Technoeconomics. 2026, Vol. 5, No. 1 (16). Pp. 32–40.
Техноэкономика. 2026, Том 5, № 1 (16). С. 32–40.

Scientific article
UDC 330.47
DOI: https://doi.org/10.57809/2026.5.1.16.3

HYBRID AI MODELS: A COMBINATION OF CLASSICAL ALGORITHMS AND NEURAL NETWORKS TO ENHANCE INTERPRETABILITY

Saveliy Cherepanov ✉

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

✉ cherepanov.sv@edu.spbstu.ru

Abstract. Modern Deep Learning neural networks demonstrate high accuracy in classification and forecasting tasks, but significant limitations remain in the interpretability of the results. This creates an obstacle to application in critical areas where maximum transparency of decisions is required. In this paper, we propose the use of a hybrid approach that combines both feature extraction methods based on neural networks and classical interpreted machine learning algorithms. In the course of the work, an architecture was developed in which a neural network forms a compact representation of data, and the final decision is made by an interpreted model in the form of a decision tree or logical regression. Experiments have been conducted on open datasets, confirming that the proposed approach allows for increased interpretability while maintaining accuracy comparable to Deep Learning models. The results demonstrate the promise of hybrid architectures for areas requiring transparency and explainability of the results.

Keywords: neural networks, interpretability, hybrid models, data analytics, decision-making

Citation: Cherepanov S. 2026. Hybrid AI models: a combination of classical algorithms and neural networks to enhance interpretability. Technoeconomics 5, 1 (16), 32−40. DOI: https://doi.org/10.57809/2026.5.1.16.3

Научная статья
УДК 330.47
DOI: https://doi.org/10.57809/2026.5.1.16.3

ГИБРИДНЫЕ МОДЕЛИ ИИ: СОЧЕТАНИЕ КЛАССИЧЕСКИХ АЛГОРИТМОВ И НЕЙРОСЕТЕЙ ДЛЯ ПОВЫШЕНИЯ ИНТЕРПРЕТИРУЕМОСТИ

Савелий Черепанов ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия

✉ cherepanov.sv@edu.spbstu.ru

Аннотация. Современные нейронные сети с глубоким обучением (Deep Learning) демонстрируют высокую точность в задачах классификации и прогнозирования, однако остаются существенные ограничения в интерпретируемости результатов. Это создает препятствие к применению в критически важных областях, где требуется максимальная прозрачность принимаемых решений. В данной работе предлагается применение гибридного подхода, сочетающего в себе как методы извлечения признаков на основе нейронных сетей, так и классические интерпретируемые алгоритмы машинного обучения. В ходе работы была разработана архитектура, в которой нейросеть формирует компактное представление данных, а финальное решение принимает интерпретируемая модель в виде дерева решений или логической регрессии. Проведены эксперименты на открытых наборах данных, подтверждающие, что предложенный подход позволяет добиться увеличения интерпретируемости при сохранении точности, сопоставимого с Deep Learning моделями. Результаты демонстрируют перспективность гибридных архитектур для областей, требующих прозрачности и объяснимости результатов.

Ключевые слова: нейронные сети, интерпретируемость, гибридные модели, аналитика данных, принятие решений

Для цитирования: Черепанов С.В. Гибридные модели ИИ: сочетание классических алгоритмов и нейросетей для повышения интерпретируемости // Техноэкономика. 2026. Т. 5, № 1 (16). С. 32–40. DOI: https://doi.org/10.57809/2026.5.1.16.3

Introduction

The growth of computing capabilities and the emergence of large datasets have led to the widespread and almost mandatory use of deep learning networks, which are well suited for solving problems of analysis, forecasting, etc. However, the main problem with using such technologies in their basic form is that the high accuracy of the data obtained is ensured by the low interpretability of the results. This behavior of technologies creates difficulties in their use in critically important government areas such as medicine, economics, law and public administration, where the results obtained must be extremely clear and understandable to a person so that subsequent decisions have a clear justification. The popularization of AI, in addition to the complexity of its use, creates the need to develop a new solution that will meet the needs of not only key stakeholders, but also others where technology is already actively used (Ignatiev and Levina, 2024; Klimentov, 2025; Kutuzova, 2024). One of the promising solutions to the above problem is the creation of hybrid methods that combine the advantages and speed of neural network analysis with classical, explicable algorithms. In such a system, the neural network performs the functions of extracting specified features, and the interpreted model makes a de-

cision based on a more transparent representation (Bouwmans et al., 2019; Vilone et al., 2021). In the course of the work, a hybrid architecture based on the above approach was investigated and proposed.

The goals and objectives of this study are:

1. Analysis of existing approaches to ensuring interpretability in AI and identification of advantages of hybrid architectures.

2. Development of the architecture of a hybrid AI model using a neural network to extract compact features (embeddings), and an interpreted classifier to make a decision.

3. Assessment of the impact of embeddings on key metrics: classification accuracy, complexity and stability of the interpreted model (decision tree), as well as on qualitative interpretability.

4. A description of the limitations identified during the study.

Materials and Methods

To solve the tasks set, the study is structured in several steps. To begin with, a theoretical analysis and synthesis of existing methods of interpreted AI is carried out. And then a hybrid architecture is implemented, consisting of a two-component system: a neural network feature extractor (MLP) and an interpreter classifier (decision tree).

The key methodological idea is to train a neural network to form embeddings that not only preserve the information necessary for classification, but also provide a more stable and simple logical representation for subsequent training of the decision tree (Adadi and Berrada, 2018; Skatova, 2024). A comparative analysis will be conducted between a reference interpreted model trained on the initial features and a hybrid model trained on embeddings obtained from a neural network.

The interpretability of AI can be divided into intrinsic (intrinsic) and resultant (post-hoc). Intrinsic models are logical regression, decision trees, or linear models that have unconditional transparent logic. Post-hoc models, such as LIME or SHAP, do not change the model, but explain its behavior from the outside, but the explanations themselves can only be approximate and do not provide a correct interpretation (Adler et al., 2018; He et al., 2016; Molnar, 2022).

The proposed hybrid approach is based on current trends in the field of explicable artificial intelligence. Unlike post-hoc methods such as LIME and SHAP, which provide only local and approximate explanations for complex models, this study focuses on creating an intrinsically interpretable model. The work is based on the concept formulated by Rudin, who argues that interpretable models should be used for high-risk solutions instead of trying to explain "black boxes" (Rudin, 2019). The proposed architecture echoes the direction described by Al-Shedivat, where neural networks are used for contextual representation of data, but in this work the emphasis is on separation of functions: the neural network is responsible for representation, and the decision tree is responsible for interpreted classification (Al-Shedivat et al., 2020). In addition, the work expands Caruana's ideas about the use of interpreted models in medicine, showing that even simple neural network preprocessing can improve the stability and transparency of final decisions without significant loss of accuracy (Caruana et al., 2015). The approach to feature extraction through a hidden layer of a neural network for subsequent training of simple models is also reflected in Zhang's work on learning compact representations, which emphasizes that reducing dimensionality and eliminating noise increases interpretability (Zhang et al., 2021).

Deep Learning networks are high—performance, but they have a number of limitations - their internal representations are extremely difficult to interpret, they are subject to shifting and unstable decisions, and the process of making these decisions is difficult for the user to explain. These limitations create promising research in the application of hybrid architectures (Lapus-

chkin et al., 2019; Nguyen et al., 2015; Plumerault et al., 2020).

Approaches combining neural networks and symbolic methods have received attention in recent years, among them are (Alonso, 2020; Lakkaraju et al., 2019; Pochetniy, 2025):

1. Rule extraction, which try to extract rules from trained networks;

2. Surrogate models, which train an interpreted model based on neural network predictions.

However, the above-mentioned methods are resultant, not embedded, which also remain insufficiently studied. In these methods, the classical model makes a decision based on the features that the trained neural network selects. Based on this, an experiment was conducted to improve the interpretability of the result using a hybrid architecture.

Experimental data

During the evaluation of the model, a data set such as breast cancer statistics for St. Petersburg and the Leningrad region (biomedical diagnostics) was used. The criterion for choosing a set is the requirement for clarity of the results obtained, the need to ensure high accuracy, as well as the openness of the data.

To conduct the experiment, a model was built with the following settings: a neural network was built based on MLP with hidden layers, one of which is 16 neurons in size, as well as an interpreted model in the form of a decision tree up to 4 levels deep. The criteria for analyzing the results obtained were the following characteristics: interpretability of the final model, estimation accuracy, sample stability, complexity of the decision tree, stability of the tree structure during retraining of the model.

In addition, another goal of the experiment was to test the hypothesis that compact embeddings formed by a trained neural network improve the quality of interpretation while reducing complexity.

For the tests, a set containing 569 samples and 30 valid signs obtained from the analysis of medical images from real practice was used. The target sample variable is 0 for the condition that the tumor is benign and 1 for a malignant tumor. Before the experiment, data was preprocessed, which included normalization of features (StandardScaler), splitting into training and test samples in a 70/30 ratio, as well as elimination of outliers by threshold methods (IQR). As a result of these operations, a balanced data set (class ratio ~ 37/63) was obtained, which increases stability and reduces the risk of model bias.

The hybrid model used in the study consists of a neural network (MLP) As a feature extraction tool, it has the following characteristics: input of 30 features, a hidden layer of 16 neurons, ReLU activation, output by 16−dimensional embedding, Adam optimization, and an epoch limit of 800. The neural network was trained as a regular classifier, then only a hidden layer was used, which extracts structured data. This made it possible to reduce the dimension and eliminate linear and nonlinear dependencies between the initial features. The model also includes an interpretation that, before hybridization, was a classic A tree with a depth of up to 4 levels and a decision tree with a depth of up to 3 levels, trained on embeddings after. Thus, the tree in system B is more depth-limited, but it benefits from better feature factorization.

To assess the quality of the data obtained, metrics such as Accuracy, Precision, Recall, F1-score, ROC-AUC, number of leaves, depth of the tree, number of rules, and stability of the tree were used.

Results and Discussion

You can see the results of the interpretation below and ROC-curves.

Tree A is shown in Figure 1. This is a classic Decision Tree, trained directly on the initial 30 features of the dataset. The depth of the tree was limited to 4 levels. In the experiment, this model serves as a benchmark for "default interpretability" and a reference point for comparing

complexity and accuracy.

Tree B is a decision tree with a depth of up to 3 levels, which is trained not on the initial features, but on 16-dimensional embeddings extracted by a trained neural network (MLP).

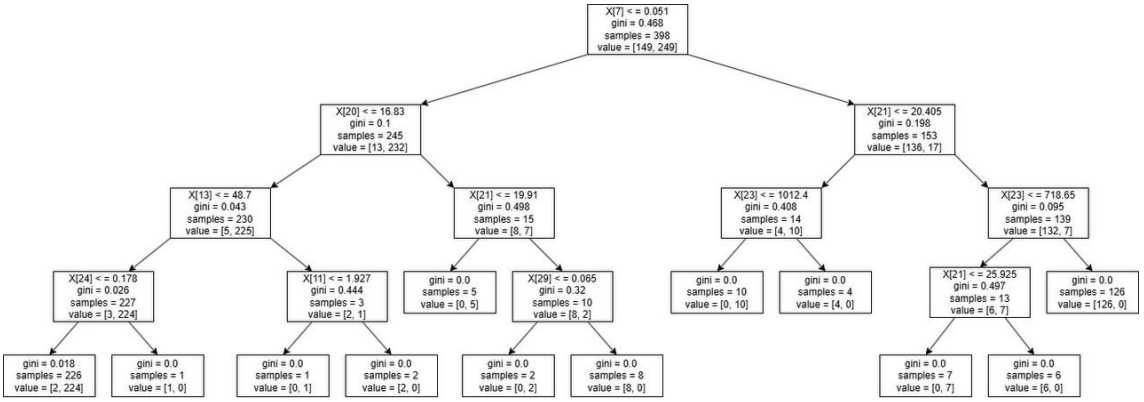


Fig. 1. Tree A.

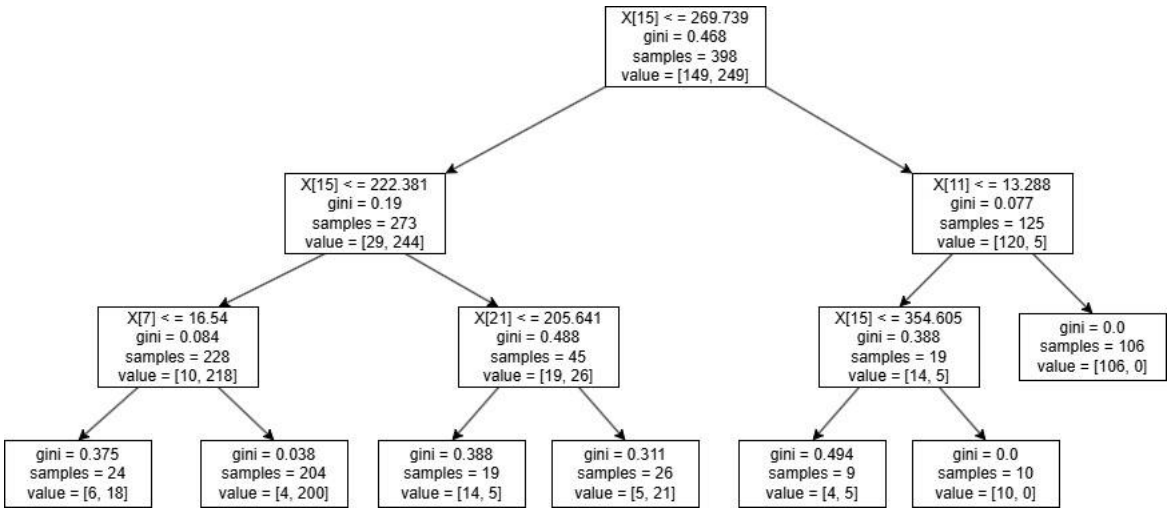


Fig. 2. Tree B.

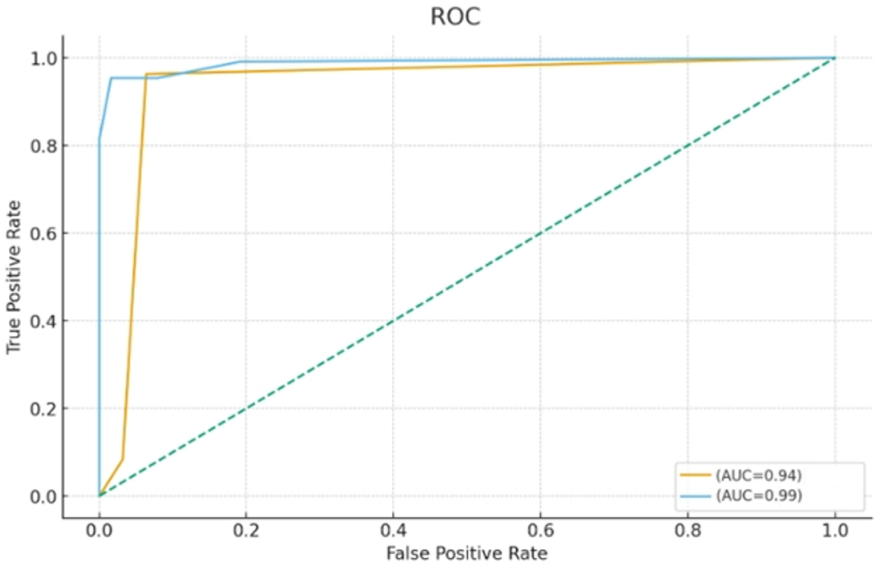
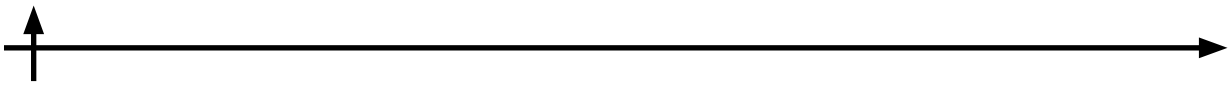


Fig. 3. Graph of ROC curves.



The main results obtained during the experiment are presented below in the form of tables 1, 2, 3, 4 with a brief explanation.

Table 1. Classification quality.

Model	Accuracy	ROC-AUC
Tree A	0,96	0,97
Tree B	0,95	0,96

Based on this, the accuracy of the hybrid model with tree B is inferior to the original one, which reduces the quality of the assessment, but remains at a high and comparable level.

Table 2. The complexity of the tree.

Model	Depth	Number of rules	Number of leaves
Tree A	4	12	7
Tree B	3	6	4

The tree structure of the hybrid model becomes almost twice as short, which increases interpretability and is the main task of the study.

Table 3. Stability during repeated training.

Model	Stability of the structure
Tree A	Low – high variability with the same sample
Tree B	High – minimal changes

The same conditions were set for both models – the same data sample and the number of runs (10 times). Based on this, it can be concluded that embedding reduces the sensitivity of the model structure to noise.

Table 3. Stability during repeated training.

Model A	1. The solution depends on the specific numbers of the initial features 2. The tree relies on combinations of parameters
Model B	1. Solutions are based on more general embeddings 2. Factors consolidate groups of features, highlighting patterns 3. Increased interpretability due to less overloading of rules

It is worth mentioning that the following restrictions were adopted during the study:

1. The analysis was performed on a single dataset, so the accuracy of the system must be confirmed by increasing the number of samples and the data extracted from them;

2. A simple MLP was used, which, although it simplifies the research task, but when using the logic of building on more complex architectures, it is possible to achieve comparatively better embeddings;

3. The interpretation of the resulting embeddings still needs the use of additional methods such as PCA, SHAP, CCA, etc.

Conclusion

In this study, interpreted models and their advantages over post-hoc methods were considered. This made it possible to justify the need for the implementation and subsequent use of interpreted models.

A two-component model was successfully implemented and tested, where a neural network (MLP) performs the role of extracting compact and structured features, and a decision tree acts as a transparent classifier. This confirmed the fundamental possibility of separating the functions of "presentation" and "decision-making".

The obtained results provide clear answers to research questions. The use of embeddings made it possible to reduce the complexity of the decision tree (the depth was reduced from 4 to 3, the number of rules — from 12 to 6), as well as to increase the stability of its structure during repeated training, which indicates a reduction in the influence of noise. The accuracy of the hybrid model (Accuracy 0.95) remained at a level comparable to the baseline (0.96), which confirms the hypothesis of maintaining the quality of forecasting.

Despite the results achieved, limitations have been identified that require further study. The task of interpreting the embeddings themselves remains open and requires the use of additional tools (for example, SHAP or CCA). In addition, it is necessary to validate the proposed approach on a wider range of datasets and using more complex neural network architectures (for example, convolutional or recurrent), which is the vector of future research. Thus, all the research tasks have been solved, and the proposed hybrid model has demonstrated its promise for creating transparent and reliable AI systems.

REFERENCES

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

Adler P., Falk C., Friedler S.A., Nix T., Rybeck G., Scheidegger C., Smith B., Venkatasubramanian S. 2018. Auditing black-box models for indirect influence. Knowledge and Information Systems 54, 95–122.

Alonso J.M. 2020. Teaching Explainable Artificial Intelligence to High School Students. International Journal of Computational Intelligence Systems 13, 974–987.

Al-Shedivat M., Wilson A.G., Saatchi Y., Hu Z., Xing. E.P. 2020. Contextual Explanation Networks. Journal of Machine Learning Research 21, 1–44.

Bouwmans T., Javed S., Sultana M., Jung S.K. 2019. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks 1, 40. DOI: 10.1016/j.neunet.2019.04.024.

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. DOI: 10.1145/2783258.2788613;

He K., Zhang X., Ren S., Sun J. 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. DOI: 10.1109/CVPR.2016.90

Ignatiev P., Levina A. 2024. Artificial intelligence and artificial neural networks in healthcare. Technoeconomics 3, 4 (11), 28–41. DOI: https://doi.org/10.57809/2024.3.4.11.3

Klimentov A. 2025. Predicting claims in auto insurance using deep neural networks. Technoeconomics 4, 4 (15), 36–43. DOI: https://doi.org/10.57809/2025.4.4.15.2

Kutuzova A. 2024. AI-support architecture in digital marketing. Technoeconomics 3, 4 (11), 69–78. DOI: https://doi.org/10.57809/2024.3.4.11.6

Lakkaraju H., Kamar E., Caruana R., Leskovec J. 2019. Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 131–138. DOI: 10.1145/3306618.3314229.

Lapuschkin S., Waldchen S., Binder A., Montavon G., Samek W., Muller K.R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications 10, 1096.

Molnar C. 2022. Interpretable Machine Learning. 2nd ed.

Nguyen A., Yosinski J., Clune J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, 427–436.

Plumerault A., Borgne H.L., Hudelot C. 2020. Controlling generative models with continuous factors of variations. In: Proceedings of the International Conference on Learning Representations, 2020.

Pochetniy V.A. 2025. Integrating generative AI for technological trend analysis and patent research automation. Technoeconomics 4, 2 (13), 4–20. DOI: https://doi.org/10.57809/2025.4.2.13.1

Rudin C. 2019. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 5, 206–215. DOI: 10.1038/s42256-019-0048-x.

Skatova M. 2024. Assessment of requirements of regulatory documents on the use of artificial intelligence in higher education. Technoeconomics 3, 2 (9), 22–33. DOI: https://doi.org/10.57809/2024.3.2.9.2

Vilone G., Longo L. 2021. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. Machine Learning and Knowledge Extraction 3, 3, 615–661. DOI: 10.3390/make3030032.

Zhang Q., Yang Y., Ma H., Wu Y.N. 2021. A Survey on Neural Network Interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence 5, 5, 726–742. DOI: 10.1109/TETCI.2021.3100641.

СПИСОК ИСТОЧНИКОВ

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

Adler P., Falk C., Friedler S.A., Nix T., Rybeck G., Scheidegger C., Smith B., Venkatasubramanian S. 2018. Auditing black-box models for indirect influence. Knowledge and Information Systems 54, 95–122.

Alonso J.M. 2020. Teaching Explainable Artificial Intelligence to High School Students. International Journal of Computational Intelligence Systems 13, 974–987.

Al-Shedivat M., Wilson A.G., Saatchi Y., Hu Z., Xing. E.P. 2020. Contextual Explanation Networks. Journal of Machine Learning Research 21, 1–44.

Bouwmans T., Javed S., Sultana M., Jung S.K. 2019. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks 1, 40. DOI: 10.1016/j.neunet.2019.04.024.

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. DOI: 10.1145/2783258.2788613;

He K., Zhang X., Ren S., Sun J. 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. DOI: 10.1109/CVPR.2016.90

Ignatiev P., Levina A. 2024. Artificial intelligence and artificial neural networks in healthcare. Technoeconomics 3, 4 (11), 28–41. DOI: https://doi.org/10.57809/2024.3.4.11.3

Klimentov A. 2025. Predicting claims in auto insurance using deep neural networks. Technoeconomics 4, 4 (15), 36–43. DOI: https://doi.org/10.57809/2025.4.4.15.2

Kutuzova A. 2024. AI-support architecture in digital marketing. Technoeconomics 3, 4 (11), 69–78. DOI: https://doi.org/10.57809/2024.3.4.11.6

Lakkaraju H., Kamar E., Caruana R., Leskovec J. 2019. Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 131–138. DOI: 10.1145/3306618.3314229.

Lapuschkin S., Waldchen S., Binder A., Montavon G., Samek W., Muller K.R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications 10, 1096.

Molnar C. 2022. Interpretable Machine Learning. 2nd ed.

Nguyen A., Yosinski J., Clune J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, 427–436.

Plumerault A., Borgne H.L., Hudelot C. 2020. Controlling generative models with continuous factors of variations. In: Proceedings of the International Conference on Learning Representations, 2020.

Pochetniy V.A. 2025. Integrating generative AI for technological trend analysis and patent research automation. Technoeconomics 4, 2 (13), 4–20. DOI: https://doi.org/10.57809/2025.4.2.13.1

Rudin C. 2019. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 5, 206–215. DOI: 10.1038/s42256-019-0048-x.

Skatova M. 2024. Assessment of requirements of regulatory documents on the use of artificial intelligence in higher education. Technoeconomics 3, 2 (9), 22–33. DOI: https://doi.org/10.57809/2024.3.2.9.2

Vilone G., Longo L. 2021. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. Machine Learning and Knowledge Extraction 3, 3, 615–661. DOI: 10.3390/make3030032.

Zhang Q., Yang Y., Ma H., Wu Y.N. 2021. A Survey on Neural Network Interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence 5, 5, 726–742. DOI: 10.1109/TETCI.2021.3100641.

INFORMATION ABOUT AUTHOR / ИНФОРМАЦИЯ ОБ АВТОРЕ

CHEREPANOV Saveliy V. – student.
E-mail: cherepanov.sv@edu.spbstu.ru
ЧЕРЕПАНОВ Савелий Васильевич – студент.
E-mail: cherepanov.sv@edu.spbstu.ru

Статья поступила в редакцию 16.12.2025; одобрена после рецензирования 22.02.2026; принята к публикации 15.03.2026.

The article was submitted 16.12.2025; approved after reviewing 22.02.2026; accepted for publication 15.03.2026.