UDC 004.8:007.52 DOI https://doi.org/10.35546/kntu2078-4481.2025.1.2.14

O. O. KOROSTIN
Master's Degree, Principal Software Engineer
Shipnext BV
ORCID: 0009-0007-7510-6757

EXPLAINABLE AI: NEW APPROACHES TO INTERPRETABILITY OF DEEP NEURAL NETWORKS

The relevance of this research is determined by the need to enhance the interpretability of deep learning models to ensure transparency and user trust in critical domains such as healthcare, finance, and autonomous systems. Despite the high performance achieved by deep neural networks, their «black-box» nature remains a significant obstacle to their widespread adoption. Increasing regulatory requirements and societal interest in the ethics of artificial intelligence underscore the need to develop explainable AI solutions.

This study aims to analyse current methods of deep neural network interpretability, identify their key limitations, and propose recommendations to improve the efficiency of model applications in real-world settings. A systematic approach was applied, encompassing literature review, comparative analysis of interpretation methods, and evaluation of their effectiveness in practical tasks. The study used theoretical and empirical methods to comprehensively address the issue.

The impact of interpretability on user trust has been examined in critical domains such as healthcare, where AI decision explanations facilitate diagnostic decision-making, and finance, where transparency reduces conflicts between clients and organisations. Model-agnostic approaches (e.g., SHAP, LIME), attention mechanisms, and rule-based explanations were identified as key tools for achieving interpretability. It was found that the main challenges include high computational costs, difficulty in adapting explanations for non-specialists, and risks associated with data confidentiality.

Recommendations were developed, including the integration of hybrid interpretation methods, adaptation of models to specific industry requirements, implementation of interpretability monitoring systems, and the creation of user-friendly explanations. It was demonstrated that such an approach facilitates the effective implementation of deep learning models in practical systems while maintaining their accuracy.

The prospects for further research lie in the development of new interpretation tools tailored to industry-specific needs and the creation of standards for assessing the quality of explanations. This will promote the integration of explainable AI into practical applications and ensure increased user trust in these technologies.

Key words: *transparency model, user trust, attention mechanisms, logical algorithms, ethical considerations, algorithm adaptation, computational challenges.*

О. О. КОРОСТІН
магістр, провідний інженер-програміст
Shipnext BV
ORCID: 0009-0007-7510-6757

EXPLAINABLE AI: НОВІ ПІДХОДИ ДО ІНТЕРПРЕТОВАНОСТІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Актуальність дослідження зумовлена необхідністю підвищення інтерпретованості моделей глибокого навчання для забезпечення прозорості та довіри користувачів у критично важливих галузях, таких як медицина, фінанси та автономні системи. Незважаючи на високі результати, досягнуті за допомогою глибоких нейронних мереж, їхній «чорний ящик» залишається серйозною перешкодою для широкого застосування. Зростання регуляторних вимог і суспільного інтересу до етичності штучного інтелекту підкреслює необхідність розвитку пояснюваного штучного інтелекту.

Метою дослідження є аналіз сучасних методів інтерпретованості глибоких нейронних мереж, визначення їхніх основних обмежень і розробка рекомендацій для підвищення ефективності застосування моделей у реальних умовах. У роботі застосовано системний підхід, який включає аналіз літературних джерел, порівняльний аналіз методів інтерпретації та оцінку їхньої ефективності у практичних задачах. Використано як теоретичні, так і емпіричні методи, що забезпечило всебічне висвітлення проблеми.

Досліджено вплив інтерпретованості на довіру користувачів у таких сферах, як медицина, де пояснення рішень штучного інтелекту сприяє прийняттю діагностичних рішень, та фінанси, де прозорість сприяє зниженню конфліктів між клієнтами й організаціями. Виявлено, що модель-агностичні підходи (SHAP, LIME), механізми уваги та логічні правила є ключовими інструментами забезпечення інтерпретованості. Основними проблемами визначено високі обчислювальні витрати, складність адаптації пояснень до потреб нефахівців і ризики, пов'язані з конфіденційністю даних.

Розроблено рекомендації, які включають інтеграцію гібридних методів інтерпретації, адаптацію моделей до специфіки галузей, впровадження систем моніторингу інтерпретованості та створення зрозумілих пояснень для кінцевих користувачів. Доведено, що такий підхід сприяє ефективному впровадженню моделей глибокого навчання у практичні системи, зберігаючи їхню точність. Перспективи подальших досліджень полягають у розробці нових інструментів інтерпретації, що враховують специфіку галузей, та створенні стандартів для оцінки якості пояснень.

Ключові слова: *прозорість моделей, довіра користувачів, механізми уваги, логічні алгоритми, етичні аспекти, адаптація алгоритмів, обчислювальні виклики.*

Problem statement

Deep neural networks are one of the most powerful technologies in modern artificial intelligence, demonstrating excellent results in image and text processing and forecasting tasks. However, their «black box» poses significant challenges for scientists and practitioners, as the lack of transparency in the decision-making process raises doubts about these systems' reliability and ethical use. This is especially true in industries where mistakes, such as healthcare, finance or automated control systems, can have serious consequences. Uncertainty about the mechanisms of the models limits their implementation in critical areas where it is necessary not only to get the right result but also to explain how a particular decision was made.

Solving this problem is an important scientific and practical task, as the interpretability of artificial intelligence models contributes to the growth of trust in these technologies, ensures compliance with ethical standards and allows users to more consciously evaluate the results of algorithms. Research in this area aims to develop explanatory methods that will help better understand how deep neural networks function and assess the impact of certain factors on decision-making. Thus, increasing the interpretability of models will facilitate their wider implementation and more efficient use in socially important areas.

Analysis of the latest research and publications

The issue of the explainability of deep neural networks remains an important research topic in the context of the growing complexity of artificial intelligence. The work of K. Chyzhmar et al. [1] emphasises that information security requires integrating transparent systems to reduce the risks associated with the uncontrolled use of technology. The results of their analysis indicate the effectiveness of the adaptation of the explained models in the protection of information systems. W. Samek et al. [2] investigated methods of explaining deep networks, particularly heatmaps, which allow for visualising the importance of individual characteristics in decision-making. Their results demonstrate a significant improvement in understanding the process of model operation, which increases user confidence. A. Adadi and M. Berrada [3] focused on the role of Explainable AI in medical research. Their review revealed that the interpretability of models significantly reduces the risk of making erroneous diagnostic decisions by allowing doctors to receive clear explanations for the results of predictions. C. Rudin [4] justified using interpretable models instead of black boxes. Her research proved that such approaches provide greater transparency, especially in high-risk industries such as medicine or finance, where lives or significant resources depend on decisions.

A. B. Arrieta et al. [5] proposed a conceptual classification of XAI that covers the main methods, challenges and opportunities. They found that the key challenge is the balance between models' performance and their solutions' comprehensibility. In turn, E. Tjoa and C. Guan [6] demonstrated how explainable models in the medical field increase the accuracy of diagnostic decisions and trust in artificial intelligence systems among medical personnel. R. Guidotti et al. [7] focused on explanation methods for large amounts of data. Their results show that simplifying algorithms ensures system stability without loss of accuracy. Complementing these conclusions, P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis [8] emphasise that integrating interpreted models into practical applications can significantly improve their effectiveness in real-world conditions.

F. Doshi-Velez and B. Kim [9] developed formal criteria for assessing the interpretability of models, which became the basis for creating standardised approaches in this area. L. H. Gilpin and colleagues [10] considered ways to balance accuracy and transparency, offering techniques that allow even the most complex models to be understood.

R. Saleem et al. [11] studied global explanation techniques that combine local and global approaches. Their results show that combined techniques provide better comprehensibility for users with different levels of technical knowledge. P. Angelov and E. Soares [12] presented an XDNN model demonstrating high performance and adaptability in real-time scenarios.

The work of W. Samek [13] covers the latest advances in the field of explanatory models that contribute to their reliability in large-scale computing systems. Y. Zhang, P. Tiňo, A. Leonardis and K. Tang [14] emphasised the need to develop models that simultaneously provide high transparency and performance, which is critical in big data.

In concluding the review, K. M. Richmond et al. [15] studied the connection between XAI and legal aspects. Their study confirmed that ethical standards and transparency of decisions are crucial for integrating artificial intelligence into legal processes.

Research shows significant progress in creating transparent and responsible artificial intelligence systems. Particular attention is paid to practical results that demonstrate that model interpretability contributes to increased trust, accuracy, and adaptability in areas such as medicine, law, and big data processing.

Despite the achievements in ensuring the interpretability of deep neural networks, the issues of their adaptation to the specifics of tasks and user needs remain unresolved. Existing methods, such as model-agnostic approaches and feature visualisation mechanisms, have not been sufficiently studied in the context of real-world applications, and high computational costs and complexity of integration limit their implementation in practical systems. The problem of adapting explanations to non-specialists also remains relevant, which creates barriers to the widespread use of models.

The impact of interpretability on user confidence remains poorly understood, especially in critical industries such as medicine or finance. Insufficient attention has been paid to developing comprehensive guidelines that balance model accuracy and transparency. This study aims to fill these gaps by analysing current methods, identifying limitations, and developing approaches to improve models' effectiveness and adaptability in practical settings.

Purpose of the article

This article aims to analyse current approaches to ensuring the interpretability of deep neural networks in the context of the development of explanatory artificial intelligence and to identify practical possibilities for their application in critical industries. The study aims to identify effective methods of explaining the operation of models that can increase their transparency, user trust, and compliance with ethical standards.

The following tasks are set to achieve this goal:

1. To analyse modern methods of ensuring the interpretability of deep neural networks, particularly model-agnostic approaches and methods of feature visualisation, and to evaluate their effectiveness in real-world applications.
2. To investigate the impact of AI models' interpretability on user trust in critical areas such as medicine, finance, and autonomous systems, taking into account technical limitations and challenges in their implementation.
3. To develop recommendations for optimising deep learning models, considering the balance between accuracy, transparency and adaptation to the practical needs of users.

Presentation of the main material

Despite their high performance in solving complex problems, deep neural networks remain mostly "black boxes". This means that the decision-making process remains opaque even for developers, which creates trust problems in such models in critical industries. To address this problem, methods of ensuring interpretability are being actively developed. Among the most common approaches are model-agnostic methods that allow analysis of any machine learning algorithms and feature visualisation methods that allow understanding of how the network processes input data (Table 1).

Table 1

Modern methods of ensuring interpretability of deep neural networks and their practical application

Method	How it works	Application in practice
LIME method (Local Interpretable Model-Agnostic Explanations)	Local explanation of the model's operation by creating simplified linear models for individual forecasts	It is used to explain the classification of images, texts and other data, giving weight to each factor
SHAP method (SHapley Additive exPlanations)	Using game theory to determine the contribution of each attribute to the outcome	It is used in financial systems for risk analysis and in medicine to identify key signs of disease
Saliency Maps	Visualise areas of images or parts of text that have the greatest impact on the model's decisions	They are used to diagnose the performance of models in computer vision tasks, including medical images
CAM and Grad-CAM (Class Activation Mapping)	Visualisation of activated model layers to identify significant regions of input data	They are effective in medicine for analysing MRI and CT images, allowing to detect pathologies at early stages

Source: compiled by the author on the basis of [5; 6; 8].

For example, the LIME method allows for the creation of interpretations for individual predictions, which is particularly important in the justice or finance sector, where every decision requires justification. SHAP provides a global explanation by assessing the contribution of each feature, which helps build confidence in predictions in credit scoring systems. Visualisation methods such as heat maps or Grad-CAM allow us to understand which areas of images or words in the model texts are considered most important, which helps to identify possible errors or biases [4]. In medical practice, these approaches are actively used to analyse diagnostic images, such as X-rays or MRI scans, providing doctors with additional information for decision-making. Thus, interpretability is important for implementing deep neural networks in critical industries.

Deep neural networks are increasingly being integrated into critical systems, but the complexity of their architecture makes explaining the decision-making process difficult. Attention mechanisms and rule-based explanations provide fundamental tools to improve the interpretability of such models. Attention mechanisms allow the model to focus on the most important components of the input data, providing an interpretation of their contribution to the final result. At the same time, rules logically explain the results, making them understandable to professionals using these models (Table 2).

Attention mechanisms and rules are actively used in various industries to ensure both high accuracy of models and their transparency. For example, in the field of natural language processing, attention mechanisms help identify keywords or phrases that have the greatest impact on the result. This allows you to create models that take into account important context, for example, in machine translation or sentiment analysis. Rules, on the other hand, provide detailed logical explanations that make the results understandable to non-specialists. In medicine, this helps doctors justify diagnoses,

Table 2

The effectiveness of using attention mechanisms and rules in explaining the decisions of deep neural networks

Method	How it works	Advantages	Application in practice
Mechanisms of attention	Dynamic weighting of data elements to highlight relevant information	Improve the accuracy of models and allow you to interpret their solutions while maintaining performance	They are used in machine translation systems (for example, Google Translate) and visual analysis
Self-Attention	Forming relationships between all elements of the input data to reveal their significant correlations	Ability to process long data sequences and identify global dependencies	Effective in transformer architectures (BERT, GPT) for processing text and other serial data
Multi-Head Attention	Parallel processing of multiple contexts to identify different aspects of data	Improves analysis performance and detail	It is used in real-time text generation and image analysis tasks
Rules	Drawing logical conclusions by formalising model solutions in the form of trees or logical expressions	Ensure transparency and adaptability in specific industries	They are used in medicine to explain diagnostic decisions and in finance to analyse risks
Logical decision trees	A hierarchical structure that explains how each variable affects the final result	Easy to understand for users without a technical background	They are used in credit scoring and risk forecasting
Hybrid systems of attention and rules	Combining attention mechanisms with logical inference to increase transparency of results	A balanced combination of neural network accuracy and rule clarity	They are used in multifactorial medical systems to analyse complex diagnostic cases

Source: compiled by the author on the basis of [3; 4; 7].

and in finance, it helps explain the reasons for refusing to issue loans [13]. Hybrid systems that combine these approaches create a synergy between the high accuracy of the model and the ability to explain complex multifactorial decisions. Thus, the use of these methods contributes not only to increasing the credibility of neural networks, but also to their wider implementation in practical systems. The interpretability of AI models is an important factor in ensuring their effective use in critical areas where user trust plays a key role. In industries such as healthcare, finance, and autonomous systems, the decisions made by models have a significant impact on safety, financial stability, or even human life. Interpretability allows users to understand how and why specific decisions were made, reducing the risk of errors and increasing the transparency of processes. This factor becomes especially important in cases where models play not only an auxiliary but also an autonomous role in decision-making (Table 3).

Table 3

The impact of interpretability of artificial intelligence models on user confidence in critical areas

Field of application	The role of interpretability	Impact on user confidence	Examples of practical use
Healthcare	Explanation of the factors that influenced the diagnostic or prognostic conclusion	It increases the trust of doctors and patients in decision support systems and facilitates their integration into clinical practice	Use in MRI image analysis systems to detect pathologies; explanation of risks in personalised medicine
Financial sector	Justification of decisions made in credit scoring, investment management or risk assessment processes	Allows users and regulators to better understand algorithms and reduces the likelihood of legal and reputational risks	Application in banking systems for analysing the reasons for refusing to lend; risk assessment of investment portfolios
Standalone systems	Explaining the actions and choices of autonomous vehicles in critical situations	Increases trust in autonomous vehicles, especially in cases of accidents or disputes	Integration into autonomous driving systems that explain how to react to unexpected circumstances
Law	Justification of judicial or administrative decisions made by artificial intelligence systems	It ensures transparency and verifiability of decisions, which is critical for maintaining public trust	Use in systems for analysing court precedents and preparing legal documents
Aviation	Explanation of solutions for automatic piloting and aviation safety monitoring systems	Increases pilot and controller confidence in automated systems, reducing the risk of errors	Application in flight control systems to explain trajectory changes or other automatic decisions

Source: compiled by the author on the basis of [2; 6; 13].

In practice, the interpretability of AI models is a crucial tool for ensuring transparency in their functioning, which, in turn, enhances user confidence. For instance, in medicine, explaining how a model arrived at a diagnosis enables doctors to critically evaluate the results and make informed decisions, facilitating the integration of these systems into clinical workflows. In the financial sector, interpretable decisions help customers understand the reasons for loan rejections, reducing tension between customers and banks. In autonomous transport, explaining route choices or vehicle actions in critical situations fosters trust among passengers and regulators [9]. Thus, the development and implementation of interpretability in artificial intelligence systems are essential for their effective application in critical industries.

However, implementing interpretable AI models in real-world environments faces several limitations and challenges that significantly impact their effectiveness. One key issue is finding a balance between accuracy and interpretability. Complex models, such as deep neural networks, achieve high performance on intricate tasks but remain opaque to most users. In contrast, simpler models, such as linear regression or decision trees, offer more comprehensible explanations but may fail to meet the accuracy demands of complex scenarios.

Another significant challenge is the high computational cost associated with model interpretation [7]. For example, methods like LIME or SHAP require substantial resources to generate local explanations or analyze the contributions of individual features. This limitation complicates the deployment of such models in real-time environments, where rapid decision-making is critical. Additionally, the development and implementation of interpretable models demand specialized expertise, which may not always be available within development teams. This technical barrier can hinder the adoption of new technologies in areas where they are most needed.

Legal and ethical considerations are increasingly becoming significant constraints. In regulated industries such as healthcare and finance, the need to comply with transparency and data protection standards places additional pressure on developers [15]. Moreover, excessive interpretability can inadvertently expose sensitive information, posing risks to both individual users and organizations. This creates the challenge of balancing sufficient explanations with the protection of confidentiality.

Another challenge is building user confidence in model explanations. Even high-quality, technically robust explanations may fail to resonate with users if they are overly complex or not adequately tailored to the audience. This issue is particularly pronounced in contexts where users lack technical expertise but require clear and comprehensible explanations to build trust in the technology.

Developing recommendations for optimizing deep learning models to balance accuracy and transparency is crucial for enhancing their practical application. A key area of focus is integrating interpretability methods during the model development stage. For instance, hybrid approaches that combine attention mechanisms with model-agnostic methods (e.g., SHAP, LIME) can preserve high model accuracy while ensuring transparency. This is especially critical in domains where adherence to ethical standards and regulatory requirements is essential.

Another vital recommendation is adapting models to specific usage contexts. In tasks where transparency is paramount, inherently interpretable architectures, such as decision trees or simplified neural networks, should be prioritized. Conversely, for high-precision tasks like pattern recognition or complex process prediction, more intricate models may be employed alongside integrated explanation tools to mitigate their opacity.

Attention should also be directed to the computational resources required for running interpretable models. Employing optimization algorithms, such as dimensionality reduction or model compression, is recommended to lower the cost of interpretation. These methods ensure acceptable processing speeds in real-world applications, including automated real-time data analysis.

Another crucial aspect is interaction with end users. Models should produce explanations that are comprehensible to the target audience, considering their professional background and level of technical expertise. For instance, in the medical field, this might involve visual explanations of diagnostic findings, highlighting key risk factors. In finance, models could generate logical conclusions outlining the primary reasons behind decisions.

Developing user interfaces for interacting with models is another essential component of optimization. Interactive systems that enable users to refine queries or explore alternative scenarios enhance understanding and build trust in the models. Additionally, developers are advised to implement mechanisms for monitoring and evaluating interpretability. These mechanisms should allow for regular assessment of the quality of explanations and their alignment with user requirements.

In conclusion, optimizing deep learning models to balance accuracy and transparency is a multifaceted process that involves technical, ethical, and social considerations. Implementing these recommendations will promote the effective integration of such models into practical systems, ensuring their reliability and acceptability across diverse industries.

Conclusions

The study identified the problem of interpretability in deep neural networks as a significant barrier to their adoption in critical domains such as medicine, finance, and autonomous systems. The findings confirm that model-agnostic approaches, attention mechanisms, and rule-based systems can substantially enhance user confidence and promote transparency in decision-making processes. However, high computational costs, integration complexities, ethical and legal challenges, and insufficient customization of interpretability tools to end-user needs remain key limitations for developers. Notably, existing approaches often address isolated aspects of interpretability while lacking a comprehensive framework that encompasses technical, social, and organizational dimensions.

Based on the analysis, several recommendations for optimizing deep learning models have been proposed. These include employing hybrid approaches that integrate diverse interpretability methods, tailoring models to task-specific requirements, ensuring user-friendly explanations, and implementing monitoring systems for regular evaluation of explanation quality. Additionally, incorporating ethical, legal, and social considerations during model development is essential to align models with established standards and user expectations.

Future research prospects involve designing novel methods to enhance model interpretability that account for application contexts and industry-specific requirements. Special focus should be placed on integrating explanation techniques into real-time systems and developing standardized frameworks for evaluating explanation quality in interdisciplinary research. These advancements will support the development of interpretable artificial intelligence as a vital tool for fostering transparency, trust, and efficiency in practical applications.

Bibliography

1. State information security as a challenge of information and computer technology development / K. Chyzhmar et al. *Journal of Security and Sustainability Issues*. 2020. P. 819–828. URL: https://doi.org/10.9770/jssi.2020.9.3(8) (date of access: 22.01.2025).

2. Samek W., Montavon G., Lapuschkin S., Anders C.J., Müller K.-R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*. 2021. Vol. 109, No. 3. P. 247–278. DOI: https://doi.org/ 10.1109/JPROC.2021.3060483. (date of access: 19.01.2025).

3. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018. Vol. 6. P. 52138–52160. DOI: https://doi.org/ 10.1109/ACCESS.2018.2870052. (date of access: 19.01.2025).

4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. Vol. 1. P. 206-215. DOI: https://doi.org/10.1038/s42256-019-0048-x. (date of access: 19.01.2025).

5. Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020. Vol. 58. P. 82-115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. (date of access: 19.01.2025).

6. Tjoa E., Guan C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*. 2021. Vol. 32, No. 11. P. 4793–4813. DOI: https://doi.org/ 10.1109/TNNLS.2020.3027314. (date of access: 19.01.2025).

7. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*. 2018. Vol. 51, No. 5. P. 1–42. DOI: https://doi.org/10.1145/3236009. (date of access: 19.01.2025).

8. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*. 2020. Vol. 23, No. 1. P. 18. DOI: https://doi.org/10.3390/e23010018. (date of access: 19.01.2025).

9. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017. arXiv:1702.08608. URL: https://arxiv.org/abs/1702.08608. (date of access: 19.01.2025).

10. Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018. P. 80-89. DOI: https://doi.org/ 10.1109/DSAA.2018.00018. (date of access: 19.01.2025).

11. Saleem R., Yuan B., Kurugollu F., Anjum A., Liu L. Explaining Deep Neural Networks: A Survey on the Global Interpretation Methods. *Neurocomputing*. 2022. Vol. 513. P. 165-180. DOI: https://doi.org/10.1016/j.neucom.2022.09.129 (date of access: 19.01.2025).

12. Angelov P., Soares E. Towards Explainable Deep Neural Networks (xDNN). *Neural Networks*. 2020. Vol. 130. P. 185–194. DOI: https://doi.org/10.1016/j.neunet.2020.07.010 (date of access: 19.01.2025).

13. Samek W. Explainable Deep Learning: Concepts, Methods, and New Developments. In: *Explainable Deep Learning AI.* Academic Press. 2023. P. 7-33. DOI: https://doi.org/10.1016/B978-0-32-396098-4.00008-9 (date of access: 19.01.2025).

14. Zhang Y., Tiňo P., Leonardis A., Tang K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2021. Vol. 5, No. 5. P. 726–742. DOI: https://doi.org/10.1109/TETCI.2021.3100641 (date of access: 19.01.2025).

15. Richmond K.M., Muddamsetty S.M., Gammeltoft-Hansen T., et al. Explainable AI and Law: An Evidential Survey. *Digital Society*. 2024. Vol. 3, No. 1. DOI: https://doi.org/10.1007/s44206-023-00081-z (date of access: 19.01.2025).

References

1. Chyzhmar, K., Dniprov, O., Korotiuk, O., Shapoval, R.V., & Sydorenko, O. (2020). State information security as a challenge of information and computer technology development. *Journal of Security and Sustainability Issues, 9*(3), 819–828. https://doi.org/10.9770/jssi.2020.9.3(8)

2. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

3. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

4. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

5. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

6. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

8. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*. https://arxiv.org/abs/1702.08608

10. Gilpin, L.H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

11. 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

12. Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks, 130*, 185–194. https://doi.org/10.1016/j.neunet.2020.07.010

13. Samek, W. (2023). Explainable deep learning: Concepts, methods, and new developments. In *Explainable Deep Learning AI* (pp. 7–33). Academic Press. https://doi.org/10.1016/B978-0-32-396098-4.00008-9

14. Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence, 5*(5), 726–742. https://doi.org/10.1109/TETCI.2021.3100641

15. Richmond, K.M., Muddamsetty, S.M., Gammeltoft-Hansen, T., et al. (2024). Explainable AI and law: An evidential survey. *Digital Society, 3*(1). https://doi.org/10.1007/s44206-023-00081-z