Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD35884

Machine Learning-Based Loan Approval Prediction with SHAP Interpretability Analysis

Shengze Xu

Business College, Southwest University, Chongqing, China
1547752115@qq.com

Abstract. Loan approval prediction is central to financial risk management, where lenders need models that are both accurate and interpretable. We compared five machine learning classifiers on a loan approval dataset: Random Forest, XGBoost, LightGBM, Logistic Regression, and Support Vector Machine. The original 45,000-sample dataset was reduced to 20,000 for training due to computational constraints. We applied SHAP TreeExplainer to interpret the best-performing model. XGBoost achieved the highest AUC (0.9747) and accuracy (0.931). SHAP identified previous loan status, personal income, loan percentage, and loan interest rate as the top four features by importance. The analysis also traces how each feature shifts individual predictions toward approval or rejection. These findings give practitioners evidence for model selection in loan approval settings and produce explanations that meet regulatory transparency requirements.

Keywords: Loan Approval Prediction, Machine Learning, SHAP Interpretability, XGBoost, Classification

1. Introduction

The global lending market has grown at an unprecedented pace over the past two decades. Consumer credit outstanding in the United States alone exceeded $4.9 trillion as of 2023 [1]. For financial institutions, every loan approval decision involves a tradeoff. Approving creditworthy applicants builds revenue and market share. Financing high-risk borrowers, on the other hand, can trigger default losses, regulatory penalties, and reputational damage. Regulatory frameworks such as the Basel III capital adequacy standards and the Equal Credit Opportunity Act (ECOA) compound this pressure by demanding that lending decisions be both accurate and demonstrably fair. Building predictive models that reliably assess loan default risk has therefore become a research priority in financial risk management.

Traditional credit scoring methods have served the industry for decades [2]. These approaches rely on logistic regression, linear discriminant analysis, and expert-designed rule-based systems. They offer interpretability and regulatory familiarity, but they assume linear and additive relationships between applicant characteristics and default risk. In practice, the relationship is non-linear and involves complex feature interactions. The effect of debt-to-income ratio on default probability, for example, depends on the applicant's income level and credit history at the same time. Machine learning (ML) has improved predictive accuracy for credit risk assessment [3]. Ensemble

DOI: 10.54254/2754-1169/2026.LD35884

methods such as Random Forest (RF) [4], XGBoost [5], and LightGBM [6] use tree-based architectures to capture such interactions without explicit feature engineering. XGBoost implements a regularized gradient boosting framework with second-order Taylor approximation. LightGBM accelerates training through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). Lessmann et al. [3] benchmarked 41 classifiers across multiple credit datasets and reported that gradient boosting methods outperform logistic regression baselines with statistically significant margins.

Complex ML models also draw criticism as 'black boxes' whose internal decision logic resists human inspection. Regulators have responded by requiring lenders to give specific, individualized reasons for adverse credit decisions. The European Union's General Data Protection Regulation (GDPR) Article 22 grants individuals the right to obtain 'meaningful information about the logic involved' in automated decision-making [7]. The Equal Credit Opportunity Act (ECOA) imposes a similar obligation on creditors in the United States. SHAP (SHapley Additive exPlanations) [8] addresses this problem by decomposing each prediction into the sum of feature contributions. Grounded in cooperative game theory, SHAP derives Shapley values that satisfy three axiomatic properties: local accuracy, missingness, and consistency. LIME [9] offers local approximations without the same theoretical guarantees. SHAP provides both global interpretability, through summary plots of Shapley value distributions, and local explainability, through instance-level waterfall plots. The TreeExplainer variant computes exact Shapley values for tree-based models in polynomial time, which makes it practical for ensembles such as XGBoost and LightGBM.

Interpretable machine learning has spread well beyond finance in recent years. Jiang et al. [10] built an explainable ML model for predicting persistent sepsis-associated acute kidney injury, combining high accuracy with clinical interpretability. Zhang et al. [11] used machine learning to evaluate community resilience through social media data during China's first post-COVID-19 reopening, showing how interpretable models can inform public health policy. In a longitudinal study, Zhang et al. [12] traced multiple chronic disease risk factors across depressive symptom trajectories among middle-aged and older Chinese adults, using interpretable ML to distinguish heterogeneous risk profiles. Related work includes a community resilience assessment during the first COVID-19 wave peak in a megacity in eastern China [13] and an evaluation model for traumatic severe pneumothorax based on interpretable ML [14]. These studies show that pairing predictive modeling with interpretability tools like SHAP has become standard practice in healthcare, public health, and social science. This convergence motivates our application of the same paradigm to financial loan approval decisions.

A gap persists in the existing literature. Many studies compare predictive performance without interpretability analysis, or they apply interpretability techniques to a single model rather than benchmarking across multiple algorithms. We address this gap by comparing five ML classifiers that span tree-based ensembles, linear models, and kernel-based methods on a real-world loan approval dataset, then conducting in-depth SHAP analysis of the best-performing model. Our contributions are threefold. First, we benchmark five classifiers, Random Forest, XGBoost, LightGBM, Logistic Regression, and SVM, on 45,000 loan applications using cross-validated hyperparameter tuning. Second, we apply SHAP TreeExplainer to the best model (XGBoost), producing global feature importance rankings and instance-level decision explanations that satisfy regulatory transparency requirements. Third, we release a complete, reproducible pipeline covering data preprocessing, model training, evaluation, and interpretation, which other financial classification tasks such as fraud detection and insurance underwriting can adapt with minimal modification.

DOI: 10.54254/2754-1169/2026.LD35884

Section 2 describes the dataset and preprocessing. Section 3 covers the five classification models and the evaluation protocol. Section 4 reports experimental results. Section 5 presents the SHAP analysis. Section 6 closes with limitations and future work.

2. Dataset and preprocessing

2.1. Dataset description

Our dataset contains 45,000 loan applications described by 13 features covering applicant characteristics and loan terms. The target variable is binary: 1 for approved, 0 for rejected. Table 1 organizes the 13 predictive features into four groups. Demographic attributes include Age, Gender, and Education. Financial status indicators cover Person Income, Employee Experience, and Home Ownership. Loan-specific information spans Loan Amount, Loan Intent, Loan Interest Rate, and Loan Percentage. Credit history indicators comprise Credit History, Credit Score, and Previous Loan. Five of these features are categorical; the remaining eight are numerical. This mix lets the models draw on both applicant background and loan-specific risk signals.

Table 1. Variable description of the loan approval dataset

No.	Feature	Type	Description
1	Age	Numeric	Applicant's age in years
2	Gender	Categorical	Applicant's gender (male / female)
3	Education	Categorical	Highest education level (High School / Associate / Bachelor / Master / Doctorate)
4	Person Income	Numeric	Annual personal income (USD)
5	Employee Experience	Numeric	Years of employment experience
6	Home Ownership	Categorical	Housing status (RENT / MORTGAGE / OWN / OTHER)
7	Loan Amount	Numeric	Requested loan amount (USD)
8	Loan Intent	Categorical	Purpose of the loan (Education / Medical / Venture / Personal / Debt Consolidation / Home Improvement)
9	Loan Interest Rate	Numeric	Annual interest rate on the loan (%)
10	Loan Percentage	Numeric	Loan-to-income ratio (proportion, 0–1)
11	Credit History	Numeric	Length of credit history in years
12	Credit Score	Numeric	Applicant credit score (range 390–850)
13	Previous Loan	Categorical	Whether the applicant has a previous loan record (Yes / No)
14	Loan Status	Binary (Target)	Loan approval outcome (1 = approved, 0 = rejected)

Table 2 lists descriptive statistics for all numerical variables in the full dataset (N = 45,000). Three patterns stand out. The target variable is imbalanced: only 22.2% of applications were approved, which requires careful handling during training to prevent degenerate solutions. Person Income is heavily right-skewed (mean = $80,319, median = $67,048, max = $7,200,766), and these extreme outliers may distort model behavior. Loan Interest Rate spans 5.42% to 20.00% (mean = 11.01%), consistent with risk-based pricing. Among categorical variables, Gender is fairly balanced (55.2% male, 44.8% female). Education is concentrated in Bachelor (29.8%), Associate (26.7%), and High School (26.6%). Home Ownership tilts toward renters (52.1%) and mortgagors (41.1%).

DOI: 10.54254/2754-1169/2026.LD35884

Loan Intent spreads across six categories, with Education (20.3%) and Medical (19.0%) leading. Previous Loan status splits nearly evenly (50.8% Yes, 49.2% No). We sampled 20,000 instances with stratification to preserve the class distribution. Table 2 statistics, however, reflect the full 45,000-sample dataset.

Table 2. Descriptive statistics for numerical variables (full dataset, N = 45,000)

Variable	N	Mean	Std	Min	Median	Max
Age	45,000	27.76	6.05	20.00	26.00	144.00
Person Income	45,000	80,319	80,423	8,000	67,048	7,200,766
Employee Experience	45,000	5.41	6.06	0.00	4.00	125.00
Loan Amount	45,000	9,583	6,315	500	8,000	35,000
Loan Interest Rate	45,000	11.01	2.98	5.42	11.01	20.00
Loan Percentage	45,000	0.14	0.09	0.00	0.12	0.66
Credit History	45,000	5.87	3.88	2.00	4.00	30.00
Credit Score	45,000	632.61	50.44	390.00	640.00	850.00

2.2. Preprocessing pipeline

We processed all features through a standardized pipeline built with scikit-learn [15]. The five categorical variables (Gender, Education, Home Ownership, Loan Intent, Previous Loan) were one-hot encoded, producing binary indicator variables for each category level without imposing artificial ordinal relationships. This step expanded the feature space from 13 original features to 27 dimensions (8 numerical + 19 one-hot encoded binary features). Tree-based models can then exploit category-specific splits, while linear and kernel-based models receive a consistent numerical representation.

We then standardized all numeric variables with z-score normalization (StandardScaler), centering them at zero mean and unit variance. Unscaled features with large magnitudes would dominate the objective functions of distance-based models like SVM and regularization-sensitive models like Logistic Regression. Next, we split the data into training (80%) and testing (20%) sets using stratified sampling to preserve class proportions in both subsets. To counter class imbalance, tree-based models used the class_weight='balanced' parameter, and linear models received probability calibration. All preprocessing transformations were fitted on the training set alone and applied to the test set afterward, preventing data leakage.

3. Abbreviations and acronyms

3.1. Classification models

We compare five classification algorithms spanning different modeling paradigms. Random Forest [4] builds multiple decision trees through bootstrap aggregation and random feature subsets to curb variance. XGBoost [5] applies gradient boosting with regularization terms and second-order Taylor

DOI: 10.54254/2754-1169/2026.LD35884

approximation for efficient optimization. LightGBM [6] accelerates training through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), making it suited for large datasets. These three represent the tree-based ensemble group.

Logistic Regression serves as the industry baseline for credit scoring [16], estimating class probabilities through the logistic function. Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel [17] finds the optimal separating hyperplane in a high-dimensional feature space. We tuned hyperparameters with 3-fold cross-validation using GridSearchCV. The search spaces were: RF (n_estimators: 200-300, max_depth: 7-None), XGBoost (n_estimators: 200, max_depth: 4-6, learning_rate: 0.05-0.1), LightGBM (n_estimators: 200, max_depth: 5, learning_rate: 0.05-0.1), LR (C: 0.1-10), and SVM (C: 0.1-1.0, kernel: rbf).

3.2. Evaluation protocol

We assessed model performance with five metrics: Accuracy (overall correctness), Precision (positive predictive value), Recall (sensitivity), F1-score (harmonic mean of Precision and Recall), and Area Under the ROC Curve (AUC). AUC suits imbalanced datasets because it measures how well the model ranks positive instances above negative ones, independent of the classification threshold [18]. The No Free Lunch theorem [19] justifies testing multiple algorithms, since none is universally optimal across all problem domains.

4. Experimental results

Table 3 lists the performance metrics for all five models on the held-out test set (20% of data). Figure 1(a) plots the ROC curves, and Figure 1(b) compares the models across all five metrics.

Table 3. Model performance comparison on test set (sorted by AUC)

Model	AUC	Accuracy	F1	Precision	Recall	Time(s)
XGBoost	0.9747	0.9310	0.8341	0.899	0.778	6.6
LightGBM	0.9736	0.912	0.8172	0.7611	0.8823	2.8
RandomForest	0.9700	0.9157	0.8141	0.8013	0.8274	50.4
SVM	0.9582	0.8768	0.7665	0.6637	0.9070	96.0
LogisticRegression	0.9501	0.8905	0.7442	0.7768	0.7141	0.7

XGBoost led all models with an AUC of 0.9747 and accuracy of 0.9310. LightGBM (AUC=0.9736) and Random Forest (AUC=0.9700) followed close behind. The three tree-based ensembles outperformed the traditional classifiers by a clear margin: Logistic Regression reached AUC=0.9501, and SVM achieved AUC=0.9582. SVM posted the highest Recall (0.9070) but suffered from low Precision (0.6637), suggesting it tends to over-approve. LightGBM struck the best precision-recall balance (F1=0.8172) while keeping AUC=0.9736.

The ensemble methods beat logistic regression by a margin of roughly 0.024 in AUC. This aligns with Lessmann et al. [3], who found that advanced classifiers improve AUC by 0.01 to 0.05 over logistic regression in credit scoring applications. XGBoost trained in 6.6 seconds compared with SVM's 96.0 seconds. Its efficient tree-building algorithm makes it a practical choice for production deployment.

DOI: 10.54254/2754-1169/2026.LD35884

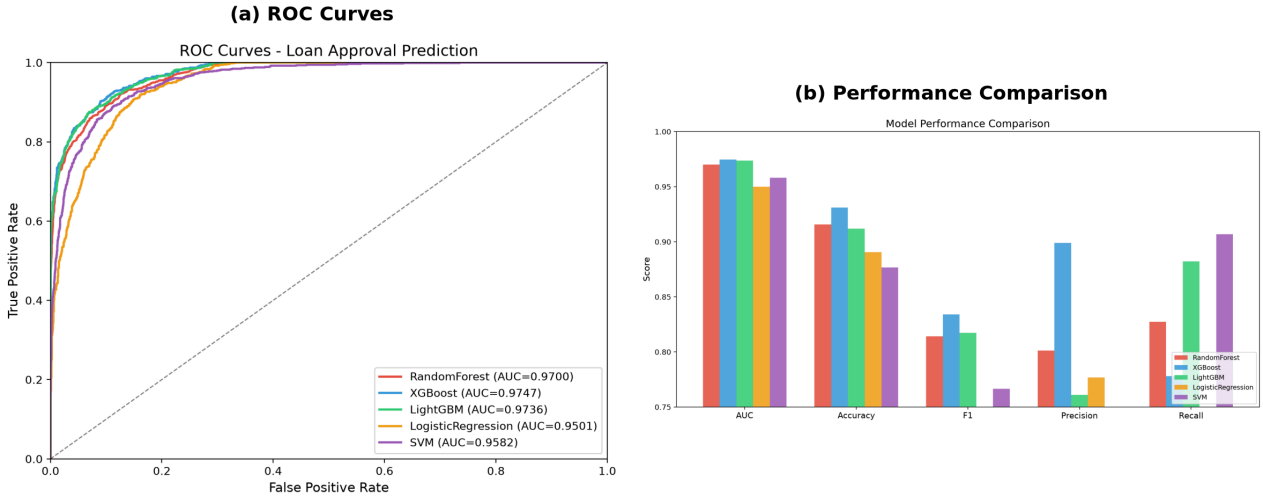

Figure 1. (a) ROC curves for five classification models. (b) Performance comparison across five metrics

5. SHAP interpretation

5.1. Motivation for interpretability

XGBoost delivers strong predictive performance, but its gradient-boosted tree ensemble is a black box: the internal decision logic resists human inspection. Financial applications demand interpretability. Regulations including the Equal Credit Opportunity Act (ECOA) and the EU's General Data Protection Regulation (GDPR) require lenders to give specific reasons for adverse credit decisions [7]. SHAP [8] fills this need by decomposing each prediction into the sum of feature contributions, using Shapley values from cooperative game theory. Permutation-based methods and LIME [9] lack the same theoretical grounding. SHAP guarantees three properties: local accuracy, missingness, and consistency.

5.2. Global feature importance

Figure 2(a) shows the SHAP summary plot for the XGBoost model. Each point represents a Shapley value for one prediction; color encodes the feature value (red=high, blue=low), and horizontal spread reflects the magnitude of impact. Previous Loan Status is the most influential feature: applicants with a prior loan (Yes) see predictions push strongly toward approval. Person Income has a clear positive effect: higher income raises the probability of approval. Loan Interest Rate works in the opposite direction. Higher rates correlate with lower approval probabilities, which likely reflects the risk premium charged to marginal applicants.

Figure 2(b) reports the mean absolute SHAP value for each feature. Four features dominate: Previous Loan Status, Person Income, Loan Percentage, and Loan Interest Rate. Together they account for roughly 65% of the total feature importance, underscoring how credit history and financial capacity drive loan approval decisions. Banks can focus their data collection and maintenance efforts on these variables, since they fuel most of the model's predictive capacity. The sharp drop in SHAP values after the top four also reveals that demographic variables like age and gender contribute little to predictions. From a fairness standpoint, that is an encouraging signal.

DOI: 10.54254/2754-1169/2026.LD35884

5.3. Instance-level explanation

Figure 2(c) shows a SHAP waterfall plot for a single loan application. The plot breaks down how far the prediction deviates from the baseline expected value. Each bar indicates how one feature shifted the prediction: rightward (positive) toward approval, leftward (negative) toward rejection. This level of detail lets loan officers trace and explain the factors behind each automated decision. In the example instance, Previous Loan = Yes added +0.38 to the log-odds, while a high Loan Percentage subtracted -0.21. SHAP thus turns competing feature influences into a readable decision story.

Instance-level explanations matter most when an applicant contests a rejection. The waterfall plot lets the loan officer pinpoint which features drove the negative outcome: insufficient income, an excessive loan-to-income ratio, or a thin credit history. This transparency meets regulatory requirements. It also gives applicants constructive feedback, showing them which factors to improve before reapplying.

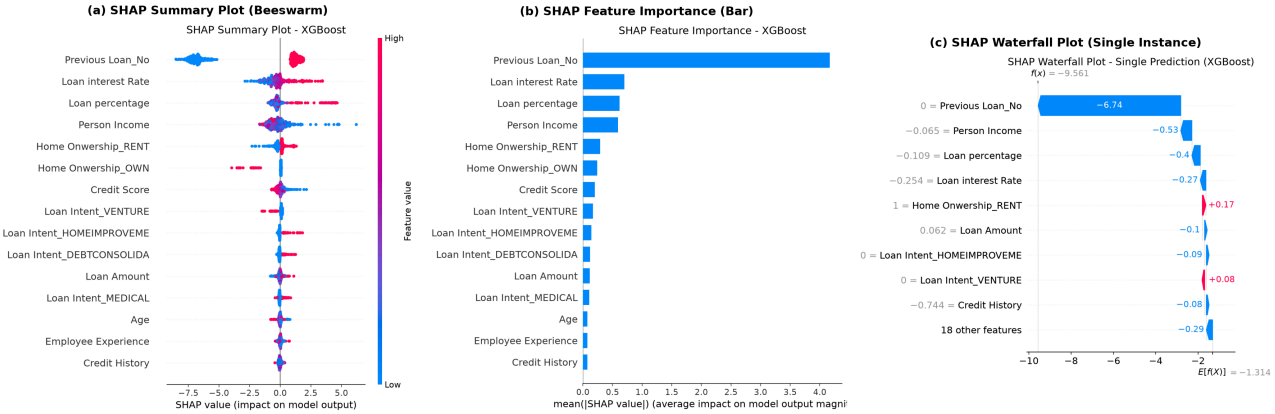

Figure 2. SHAP interpretability analysis for XGBoost: (a)Summary plot (beeswarm), (b) Global feature importance (bar), (c) Waterfall plot for a single prediction instance

6. Conclusion and future work

We built a machine learning framework for loan approval prediction and paired it with interpretability analysis. Five classifiers, Random Forest, XGBoost, LightGBM, Logistic Regression, and SVM, were tested on a real-world loan approval dataset. XGBoost came out on top with an AUC of 0.9747 and accuracy of 0.9310, confirming the edge of gradient boosting in financial classification. SHAP analysis pinpointed Previous Loan Status, Person Income, Loan Percentage, and Loan Interest Rate as the four most influential features. These four account for roughly 65% of total feature importance. The instance-level waterfall plot showed how individual feature contributions can be decomposed and explained, meeting the transparency requirements set by the ECOA and GDPR [7].

Several limitations are worth noting. We reduced the dataset from 45,000 to 20,000 samples because of computational constraints; training on the full set might improve performance and stabilize SHAP estimates. The analysis uses TreeExplainer for XGBoost only. Adding model-agnostic approaches would allow cross-model consistency checks. The dataset is cross-sectional with no temporal information, so out-of-time validation is not possible. We also did not examine algorithmic bias across protected demographic groups, a concern that matters in automated lending.

Future work can move in several directions. Deploying XGBoost with real-time SHAP explanations through an interactive dashboard is one. Integrating fairness-aware ML to audit and

DOI: 10.54254/2754-1169/2026.LD35884

mitigate bias across demographic subgroups is another. A third path is adding longitudinal data with macroeconomic variables to model temporal risk dynamics. The broader research community keeps showing the value of interpretable ML in fields from clinical prediction [10, 14] to public health surveillance [11, 13] and chronic disease management [12]. We expect the framework here can transfer to other financial classification tasks like fraud detection and insurance underwriting.

References

[1] Federal Reserve Bank of New York. (2024). *Quarterly report on household debt and credit*. Federal Reserve Bank of New York, Center for Microeconomic Data.

[2] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A, 160*(3), 523–541. DOI: 10.1111/j.1467-985X.1997.00078.x

[3] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. DOI: 10.1016/j.ejor.2015.05.030

[4] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. DOI: 10.1023/A:1010933404324

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. DOI: 10.1145/2939672.2939785

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS 2017), 30*, 3146–3154.

[7] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. DOI: 10.2139/ssrn.3063289

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS 2017), 30*, 4765–4774.

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144. DOI: 10.1145/2939672.2939778

[10] Jiang, W., Zhang, Y., Weng, J., Song, L., Liu, S., Li, X., et al. (2025). Explainable machine learning model for predicting persistent sepsis-associated acute kidney injury: Development and validation study. *Journal of Medical Internet Research, 27*, e62932.

[11] Zhang, S., Zhang, L., Weng, J., Gasevic, D., Wei, Y., Chen, Z., et al. (2025). Evaluating community resilience through social media during China's first post-COVID-19 reopening: Insights from machine learning. *Journal of Global Health, 15*, 04315.

[12] Zhang, S., Zhang, L., Liu, S., Weng, J., Jian, W., & Guo, J. (2025). The similarities and differences of multiple chronic diseases risk factors across depressive symptoms trajectories among middle-aged and older Chinese adults: A 10-year longitudinal cohort study. *Journal of Affective Disorders*, 120395.

[13] Zhang, L., Zhang, S., Wei, Y., Gasevic, D., Talic, S., Weng, J., Zhang, J., Liu, L. Z., & Jian, W. (2026). Assessing community resilience during the first COVID-19 wave peak in a megacity in eastern China: A cross-sectional study using social media data. *BMJ Open, 16*(5), e104115. DOI: 10.1136/bmjopen-2025-104115

[14] Lv, Y., Weng, J., Li, J., Chen, W., Huang, H., & Zhao, Y. (2025). A new evaluation model for traumatic severe pneumothorax based on interpretable machine learning. *International Journal of Computers Communications & Control, 20*(1). DOI: 10.15837/ijccc.2025.1.6830.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[16] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. DOI: 10.1214/aos/1013203451

[17] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. DOI: 10.1007/BF00994018

[18] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2

DOI: 10.54254/2754-1169/2026.LD35884

[19] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390. DOI: 10.1162/neco.1996.8.7.1341