Electronics and Information Technologies, 2025, 32, 87–100
http://publications.lnu.edu.ua/collections/index.php/electronics/index

ELIT

UDC: 004.8,004.9

EXPLAINABLE AND INTERPRETABLE MACHINE LEARNING MODELS FOR ANALYSIS OF OPEN BANKING DATA

Markiyan Fostyak*, *Lidiia Demkiv*
System Design Department
Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine

Fostyak M., Demkiv L. (2025). Explainable and Interpretable Machine Learning Models for Analysis of Open Banking Data. *Electronics and Information Technologies*, 32, 87–100. https://doi.org/10.30970/eli.32.6

ABSTRACT

Background. The development of artificial intelligence and machine learning models has significantly influenced financial analytics and credit decision-making. These models provide high predictive accuracy but often operate as “black boxes,” which complicates the interpretation of their internal mechanisms. In the context of open banking, where decisions directly affect users’ access to financial resources, such opacity is a substantial drawback. This creates a need for explainable and interpretable approaches that make it possible to establish causal relationships between input features and output predictions.

Materials and Methods. The research methods are based on a multi-level approach to ML model interpretation. Feature Importance is applied for a statistical assessment of feature contributions; LIME is used to provide local interpretability; and SHAP (SHapley Additive exPlanations) is employed to capture nonlinear dependencies. Structural interpretability is ensured by DNFS (Deep Neuro-Fuzzy System) through the formation of fuzzy rules, while BRB-ER (Belief Rule Base with Evidential Reasoning) adds logically consistent explanations of decisions based on a rule base.

Results and Discussion. It is shown that, after hyperparameter optimization of credit risk models trained on open banking data, the accuracy of the DNFS model becomes 4 percentage points higher than that of the Random Forest model. A global analysis of feature importance scores obtained using Feature Importance, SHAP, and DNFS demonstrates a high correlation between them (above 88%), indicating model stability. At the local level, instances that reduce model accuracy are identified. Visualizations using SHAP graphs reveal regions of linear and nonlinear feature interactions and their influence on decision-making.

Conclusion. In contrast to the traditional use of individual XAI methods to explain machine learning model outputs, this work combines global and local feature importance metrics (Feature Importance, SHAP, LIME), fuzzy rule–based metrics from DNFS, and aggregated coefficients from BRB-ER. The proposed approach makes it possible to localize the causes of accuracy degradation, identify nonlinear feature dependencies, and assess the consistency of explanations through correlation analysis across methods.

Keywords: explainable artificial intelligence, machine learning, BRB, DNFS, fuzzy logic.

INTRODUCTION

Over the past few years, machine learning (ML) models have been widely used to solve a variety of tasks. However, the limited transparency of their results creates obstacles to practical deployment, especially in domains where the explainability of outcomes is of critical importance. This primarily concerns medicine, financial analytics, and decision support systems, where trust in algorithmic predictions directly depends on the ability to interpret their internal logic. For this reason, there is growing interest in such areas as Explainable Artificial Intelligence (XAI) and the development of interpretable models. Reviews [1–3] present more than ten methods used to construct explainable models in the financial domain. In finance, XAI has the highest priority due to the complex and often opaque structure of ML models used for credit rating assessment, bankruptcy prediction, fraud detection, and credit portfolio optimization.

To explain such models, the following methods are commonly employed: LIME (Local Interpretable Model-agnostic Explanations), which builds local linear approximations to explain the behavior of complex models in the vicinity of an individual prediction; SHAP (SHapley Additive exPlanations), a method based on Shapley value theory that decomposes the model output into contributions of each feature according to its marginal importance; TreeSHAP or TreeExplainer, a modification of SHAP optimized for tree-based models (Random Forest, XGBoost), which provides exact analytical values without the need for simulations; PDP (Partial Dependence Plot), a graphical tool that visualizes the average effect of one or several features on the model prediction; Anchors (High-Precision Model-Agnostic Explanations), a technique that constructs interpretable “anchor” rules providing high-precision local explanations; Counterfactual Explanations (CF), which identify minimal changes in the input data that would lead to an alternative model output; Permutation Feature Importance (PFI), a simple global metric that measures the drop in model accuracy after random permutation of a specific feature; and Surrogate Models (SM), which approximate the behavior of a complex model with an interpretable structure, such as a decision tree or linear regression, among others.

The theoretical foundations and prospects of Explainable Artificial Intelligence (XAI) are discussed in [4,5]. The practical application of XAI in various types of financial models is examined in [6–9]. All these studies confirm that integrating XAI approaches into financial analytics increases trust in automated systems and helps maintain a balance between model accuracy and interpretability. In [6], the SHAP method is used to explain the results of a LightGBM model in a credit scoring task, which made it possible to identify key factors influencing loan approval decisions. In [7], XGBoost, LightGBM, and Random Forest algorithms are applied for predictive analysis of loan default risk in combination with XAI methods SHAP and LIME. Study [8] presents a method for credit card default prediction based on a combination of deep learning and explainable artificial intelligence (XAI) techniques.

Another direction in the development of explainable artificial intelligence is hybrid interpretable approaches. One such approach is the Belief Rule Base with Evidential Reasoning (BRB-ER). BRB-ER ensures transparency and interpretability of decision-making, since each prediction is formed based on an intelligible system of logical rules. In [9], the hierarchical BRB structure incorporates both factual and heuristic rules and can explain the chain of events leading to a decision on a loan application. In addition, performance indicators for evaluating the rule base are proposed, including quantitative, qualitative, and visual metrics in the form of a rule interaction graph [9,10]. Another line of development for interpretable models is Deep Neuro-Fuzzy Systems (DNFS). In DNFS, the hidden layers implement the fuzzification process, creating fuzzy sets for input features, after which the system automatically generates and tunes “if–then” rules that reflect causal relationships in the data [11,12]. Unlike BRB, DNFS does not require prior specification of expert rules, but instead learns them during training, which provides data-driven adaptive interpretability. The importance of each feature can be computed based on the activation strength of the corresponding rules or on the weights of neural connections.

This study aims to compare the effectiveness of interpreting ML models using BRB-ER and DNFS approaches with XAI methods, in particular SHAP and LIME, in terms of their ability to determine feature importance, and to define a consistent approach to constructing explainable and interpretable models for financial analysis in open banking systems.

MATERIALS AND METHODS

Explaining and interpreting the results of machine learning models, as well as analyzing the characteristics of the data on which these models are trained, requires a combination of classical XAI methods, hybrid neuro-fuzzy systems, and rule-based logical approaches. To this end, this study implements an integration of global and local XAI techniques with hybrid approaches and fuzzy logic, which makes it possible to analyze both the overall patterns captured by the model and individual decisions for specific instances.

The SHAP (SHapley Additive exPlanations) method quantifies how important each input variable is for the model. SHAP values are derived from Shapley values in cooperative game theory, where each feature is treated as a “player” contributing to the collective outcome, i.e., the model prediction. The key idea behind SHAP is that the Shapley value is expressed as an additive feature attribution method via a linear explanation model.

In the Python implementation of SHAP, a specialized algorithm called TreeSHAP is provided, which is optimized for tree-based ML models, in particular Decision Trees, Random Forests, and Gradient Boosted Trees (XGBoost, LightGBM, CatBoost). TreeSHAP is a computationally efficient variant of SHAP that does not compute Shapley values via exhaustive enumeration of all feature subsets, but instead relies on an analytical analysis of the tree structure [13]. The algorithm exploits the fact that a tree prediction is a deterministic function of the split conditions at its internal nodes. By analyzing all possible decision paths, it is therefore possible to determine exactly which features actually influence the model output. When SHAP values are used to explain real-world outcomes, it must be recognized that SHAP only shows what the model does in the context of the data on which it was trained. The method does not necessarily uncover the true causal relationships between variables and outcomes in the real world.

The LIME (Local Interpretable Model-agnostic Explanations) method implements the concept of local interpretability by constructing a simplified local regression model (typically linear) that approximates the predictions of the base model in the immediate neighborhood of a selected instance for a subset of points close to the observation of interest. To construct such an explanation, LIME generates a synthetic local dataset by randomly sampling points around the instance, and then fits a linear regression model whose coefficients are used as local feature attributions. The effectiveness of LIME depends strongly on the choice of its parameters: the kernel width defining the locality, the number of generated neighboring points, and the regularization parameter controlling the number of features in the local model [14]. At the same time, the fidelity of the explanation depends on the adequacy of the local linear approximation: if the global model exhibits strong nonlinearity or complex feature interactions, the fitted linear model may capture only a very small, highly local fragment of its behavior.

A separate line of XAI research concerns the construction of hybrid solutions in which interpretable or explainable models are combined with optimization algorithms. In hybrid systems of the type ML + BRB + Optimizer, the ML component is responsible for predictive accuracy, the BRB component ensures transparency via a rule base, and the optimization algorithm adjusts the structure and parameters of the rules to align them with the outputs of the “black-box” model [15]. Such an approach is used, for example, in credit scoring, where predictive accuracy must be combined with the ability to explain the result to the end user and to regulators. In [16], we described in detail a hybrid decision-support system that combines a Belief Rule Base (BRB) model, a machine learning (ML) model, and Particle Swarm Optimization (PSO). At the first stage, reference values of the input features are defined to represent the linguistic states of the parameters. At the second stage, based on these reference values, a BRB rule base is constructed, where each rule encodes the relationship between a combination of input attribute states and the corresponding class. Each rule is characterized by a rule weight and attribute weights. The Evidential Reasoning (ER) mechanism is used to aggregate the activated rules into a final output. At the third stage, the machine learning model (ML) is trained. At the final stage, Particle Swarm Optimization (PSO) is applied to automatically tune the parameters. Analyzing the parameters learned makes it possible to identify which rules or attributes have the greatest influence on the result. Consequently, the BRB–ML–PSO system produces not only a numerical output, but also a logical interpretation of this output via the rule base.

Deep Neuro-Fuzzy System (DNFS) is a hybrid approach that combines fuzzy rules (fuzzy logic) with deep neural networks (deep learning). Unlike BRB, DNFS does not explicitly store all possible rules. Instead, a neural network automatically learns the relevance of rules, prunes insignificant combinations (through the weights), and generalizes, so that the effective number of rules does not grow exponentially.

Structurally, DNFS consists of several interconnected layers [17]. At the first level, each feature passes through a separate fuzzification layer, where it is transformed into fuzzy terms representing the linguistic states of the variable. This is implemented via sigmoidal activation functions. As a result, fuzzy features are formed that represent fuzzy intervals of the values of each input variable. The generated fuzzy features are concatenated into a vector in a fuzzy feature space, which is fed into the rule layer, a layer of neurons. Each neuron corresponds to a single fuzzy IF–THEN rule, and its weights determine the importance of individual terms in the rule combination. During training, the parameters of the fuzzy terms and the rule weights are optimized using a gradient-based method (the Adam optimizer). After computing rule activation degrees, they are normalized using the Softmax function, which transforms them into a distribution of influence weights over the rules. The final output layer aggregates these normalized rule contributions and, in the case of classification, again applies Softmax to produce a probabilistic class prediction.

As a result, we obtain an interpretable output in the form of rules that are generated automatically based on the network’s learned parameters, as well as rule activation degrees and normalized rule weights. Feature importance in DNFS is computed from the aggregate influence of the weights connecting fuzzy terms to the rule layer. Since each feature is represented by several terms, its integral importance is defined as the sum of the absolute weights linking its terms to all rules. This approach makes it possible to quantify the contribution of each feature to the system’s decisions and to identify which rules determined the outcome and why. In this way, the learned parameters can be used for explainability (XAI), analogous to SHAP or BRB. Thus, the DNFS architecture combines the high predictive accuracy of a neural model with the interpretability of fuzzy logic, which makes it effective for financial and medical applications where not only predictive accuracy but also an explanation of the underlying reasons is crucial.

With the development of financial technologies (FinTech) and the digitalization of banking services, the role of ML models in decision-making related to lending, risk assessment, and forecasting clients’ financial behavior has increased significantly. One of the modern approaches to building credit rating models is the use of open banking API data, which contains detailed information on client transactions. At the data preparation stage, the raw open banking API data are preprocessed and transformed into features such as income stability (SI), mandatory spending (MS), non-mandatory (discretionary) spending (DS), risky spending (RS), and account balance [18,19]. To reduce the total number of BRB rules, further categorization of financial features is performed, which we described in detail in [16]. As a result, four features are obtained and expressed in relative units: stability of income (SI), stability of discretionary spending (SDS = 1 − SD), stability of risky spending (SRS = 1 − RS), and the difference between income and expenditure (Diff)

Achieving high classification accuracy on open banking data is a challenging task due to the nature of transactional data itself [20], which is characterized by high variability and low correlation between individual features and the target variable (Fail). Here, Fail = 1 corresponds to the highest credit risk level. Analysis of the correlation matrix shows that the maximum correlation of 0.52 is observed between Diff and Fail, while all other correlation values between variables are below 0.40.

All experiments were implemented in Python 3.8.6. The following libraries were used: NumPy [1.24.3], pandas [1.5.3], matplotlib [3.7.5], scikit-learn [1.3.2], SHAP [0.44.1], LIME [0.2.0], and TensorFlow/Keras [2.13.0]. In addition to using these established packages for model training and explainability, custom Python scripts were implemented for the core study workflow, including: data loading from CSV and preprocessing (feature extraction/ordering for SI, SDS, SRS, DiFF and label handling), training and evaluation of a Random Forest classifier (train/test split, confusion matrix, accuracy, classification report, ROC-AUC, and partial dependence plots), generation of explanation artifacts (SHAP global summaries and dependence plots, SHAP waterfall plots for individual instances, and LIME local surrogate explanations), implementation of a Belief Rule Base inference engine comprising triangular fuzzification (Low/Medium/High on [0,1]), a complete rule base with belief degrees and optional rule weights, rule activation and belief aggregation using weighted averaging and an Evidential Reasoning style combination, and development of a DNFS (Deep Neuro-Fuzzy System) model in TensorFlow/Keras, including fuzzification layers, a rule layer, softmax-based normalization, and rule extraction from learned weights for interpretability.

RESULTS AND DISCUSSION

Table 1 presents the classification accuracy of the Random Forest (RF) and Deep Neuro-Fuzzy System (DNFS) models before and after hyperparameter optimization. The analysis is performed on open banking data containing four key financial parameters of clients (SD, SDS, SRS, Diff) and a binary target variable (Fail) that characterizes the level of credit risk. Among the ensemble machine learning models considered — Gradient Boosting (XGBClassifier), AdaBoostClassifier, and CatBoostClassifier — the Random Forest model demonstrated the highest classification accuracy.

Particle Swarm Optimization (PSO) was used to determine the optimal hyperparameter values for both RF and DNFS. For RF, the following hyperparameters were optimized: the number of trees (n_estimators), maximum tree depth (max_depth), minimum number of samples required to split an internal node (min_samples_split), and minimum number of samples at a leaf node (min_samples_leaf). For DNFS, the optimized parameters included the number of terms per feature (terms_per_feature), the number of rules (number_of_rules), the learning rate, and the dropout rate.

Table 1. **Accuracy of Random Forest (RF) and Deep Neuro-Fuzzy System (DNFS) models before and after the optimization procedure**

	RF	RF (opt)	DNFS	DNFS (opt)
Accuracy	0.84	0.88	0.83	0.91

As can be seen from **Table 1**, both models exhibit an increase in accuracy after optimization. The DNFS model shows a more pronounced improvement in accuracy, which can be attributed to the structural alignment of its rules with the data.

Global explanation

Table 2 presents the feature importance coefficients for the variables SD, SDS, SRS, and Diff, computed using three different approaches (Feature Importance, SHAP, DNFS). The Feature Importance coefficients capture the global statistical contribution of each feature to classification accuracy over the entire dataset. They are obtained from the

Random Forest model using the feature_importances_ attribute, which calculates the mean decrease in impurity (Gini importance) across the decision trees.

Global SHAP coefficients are derived by aggregating local SHAP values for each data instance using shap.Explainer(model), which makes it possible to estimate the average effect of each feature on the model prediction. Global feature importance in DNFS is determined based on the weights of fuzzy terms in the learned rules. Example rules for different classes have the following form:

- Rule 4: IF Diff_Low (w = −0.69) AND SI_High (w = −0.54) AND SD_High (w = −0.45) AND SI_Low (w = 0.35) AND SR_High (w = 0.31) AND SR_Low (w = 0.24) THEN → Class 1;
- Rule 5: IF Diff_High (w = 0.56) AND Diff_Low (w = 0.47) AND SI_Low (w = 0.42) AND SI_High (w = 0.40) AND SD_High (w = 0.39) AND SD_Low (w = −0.39) THEN → Class 0.

For each rule, the absolute values of the weights of the terms associated with a given feature were summed, and these sums were then aggregated across the entire rule set. Subsequent normalization of the resulting values made it possible to obtain an integral rule-based estimate of the importance of each attribute. The resulting aggregate coefficients reflect the degree of influence of each feature on the final model decision.

The DNFS architecture consists of four sequential layers: the input layer, the fuzzy layer (12 neurons, sigmoid), the rule layer (45 neurons, sigmoid), and the decision layer (2 neurons, softmax). The results show that the obtained DNFS-based importance coefficients are highly correlated with the global SHAP values, which confirms the methodological consistency of the different levels of explanation.

A three-component Pearson correlation matrix was computed for the feature importance coefficients. The high correlation between the results of all three approaches (greater than 0.88) indicates that the overall pattern of how influence is distributed across features remains consistent.

Table 2. **Global feature importance coefficients**

	Feature importance	SHAP	DNFS
Diff	0.56	0.36	0.37
SRS	0.19	0.08	0.23
SDS	0.17	0.05	0.21
SI	0.06	0.02	0.19

For a more detailed interpretation of the global impact of the features, a beeswarm plot (**Fig. 1**) was constructed to visualize the distribution of local SHAP values for each feature. The Y-axis lists the features in descending order of their SHAP-based importance, while the X-axis shows the SHAP values, which reflect the direction and magnitude of their influence on the probability of the class Fail = 1. Each point corresponds to an individual data instance, and its color encodes the actual value of the feature. The plot clearly shows that low values of all variables (except for SDS) increase credit risk in a more linear manner and much more rapidly than high values of these variables. For high values of all features, the interactions are more strongly nonlinear than for low values.

The feature SDS exhibits a strongly nonlinear effect across its entire range of values. The atypical influence of SDS (discretionary spending) on credit risk at low spending levels is associated with the borrower’s limited ability to redirect these expenditures towards loan repayment.

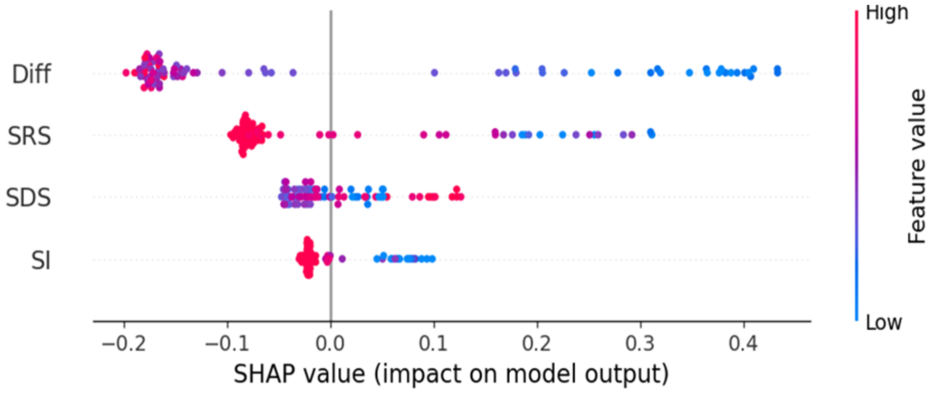

Fig. 1. Distribution of local SHAP values for each feature.

To further assess the global behavior of the model, Partial Dependence Plots (PDPs) (**Fig. 2**) were constructed for the four input features. PDPs make it possible to visually examine how changes in the value of a single feature affect the model's prediction while holding the remaining variables at their average levels. The X-axis represents the values of the corresponding feature, and the Y-axis shows the mean predicted value of the target variable, averaged over all observations.

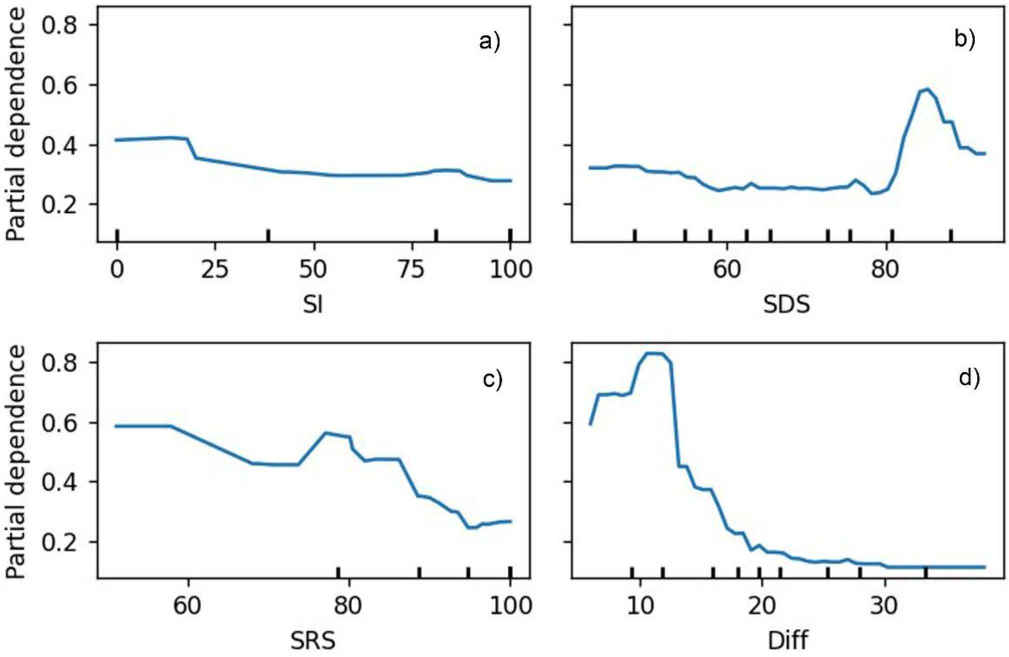

Fig. 2. Partial Dependence Plot: a – Partial Dependence–Si, b – Partial Dependence–SDS, c – Partial Dependence–SRS, d – Partial Dependence–Diff.

As income stability (SI) and stability of risky spending risky spending (SRS) increases, an almost smooth reduction in credit risk is observed. For SDS, two distinct regions of small and large values (i.e., very low and very high discretionary spending) can be seen, both of which are associated with an increase in credit risk. For the difference between income and expenditure (**Fig. 2d**), two distinct dependence regions are observed. For Diff > 12, credit risk decreases in an almost exponential manner up to a point beyond which further increases in the income–expenditure gap no longer improve the risk level. This behavior reflects a zone of financial stability, beyond which additional resources do not increase the probability of loan approval.

For Diff < 12, a threshold-type relationship between Diff and credit risk is observed, namely, a decrease in risk as the difference decreases. Such atypical behavior may be driven by various underlying factors, which require more detailed investigation at the local level.

Local explanation

To conduct a detailed analysis of the local, instance-specific effects of features on the model's predictions, the Individual Conditional Expectation (ICE) method was employed. In contrast to global dependencies (**Fig. 2**), the ICE plot in **Fig. 3** depicts the trajectory of the model's predicted value for a single observation when only one selected feature is varied, while all other characteristics are held fixed. One of the characteristics of the average ICE curve is its mean slope. The largest mean slope, 0.023, corresponds to the feature Diff (**Fig. 3d**).; identical mean slopes of 0.019 are observed for the SI (**Fig. 3a**) and SRS (**Fig. 3c**) features, and the smallest mean slope, 0.010, is obtained for the SDS feature (**Fig. 3b**).

The application of ICE in the context of credit risk assessment made it possible to detect local nonlinearities that are not apparent in the PDPs (**Fig. 2**) or in the average curve (**Fig. 3**). The presence of an S-shaped structure in the ICE plot for Diff indicates that even small increases or decreases in Diff substantially change the level of credit risk. This behavior suggests that the model responds to this feature in a non-linear, threshold-like manner.

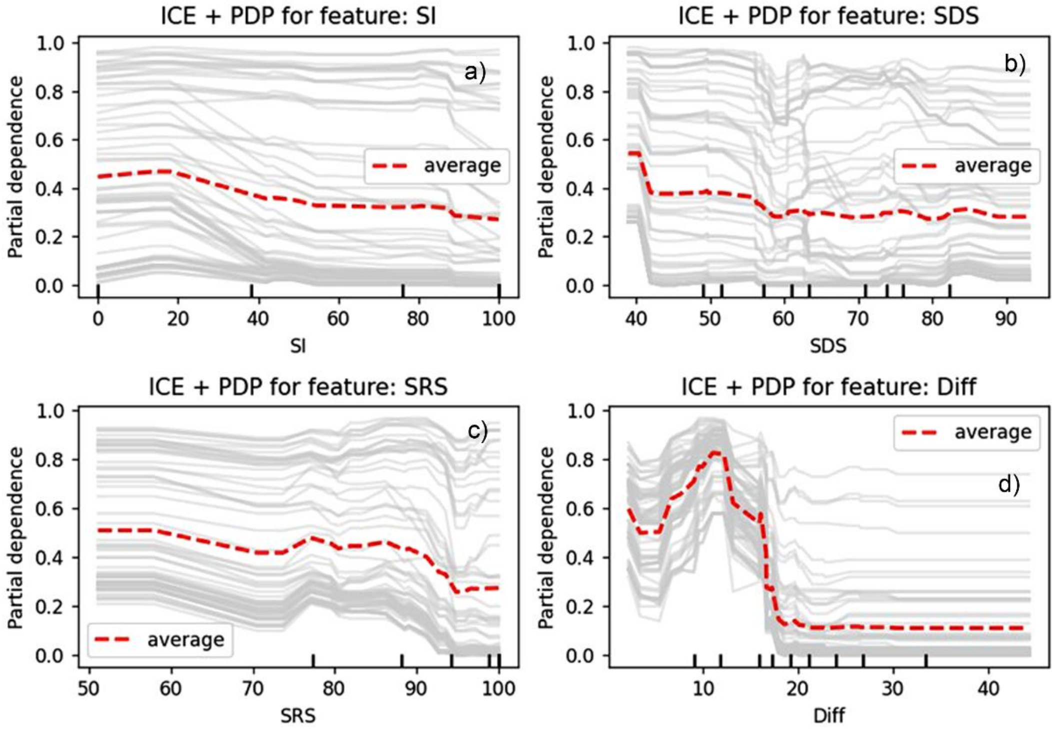

Fig. 3. Individual Conditional Expectation (ICE): a – SI, b – SDS, c – SRS, d – Diff.

Analysis of the ICE curves for SDS revealed two distinct types of client behavior in the region of medium SDS values. For the first group, an upward kink in the curve is observed, where even a slight increase in SDS immediately amplifies credit risk. In contrast, the second group exhibits a downward kink, where a small increase in spending instability improves the prediction, which may indicate the activation of positive behavioral patterns. As a result of averaging, the peak that is clearly visible on the corresponding SDS curve in the PDP plot (**Fig. 2b**) disappears from the average ICE curve. This behavior highlights the heterogeneity of borrowers' spending strategies and the nonlinear nature of the impact of SDS on credit risk. This relationship can be seen in more detail in the SHAP dependence plot in **Fig. 4**.

The SHAP dependence plot illustrates the relationship between the actual value of a feature and its local contribution to the model's prediction, represented through SHAP coefficients (**Fig. 4**). The X-axis shows the feature values, while the Y-axis shows the corresponding SHAP values, which indicate the extent to which this feature increases or decreases the predicted credit risk. The color of the points encodes the value of a second interacting feature, allowing nonlinear and context-dependent interactions between factors to be visualized. SHAP dependence plots were obtained and analyzed for all pairs of features. **Fig. 4** presents the dependence on Diff, as this feature exhibits the strongest nonlinear interactions with other variables.

The nonlinear dependence observed for SDS shows that very high and very low discretionary spending led to a sharp increase in credit risk only for samples with relatively small values of Diff. The slight decrease in credit risk at the smallest Diff values is related to the fact that the data do not uniformly populate the feature space. In the region of the smallest Diff values, there are few data samples, and the available samples are characterized by high values of SI, SDS, and SRS, which contribute to a reduction in credit risk. A joint analysis of all SHAP dependence plots makes it possible to capture both the overall trend of how each feature affects the model and the individual behavior of specific instances through their interactions with other features.

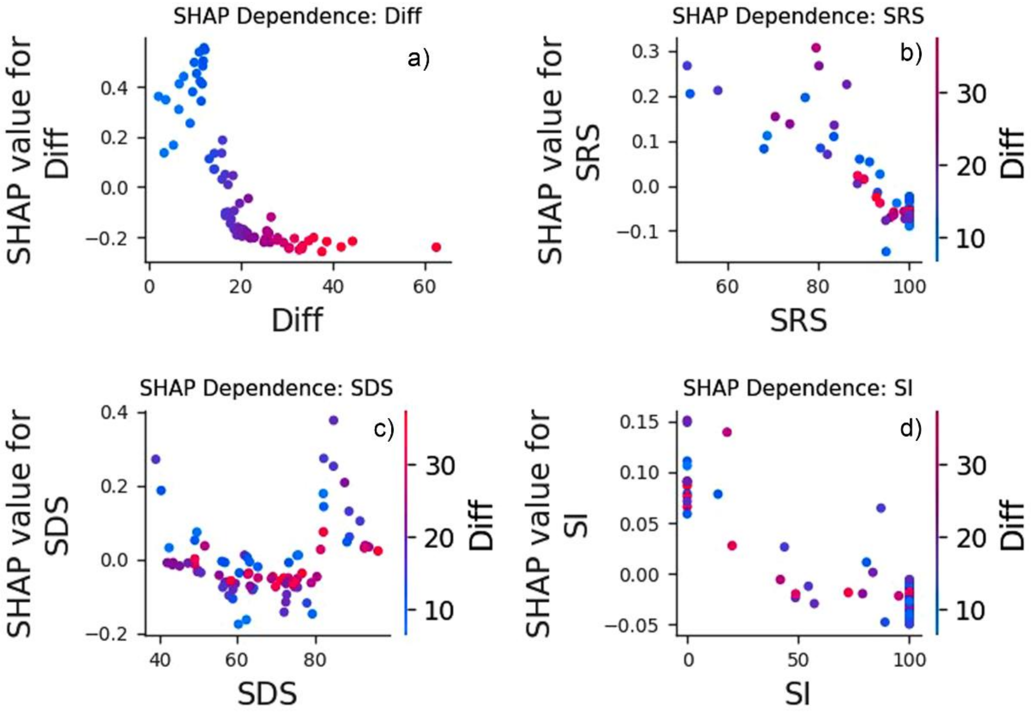

Fig. 4. SHAP dependence plots: a – Diff, b – SRS, c – SDS, d – Si.

In this work, a local analysis of feature importance coefficients was carried out for all samples in the test set, taking into account the distribution of their credit risk levels. **Table 3** presents the local feature importance values obtained using SHAP, LIME, DNFS, and BRB-ER for two different data samples.

Table 3. **Local feature importance coefficients for two samples**

Feature	Sample 1				Sample 2			
	LIME	SHAP Local	DNFS	BRB ER	LIME	SHAP Local	DNFS	BRB ER
Diff	–0.03	0.14	0.26	0.14	0.50	0.31	0.21	0.24
SRS	–0.22	–0.01	–0.12	0.08	0.20	0.11	0.08	0.12
SDS	–0.10	–0.08	0.02	0.1	0.03	0.01	0.05	0.02
SI	0.09	0.09	0.1	0.08	0.08	0.08	0.11	0.14

In **Fig. 5** on the right, the numerical feature values for these two samples are shown, while the corresponding credit risk levels are shown on the left. The central LIME plot illustrates the local contribution of each feature to the model prediction for the selected sample. **Fig. 6** shows SHAP waterfall plots for sample 1 (**Fig. 6a**) and sample 2 (**Fig. 6b**). The SHAP waterfall plot represents the stepwise construction of the model prediction, starting from the baseline (expected value) and adding the contributions of individual features that shift the prediction towards higher or lower credit risk.

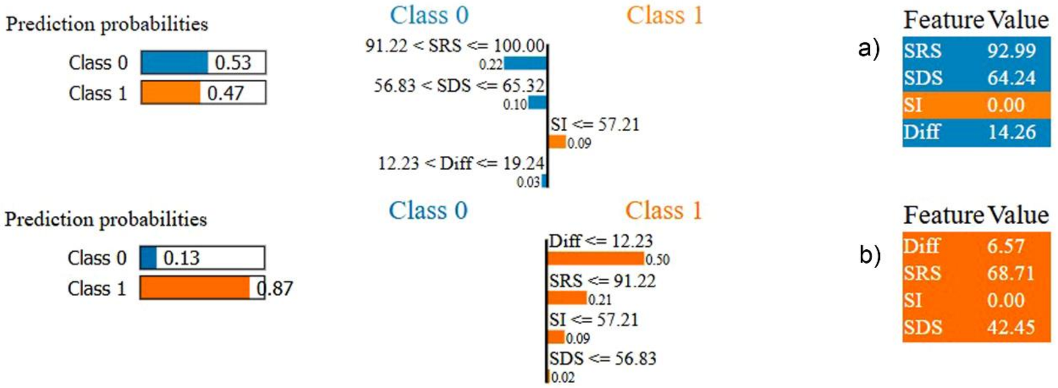

Fig.5. LIME-based local feature importance: a – sample 1, b – sample 2.

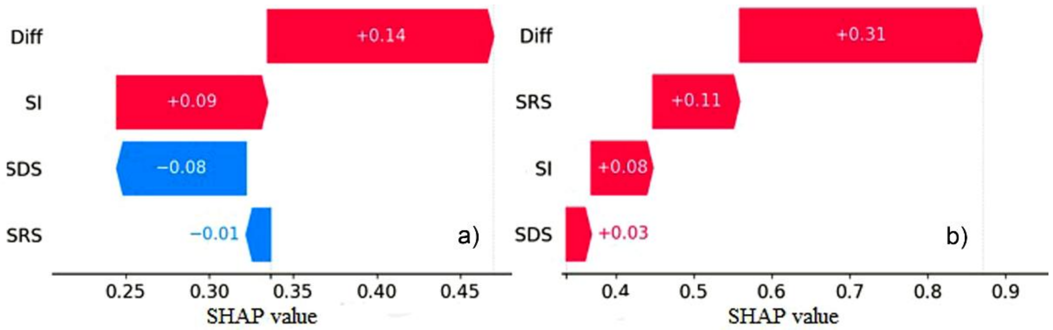

Fig.6. SHAP waterfall plots: a – sample 1, b – sample 2.

Both DNFS and BRB also provide local feature contribution coefficients. In DNFS, these are derived from the activations of fuzzy terms and the corresponding rules in the deep neural structure for each sample, whereas in BRB they result from the combination of rule matching degrees, belief weights, and aggregated belief levels. The results obtained using all methods are summarized in **Table 3**.

For sample 1, the local effects of the features act in different directions (**Fig. 5** and **Fig. 6**). Some features increase the prediction in LIME, whereas in SHAP they exhibit the opposite effect. In particular, the feature Diff has the strongest positive contribution to credit risk in SHAP, while in LIME it has only a minor contribution in the direction of reducing credit risk. The numerical value of Diff for sample 1 is 14.26. As follows from **Fig. 2d** and **Fig. 3d**, this value lies in a region where credit risk changes rapidly. Such rapid variation in credit risk indicates nonlinear local feature interactions and reduced model stability in the neighborhood of this instance, making it an important candidate for further analysis as a potentially unstable or borderline (between-class) case. Correlation analysis of the feature importance coefficients for sample 1 (**Table 3**) shows only moderate agreement between methods (0.58–0.76).

For sample 2, the directions of the local feature effects in LIME and SHAP are consistent (**Fig. 5** and **Fig. 6**), with a high pairwise correlation of 0.98 between them. In contrast to sample 1, the model clearly assigns sample 2 to a particular class. This is reflected in the consistency and high correlation of the feature importance coefficients (0.88–0.98) obtained from all local methods reported in **Table 3**.

CONCLUSION

This study proposes a multi-level approach to explaining credit scoring models based on open banking data, which combines global XAI methods (Feature Importance, SHAP) with interpretable DNFS and BRB-ER structures. The classification models for credit risk assessment (Random Forest and Deep Neuro-Fuzzy System) demonstrate that, after hyperparameter optimization, classification accuracy increases to 0.88 for RF and 0.91 for DNFS, with DNFS providing a better trade-off between predictive performance and structural interpretability due to its automatically generated fuzzy rules.

The global feature importance analysis shows a high correlation between the coefficients obtained using different methods (Feature Importance, SHAP, and DNFS). This indicates consistent model behavior and confirms that the key features drive the prediction irrespective of the explanation technique used. In addition, DNFS produces a set of fuzzy rules that allows the prediction to be interpreted in a logically transparent form, combining quantitative importance scores with textual rule-based explanations.

The analysis of local feature importance coefficients obtained with LIME, local SHAP, DNFS, and BRB-ER shows that interpretable models can adequately capture local feature contributions while providing enriched explanations through rule bases, belief degrees, and fuzzy term activations. A comparison of local feature importance for two test samples demonstrates that, in regions of strong nonlinearity and feature interactions, local discrepancies between methods may arise. This highlights the need to combine global and local explanations to achieve a robust and reliable interpretation of model decisions.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and/or publication of this article.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [M.F., L.D.]; methodology, [M.F., L.D.]; validation, [M.F., L.D.]; formal analysis, [M.F., L.D.]; investigation, [M.F., L.D.]; writing – review and editing, [M.F., L.D.]; visualization, [M.F., L.D.];

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Martins T., de Almeida A. M., Cardoso E., L. Nunes (2024). Explainable Artificial Intelligence (XAI): A Systematic Literature Review on Taxonomies and Applications in Finance, in IEEE Access, vol.12, pp.618-629. https://doi.org/10.1007/s10462-024-10854-8

[2] Černevičienė, J., Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artif Intell Rev 57, 216. https://doi.org/10.1007/s10462-024-10854-8

[3] Yeo, W.J., Van Der Heever, W., Mao, R. et al. (2025). A comprehensive review on financial explainable AI. Artif Intell Rev 58, 189 https://doi.org/10.1007/s10462-024-11077-7

[4] Chinnaraju A. (2025). Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. World Journal of Advanced Engineering Technology and Sciences, 14(03), 170-207. https://doi.org/10.30574/wjaets.2025.14.3.0106

[5] Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K. and Menegaz, G. (2025), A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Adv. Intell. Syst., 7: 2400304. https://doi.org/10.1002/aisy.202400304

[6] de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. Journal of Risk and Financial Management, 15(12), 556. https://doi.org/10.3390/jrfm15120556

[7] Shreya, Harsh Pathak. (2025). Explainable Artificial Intelligence Credit Risk Assessment using Machine Learning. Computer Science Machine Learning. arXiv:2506.19383 https://doi.org/10.48550/arXiv.2506.19383

[8] Talaat, F.M., Aljadani, A., Badawy, M. et al. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Comput & Applic 36, 4847–4865 https://doi.org/10.1007/s00521-023-09232-2

[9] Aosen Gong, Wei He, You Cao, Guohui Zhou, Hailong Zhu. (2025). Interpretability metrics and optimization methods for belief rule based expert systems, Expert Systems with Applications, Volume 289, 128363, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2025.128363

[10] Yaqian You, Jianbin Sun, Ruirui Zhao, Yuejin Tan, Jiang Jiang, (2024). A rule reasoning diagram for visual representation and evaluation of belief rule-based systems, Expert Systems with Applications, Volume 255, Part D, 124806, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2024.124806

[11] Talpur, N., Abdulkadir, S.J., Alhussian, H. et al. (2022). A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods. Neural Comput & Applic 34, 1837–1875. https://doi.org/10.1007/s00521-021-06807-9

[12] Talpur, N., Abdulkadir, S.J., Alhussian, H. et al. (2023). Deep Neuro-Fuzzy System application trends, challenges, and future perspectives: a systematic survey. Artif Intell Rev 56, 865–913. https://doi.org/10.1007/s10462-022-10188-3

[13] Yang, J. (2021). Fast TreeSHAP: Accelerating SHAP Value Computation for Trees. ArXiv. https://arxiv.org/abs/2109.09847

[14] Aljadani, A.; Alharthi, B.; Farsi, M.A.; Balaha, H.M.; Badawy, M.; Elhosseini, M.A. Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach. Mathematics 2023, 11, 4055. https://doi.org/10.3390/math11194055

[15] Badhon, B., Chakrabortty, R.K., Anavatti, S.G., Vanhoucke, M., IRAF-BRB: An explainable AI framework for enhanced interpretability in project risk assessment, Expert Systems with Applications, Volume 285, 2025, 127979, https://doi.org/10.1016/j.eswa.2025.127979

[16] Fostyak, M. Demkiv, L. (2025). Hybrid Optimized BRB–ML Model for Credit Rating Prediction in Open Banking Systems. Artificial Intelligence Stuc. intelekt. ; 30; (3):110-118 https://doi.org/10.15407/jai2025.03.110

[17] Rawal, R., Chug, R, Singh A., Prakash A., (2025). Review of Deep Learning Revolution on Neuro-Fuzzy Systems, Advances in Data Science and Adaptive Analysis, V.17(03). https://doi.org/10.1142/S2424922X25300015

[18] Fostyak, M. (2024). Development of an ai domain in a data mesh network for customer credit classification using transaction data IEEE 19th International Conference on Computer Science and Information Technologies (CSIT) IEEE Lviv Polytechnic Week 16-19 October, Lviv, DOI:10.1109/CSIT65290.2024.10982569

[19] Fostyak, M., Demkiv L. (2025). A data-centric approach to building ai models for determining the credit rating of fintech company clients based on open banking. ISSN 2710 – 1673 Artificial Intelligence Stuc. intelekt № 1. https://doi.org/10.15407/jai2025.01.132

[20] Hielkrem, L.O., Lange, P. E. d. (2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. Journal of Risk and Financial Management*, 16*(4), 221. https://doi.org/10.3390/jrfm16040221

ПОЯСНЮВАЛЬНІ ТА ІНТЕРПРЕТОВАНІ МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ВІДКРИТИХ БАНКІВСЬКИХ ДАНИХ

Маркіян Фостяк, Лідія Демків*

Львівський національний університет імені Івана Франка,
кафедра системного проєктування,
вул. М. Драгоманова, 50, Львів, 79005.
**lidiya.demkiv@lnu.edu.ua*

АНОТАЦІЯ

Вступ. Розвиток моделей штучного інтелекту та машинного навчання значно вплинув на фінансову аналітику і прийняття кредитних рішень. Моделі забезпечують високу точність, проте функціонують як «чорні скриньки», що ускладнює інтерпретацію внутрішніх механізмів їх роботи. У сфері відкритого банкінгу, де рішення безпосередньо впливають на доступ користувачів до фінансових ресурсів, така непрозорість є суттєвим недоліком. Тому виникає потреба у розвитку пояснювальних та інтепретованих підходів, які дозволяють встановити причинно-наслідкові зв’язки між вхідними ознаками та вихідними прогнозами.

Матеріали та методи. Методи досліджень ґрунтуються на багаторівневому підході до інтерпретації ML-моделі. Застосовано Feature Importance для статистичної оцінки внеску ознак, LIME для локальної інтерпретованості, SHAP (SHapley Additive exPlanations) для виявлення нелінійних залежностей. Структурну інтерпретацію забезпечує DNFS (Deep Neuro-Fuzzy System) через формування нечітких правил, а BRB-ER (Belief Rule Base with Evidential Reasoning) додає логічно узгоджене пояснення рішень на основі бази правил.

Результати. Показано, що після оптимізації гіперпараметрів моделей кредитного ризику на основі даних відкритого банкінгу точність моделі DNFS зростає на 4% більше ніж для моделі Random Forest. Глобальний підхід до визначення коефіцієнтів важливості ознак, отриманих за допомогою Feature Importence, SHAP та DNFS, показав їх високу кореляцію (більше 88%), що свідчить про стабільність моделі. На локальному рівні визначено зразки, які зменшують точність моделі. Візуалізація SHAP графіків розкрила області лінійної та нелінійної взаємодії ознак та їх вплив на прийняття рішень.

Висновки. На відміну від традиційного використання окремих XAI-методів для пояснення результатів моделей машинного навчання, у роботі поєднано глобальні та локальні метрики важливості ознак (Feature Importance, SHAP, LIME), нечітко-правилові метрики DNFS та агреговані коефіцієнти BRB-ER. Запропонований підхід дає змогу локалізувати причини зниження точності, визначати нелінійні залежності ознак, а також оцінити узгодженість пояснень через кореляційний аналіз між методами.

Ключові слова: пояснення штучного інтелекту (XAI), машинне навчання, база правил переконань (BRB), глибока нейронна нечітка система (DNFS), нечітка логіка.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
03 December, 2025	12 December, 2025	14 December , 2025	25 December , 2025