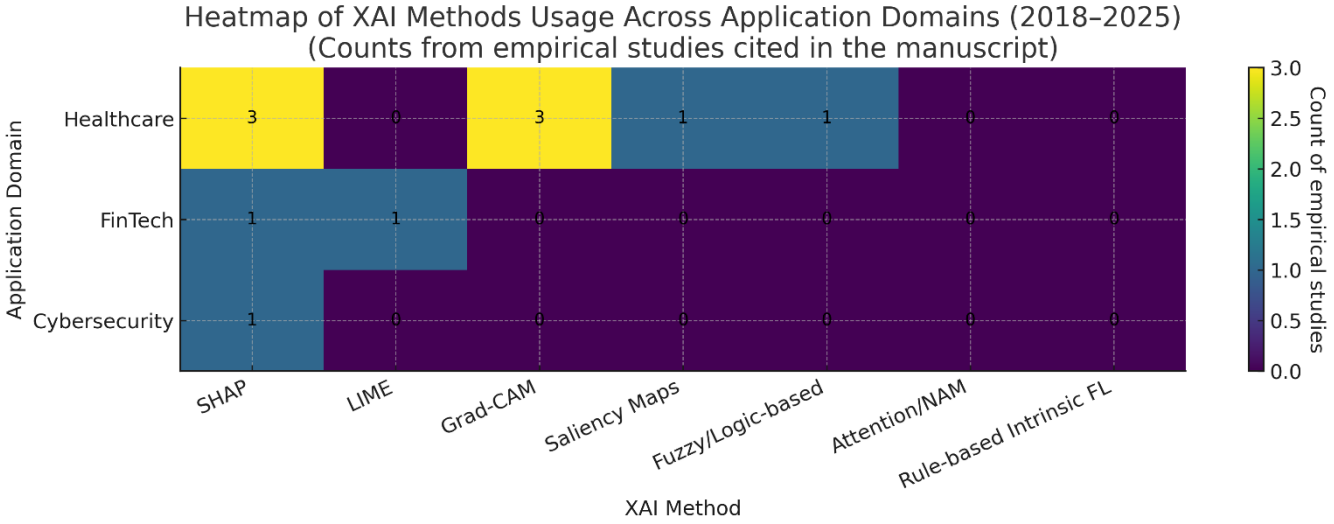

Fig 1: Heatmap of XAI methods usage across application domains (2018-2025)

5. GAPS AND CHALLENGES

There are still a number of research gaps at the nexus of FL and XAI, despite encouraging advancements. First, the majority of previous research is limited to small-scale federations and limited datasets, which limits its applicability to real-world deployments [1, 2, 5, 28, 32]. Heterogeneous, large-scale settings are still poorly understood. Second, both the fidelity of explanations and the accuracy of federated models are still under threat from non-IID data and client heterogeneity [7,10,12]. Third, although the majority of recent research is focused on post-hoc techniques like SHAP, LIME, and Grad- CAM [17,21,25], these techniques only offer approximations of model behavior and may not be consistent across federated clients. Despite their conceptual strength, intrinsic explainability techniques have to compromise on scalability and predictive performance [8,9,19].

The absence of standardized evaluation metrics for interpretability represents another significant gap. Qualitative case studies or visualizations are frequently presented in current works without systematic, quantitative benchmarks [14,15,25,26]. Furthermore, little is known about how FL parameters—like communication rounds, aggregation tactics, and privacy-preserving mechanisms—interact with explanation quality [7,32]. Lastly, there are still unresolved issues regarding explanations that compromise privacy and their susceptibility to hostile manipulation [30, 32].

6. FUTURE SCOPE

In order to compare FL+XAI frameworks fairly across domains, future research should concentrate on creating standardized interpretability metrics [25,26]. It will be essential to develop privacy-preserving explanation strategies that guarantee accuracy without disclosing private information [30, 32]. Furthermore, hybrid approaches that balance accuracy, transparency, and scalability by combining post-hoc flexibility with intrinsic interpretability are required [8,9].

There is also potential in investigating domain-specific adaptations, such as feature-attribution techniques in finance where auditability is necessary for regulatory compliance [6,28] or rule-based or symbolic explanations for clinical interpretability in healthcare [3,4,5]. Likewise, cybersecurity applications require explanations that are resistant to hostile attacks [32]. Another approach is to combine explainability modules and federated frameworks with sophisticated architectures like transformers [23] and graph neural networks [22]. Last but not least, developing extensive benchmark datasets and simulation platforms for federated XAI would promote equitable evaluation of competing approaches, expedite research, and enable reproducibility [7,10].

7. CONCLUSION

In order to develop AI systems that are both interpretable and privacy-preserving, federated learning and explainable AI must come together. Previous studies show encouraging outcomes in cybersecurity [32], healthcare [1–5,8], and finance [6,28], where XAI offers the transparency needed for accountability and trust, and FL permits collaborative training without jeopardizing sensitive data. Current research is dominated by post-hoc methods [17,21,25], but intrinsic approaches are becoming more popular [8,9,19]. Scalability, robustness, evaluation metrics, and privacy–interpretability trade-offs are still issues, though [7,10,12,30].

Although FL+XAI is still in its early stages, the direction of research suggests that it has a lot of promise. Future work can guarantee the deployment of reliable, transparent, and useful federated AI systems in high-stakes domains by filling in the existing gaps through standardized frameworks, hybrid approaches, and extensive evaluations.

8. ACKNOWLEDGEMENTS

I would like to sincerely thank my mentor, Ms. Swati Joshi for her guidance, motivation, and insightful suggestions during the preparation of this paper. I also extend my gratitude to Thakur College of Engineering and Technology for providing the

facilities and resources that supported this work.

9. REFERENCES

[1] Briola E, Nikolaidis CC, Perifanis V, Pavlidis N, Efraimidis P. A federated explainable AI model for breast cancer classification. In: *Proc Eur Interdiscip Cybersecurity Conf*. 2024 Jun 5;194–201.

[2] Dipto SM, Reza MT, Mim NT, Ksibi A, Alsenan S, Uddin J, Samad MA. An analysis of decipherable red blood cell abnormality detection under federated environment leveraging XAI incorporated deep learning. *Sci Rep*. 2024 Oct 27;14(1):25664.

[3] Jabarulla MY, Uden T, Jack T, Beerbaum P, Oeltze-Jafra S. Artificial intelligence in pediatric echocardiography: exploring challenges, opportunities, and clinical applications with explainable AI and federated learning. *arXiv preprint* arXiv:2411.10255. 2024 Nov 15.

[4] Raza A, Tran KP, Koehl L, Li S. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. *Knowl Based Syst*. 2022 Jan 25;236:107763.

[5] Mastoi QU, Latif S, Brohi S, Ahmad J, Alqhatani A, Alshehri MS, Al Mazroa A, Ullah R. Explainable AI in medical imaging: an interpretable and collaborative federated learning model for brain tumor classification. *Front Oncol*. 2025 Feb 27;15:1535478.

[6] Sharma MA, Raj BG, Ramamurthy B, Bhaskar RH. Credit card fraud detection using deep learning based on auto-encoder. In: *ITM Web Conf*. 2022;50:01001.

[7] Lopez-Ramos LM, Leiser F, Rastogi A, Hicks S, Strümke I, Madai VI, Budig T, Sunyaev A, Hilbert A. Interplay between federated learning and explainable artificial intelligence: a scoping review. *arXiv preprint* arXiv:2411.05874. 2024 Nov 7.

[8] Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of XAI models in healthcare: a case study on Parkinson's disease. *Cogn Comput*. 2024 Nov;16(6):3051–76.

[9] Bárcena JL, Daole M, Ducange P, Marcelloni F, Renda A, Ruffini F, Schiavo A. Fed-XAI: Federated learning of explainable artificial intelligence models. In: *Proc XAI.it@AI* 2022;104–117.

[10] López-Blanco R, Alonso RS, González-Arrieta A, Chamoso P, Prieto J. Federated learning of explainable artificial intelligence (FED-XAI): A review. In: *Int Symp Distrib Comput Artif Intell*. Cham: Springer; 2023 Jul 12. p. 318–26.

[11] Hickman E, Petrin M. Trustworthy AI and corporate governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *Eur Bus Organ Law Rev*. 2021 Dec;22(4):593–625.

[12] Silva PR, Vinagre J, Gama J. Towards federated learning: An overview of methods and applications. *WIREs Data Min Knowl Discov*. 2023 Mar;13(2):e1486.

[13] Beutel DJ, Topal T, Mathur A, Qiu X, Fernandez-Marques J, Gao Y, Sani L, Li KH, Parcollet T, de Gusmão PP, Lane ND.Flower: A friendly federated learning framework. *arXiv preprint* arXiv:2007.14390. 2020.

[14] Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal*. 2022 Jul 1;79:102470.

[15] Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F. Explainable AI in medical imaging: An overview for clinical practitioners – beyond saliency-based XAI approaches. *Eur J Radiol*. 2023 May 1;162:110786.

[16] Konečný J, McMahan HB, Ramage D, Richtárik P. Federated optimization: Distributed machine learning for on- device intelligence. *arXiv preprint* arXiv:1610.02527. 2016 Oct 8.

[17] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? *arXiv preprint* arXiv:1611.07450. 2016 Nov 22.

[18] Assaf R, Giurgiu I, Bagehorn F, Schumann A. Mtex-CNN: Multivariate time series explanations for predictions with convolutional neural networks. In: *IEEE Int Conf Data Min (ICDM)*. 2019 Nov 8. p. 952–57.

[19] Deshpande RS, Ambatkar PV. Interpretable deep learning models: Enhancing transparency and trustworthiness in explainable AI. In: *Proc Int Conf Sci Eng*. 2023 Feb;11(1):1352–63.

[20] Kothandaraman D, Praveena N, Varadarajkumar K, Madhav Rao B, Dhabliya D, Satla S, Abera W. Intelligent forecasting of air quality and pollution prediction using machine learning. *Adsorpt Sci Technol*. 2022 Jun 27;2022:5086622.

[21] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. 2016 Aug 13. p. 1135–44.

[22] Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M. Weisfeiler and Leman go neural: Higher- order graph neural networks. In: *Proc AAAI Conf Artif Intell*. 2019 Jul 17;33(1):4602–09.

[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.

[24] Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag*. 2019 Jun 24;40(2):44–58.

[25] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018 Sep 16;6:52138–60.

[26] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: *IEEE 5th Int Conf Data Sci Adv Anal (DSAA)*. 2018 Oct 1. p. 80–89.

[27] Renda A, Ducange P, Marcelloni F, Sabella D, Filippou MC, Nardini G, Stea G, Virdis A, Micheli D, Rapone D, Baltar LG. Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. Information. 2022 Aug 20;13(8):395.

[28] Aljunaid SK, Almheiri SJ, Dawood H, Khan MA. Secure and transparent banking: explainable AI-driven federated learning model for financial fraud detection. J Risk Financ Manag. 2025 Mar 27;18(4):179.

[29] Khan MA, Azhar M, Ibrar K, Alqahtani A, Alsubai S,

Binbusayyis A, Kim YJ, Chang B. COVID-19 classification from chest X-ray images: a framework of deep explainable artificial intelligence. Comput Intell Neurosci. 2022;2022(1):4254631.

[30] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021 Jan 1;113:103655.

[31] Zhang QS, Zhu SC. Visual interpretability for deep learning: a survey. Front Inf Technol Electron Eng. 2018 Jan;19(1):27–39.

[32] Timofte EM, Dimian M, Graur A, Potorac AD, Balan D, Croitoru I, Hrițcan DF, Pușcașu M. Federated learning for cybersecurity: a privacy-preserving approach. Appl Sci. 2025 Jun 18;15(12):6878