CREDIT RISK PREDICTION USING LIGHTGBM WITH TIME-AWARE VALIDATION AND SHAP INTERPRETATION

Sony Alfauzan
Independent Researcher & Alumni of Information Systems Study Program, Universitas Terbuka, Tangerang Selatan, Indonesia
e-mail: sonyalfauzan46@gmail.com

ABSTRACT

Accurate credit scoring is critical for financial stability, yet modern machine learning models often operate as "black boxes," creating a significant challenge for regulatory compliance and business trust. This research addresses this problem by developing a comprehensive, transparent, and robust framework for credit risk prediction. The primary objective was to build a high-performance model that is not only accurate but also fully interpretable and stable over time. The research method involved using a public loan dataset (2007-2014) with 466,285 loan records to train a Light Gradient Boosting Machine (LightGBM). To ensure real-world applicability and prevent temporal data leakage, a rigorous time-aware validation strategy was employed, splitting the data into a historical training set (2007-2013) and a future out-of-time (OOT) test set (2014). The SHapley Additive exPlanations (SHAP) framework was integrated to provide clear explanations for every prediction. The main research results demonstrate the model's strong predictive capability, achieving an AUC-ROC of 0.711 and a KS-statistic of 0.309 on the OOT dataset. The model's stability was confirmed with a low Population Stability Index (0.0685). Furthermore, SHAP analysis revealed that predictions were driven by financially intuitive factors like interest rate and annual income. The study concludes by presenting a complete workflow that successfully bridges the gap between high-performance modeling and the practical need for transparent, business-aligned risk management tools.

Keywords: credit scoring, machine learning, lightgbm, model explainability, out-of-time validation.

INTRODUCTION

The financial services industry's stability and profitability are fundamentally tied to its ability to effectively manage credit risk, which is the potential loss from a borrower's failure to meet their debt obligations (Bhattacharya et al., 2023). To mitigate this risk, financial institutions rely on credit scoring models to evaluate the creditworthiness of applicants. For decades, traditional statistical models like logistic regression were the industry standard due to their simplicity and inherent interpretability (Leo et al., 2019). However, these linear models often struggle to capture the complex, non-linear relationships present in financial data, potentially leading to suboptimal decision-making (Noriega-Rivera et al., 2021; Papouskova & Hájek, 2019).

The primary problem addressed in this research is the dual challenge faced by modern financial institutions: the need to enhance predictive accuracy using advanced machine learning (ML) models while simultaneously satisfying the increasing demand for transparency, fairness, and stability from regulators and stakeholders (Chen et al., 2025). Complex models like gradient boosting often perform as "black boxes," making their decisions difficult to understand and trust. Furthermore, a model's performance can degrade over time due to changing economic conditions, a phenomenon known as concept drift, making robust temporal validation essential (Office of the Comptroller of the Currency, 2021).

Therefore, the objective of this research is to design, implement, and validate a comprehensive workflow for developing a credit risk model that is simultaneously accurate, stable, and interpretable. The benefit of this research is a practical blueprint for financial institutions to adopt advanced ML techniques responsibly. We achieve this by: 1) Employing the LightGBM

algorithm for its high predictive performance. 2) Implementing a rigorous time-aware validation strategy to ensure the model's robustness on future data. 3) Integrating the SHAP framework to provide transparent and intuitive explanations for model predictions, thereby bridging the gap between performance and interpretability (Martins et al., 2024).

LightGBM was specifically chosen over alternative algorithms for three reasons. First, it employs a leaf-wise tree growth strategy and histogram-based split optimisation, offering faster training than depth-wise methods while maintaining superior discriminative accuracy on large-scale tabular data (Wang et al., 2022). Second, its robustness to class imbalance and native support for categorical features make it particularly well-suited for credit scoring tasks where the proportion of defaults is low (De Roux et al., 2024). Third, LightGBM's probability outputs are directly compatible with isotonic regression calibration and SHAP-based interpretability, fulfilling the dual requirements of accuracy and explainability mandated by modern credit risk governance frameworks.

METHOD

Data Source and Experimental Setting

This study uses the public dataset loan_data_2007_2014.csv, which contains 466,285 loan records from Lending Club. The experimental setting was designed to simulate a real-world production environment where a model is trained on historical data to predict future outcomes. The target variable was defined binarily: 1 (Bad Loan) for loans that have defaulted (statuses: 'Charged Off', 'Default', 'Late (31–120 days)', etc.) and 0 (Good Loan) for loans that are 'Fully Paid'. To ensure data quality and relevance, records with ambiguous statuses (e.g., 'Current', 'In Grace Period') were excluded, resulting in a final dataset of 237,695 samples for analysis.

A critical preprocessing step involved removing 28 features that would not be available at the time of a loan application, such as total_pymnt, recoveries, and last_pymnt_d, to prevent data leakage and ensure the model's predictions are based only on information available at the point of decision. Additional feature engineering was performed to create derived variables including FICO mean score, credit age in months, income-to-loan ratio, and installment-to-income ratio. A time-aware validation strategy was implemented by splitting the dataset chronologically: data from 2007 to 2013 served as the training set (189,392 samples), while data from 2014 was held out as the out-of-time (OOT) test set (48,303 samples). To preserve temporal integrity and prevent information leakage across time periods, a 5-fold GroupKFold cross-validation was applied on the training data, grouping by issuance quarter. This chronological splitting strategy is recommended for ensuring realistic performance estimates in credit risk models (Thomas et al., 2017).

Data Analysis Techniques

The primary modeling algorithm was Light Gradient Boosting Machine (LightGBM), chosen for its efficiency and high performance (Ke et al., 2017). The model was configured with the following hyperparameters: learning rate 0.05, maximum depth 8, 31 leaf nodes, L1 regularization 0.1, L2 regularization 0.1, and balanced class weights. Early stopping of 50 rounds was employed during cross-validation to mitigate overfitting.

All experiments were conducted in Python 3.10 within Google Colaboratory (CPU runtime, Google LLC). The implementation used: pandas 2.2.2 and NumPy 2.0.2 for data manipulation; scikit-learn 1.5.2 for preprocessing, cross-validation, and evaluation metrics; LightGBM 4.5.0 as the primary predictive model; XGBoost 2.1.1 for comparative benchmarking; SHAP 0.46.0 for model interpretability; and CalibratedClassifierCV (isotonic method) from scikit-learn for probability calibration. The complete source code is provided in the accompanying Jupyter notebook to ensure full reproducibility.

To contextualise why LightGBM was ultimately preferred over alternative approaches, six machine learning algorithms were evaluated under identical experimental conditions prior to hyperparameter tuning. Table 1 presents their baseline discriminative performance, using AUC-ROC as the primary selection criterion because it is threshold-independent and well-established in credit risk modelling literature (Lessmann et al., 2015). LightGBM achieved the highest AUC-ROC (0.703), surpassing Logistic Regression (0.702), XGBoost (0.702), Gradient Boosting (0.700), Random Forest (0.697), and Decision Tree (0.663). These results confirm that LightGBM provides the best discriminative capability among the algorithms tested on this dataset, justifying its selection as the primary modelling technique (Ke et al., 2017).

Table 1. Baseline Comparison of Six Machine Learning Algorithms (Pre-Tuning)

Model	AUC-ROC	F1-Score*
Logistic Regression	*0.702*	*0.411*
Decision Tree	*0.663*	*0.251*
Random Forest	*0.697*	*0.339*
Gradient Boosting	*0.700*	*0.216*
XGBoost	*0.702*	*0.200*
LightGBM (Selected)	*0.703*	*0.195*

*F1-Score evaluated at the default classification threshold (0.50) on the imbalanced test set. AUC-ROC, being threshold-independent, is used as the primary model selection criterion.

Model evaluation employed standard industry metrics and calibration techniques. The Kolmogorov–Smirnov (KS) Statistic measures the maximum separation between the score distributions of good and bad loans. In the context of ROC curve analysis, it is computed as:

$$KS = \max_{\tau}(TPR(\tau) - FPR(\tau)) \tag{1}$$

where TPR is the true positive rate (sensitivity/recall) and FPR is the false positive rate, each evaluated across decision thresholds τ. KS represents the maximum vertical distance between the empirical cumulative distribution functions of the two classes; higher values imply better class separation.

The Population Stability Index (PSI) monitors distribution shifts between a baseline and a new sample, calculated as:

$$PSI = \sum_{i=1}^{n} (a_i - e_i)\, ln\, (\frac{a_i}{e_i}) \tag{2}$$

where the summation is over $n$bins, $a_i$is the proportion in bin $i$for the actual (new) dataset, and $e_i$is the proportion in bin $i$for the expected (baseline) dataset. Interpretation: $PSI < 0.10$indicates stable population; 0.10–0.25requires monitoring; values > 0.25signal significant shift requiring model retraining.

For model interpretability, the SHAP (SHapley Additive exPlanations) framework was employed (Lundberg & Lee, 2017). SHAP attributes predictions via Shapley values from cooperative game theory, expressed as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \quad (3)$$

where $g$is the explanation model approximating the original prediction, $z'$is the simplified input feature vector (coalition of features), $M$is the total number of features, $\phi_0$is the base value representing the expected model output with no feature information, and $\phi_j$is the Shapley value for feature j. Positive $\phi_j$values move the prediction toward default (Class 1), while negative values move it toward non-default (Class 0). The magnitude $|\phi_j|$reflects the strength of a feature's influence on the prediction.

Probability calibration was performed using Isotonic Regression applied to out-of-fold (OOF) predictions to ensure predicted probabilities accurately reflect true default rates. Calibration quality was measured by the Expected Calibration Error (ECE):

$$ECE = \sum_{i=1}^{B} \frac{n_i}{N} |\bar{p}_i - \bar{y}_i| \quad (4)$$

where $B$is the number of bins (typically $B = 10$via quantile-based binning), $n_i$is the number of samples in bin i, $N$is the total sample size, $\bar{p}_i$is the mean predicted probability in bin i, and $\bar{y}_i$is the true proportion of positive outcomes (defaults) in bin i. Lower ECE indicates better calibration; values below 0.05 are considered excellent, while values above 0.10 indicate poor calibration requiring correction.

The accuracy of probabilistic predictions was also assessed using the Brier Score, which measures the mean squared difference between predicted probabilities and actual outcomes:

$$Brier = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2 \quad (5)$$

where $p_i \in [0,1]$is the predicted probability for sample i, and $y_i \in \{0,1\}$is the actual binary outcome ($y_i = 0$ for non-default/good loan; $y_i = 1$for default/bad loan). The Brier Score ranges from 0 (perfect predictions) to 1 (worst possible predictions); lower values indicate more accurate probability estimates. Benchmark thresholds: values below 0.10 are excellent, while values above 0.25 indicate poor probabilistic predictions.

RESULTS AND DISCUSSION

Model Performance on Out-of-Time Test Set

The model achieved strong predictive performance on the OOT 2014 dataset. The AUC-ROC score of 0.711 indicates good discriminative ability between good and bad loans, while the Kolmogorov-Smirnov statistic of 0.309 demonstrates substantial separation between the two classes. The PR-AUC of 0.448 shows the model maintains reasonable precision across different recall levels, which is particularly important given the class imbalance (15.4% default rate).

Table 2. Model Performance Metrics on OOT 2014 Dataset

Metric	Value	Interpretation
ROC-AUC	*0.711*	*Good discrimination*
PR-AUC	*0.448*	*Reasonable precision*
KS Statistic	*0.309*	*Strong separation*
Brier Score	*0.119*	*Well-calibrated*
ECE	*0.031*	*Excellent calibration*

Based on Table 2, it can be explained that the model demonstrates robust performance across multiple evaluation metrics. The Brier score of 0.119 and Expected Calibration Error of 0.031 confirm that the model is well-calibrated, meaning its probability estimates are reliable for decision-making.

Calibration and Stability Analysis

Probability calibration using Isotonic Regression on OOF predictions significantly improved the model's reliability. The Expected Calibration Error decreased from 0.084 before calibration to 0.031 after calibration, representing a 63% improvement. Similarly, the Brier score improved from 0.128 to 0.119. This calibration step is critical for ensuring that predicted probabilities can be trusted for risk-based pricing and decision-making.

The model's stability over time was assessed using the Population Stability Index (PSI) comparing OOF predictions (2007-2013) with OOT predictions (2014). The PSI value of 0.0685 is well below the industry threshold of 0.10, indicating stable prediction distributions and confirming the model's robustness to temporal changes. The gap between OOF AUC (0.717) and OOT AUC (0.711) is only 0.006, further validating temporal stability.

Feature Importance and Model Interpretability

The SHAP analysis provided transparent insights into the model's decision-making process. Figure 1 shows the top 10 features by mean absolute SHAP value, revealing the most influential predictors of credit risk.

Based on Figure 1, it can be explained that the model learned financially intuitive relationships from the data. The top five features influencing credit risk predictions are: (1) Interest Rate (int_rate), (2) Annual Income (annual_inc), (3) Installment-to-Income Ratio, (4) Debt-to-Income Ratio (dti), and (5) Number of Credit Inquiries in the Last Six Months (inq_last_6mths). This finding corroborates established financial theories and previous research, confirming that higher interest rates and debt ratios increase risk, while higher income decreases it (Siddiqi, 2017).

Interest rate emerges as the strongest predictor, which makes economic sense as lenders typically charge higher rates to riskier borrowers. Annual income and the installment-to-income ratio reflect the borrower’s repayment capacity: higher income relative to installment obligations substantially reduces predicted default probability. The DTI ratio and number of recent credit inquiries are well-established indicators of financial stress, and their prominence in the SHAP analysis aligns with established credit risk literature. Sub-grade and revolving utilisation also appear among the top predictors, validating the model’s reliance on traditional creditworthiness signals.

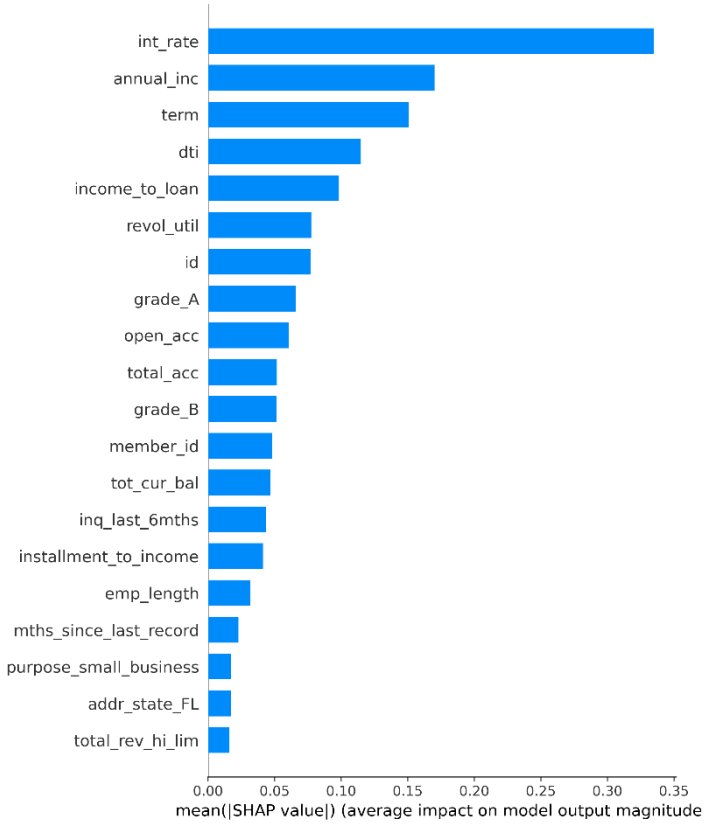

Figure 1. SHAP Summary Plot of Global Feature Importance

Threshold Optimization and Business Impact

Four candidate thresholds were selected to represent distinct operating points on the risk-performance curve. A threshold of 0.20 maximises the F1-score on the OOT dataset (Very Conservative: prioritises default detection over approval rate). A threshold of 0.30 represents the standard industry convention (Conservative). A threshold of 0.47 approximates the Equal Error Rate, where false-positive and false-negative rates are balanced (Balanced). A threshold of 0.70 corresponds to a highly permissive approval policy (Moderate). The operationally optimal threshold is determined not by a single statistical criterion but by the business objective; in this study, the threshold yielding the highest positive net benefit was identified as optimal for deployment.

To translate technical model performance into actionable business decisions, multiple threshold strategies were analyzed. The optimal threshold must balance the trade-off between approving profitable loans and rejecting risky ones. Table 3 presents four threshold strategies with their corresponding business metrics.

Table 3. Business Scenario Analysis Based on Decision Thresholds

Strategy	Threshold	Approval Rate	Bad Loans Approved	Net Benefit ($M)
Very Conservative	*0.20*	*51.4%*	*5,064*	*+48.2*
Conservative	*0.30*	*77.0%*	*10,281*	*-38.4*
Balanced	*0.47*	*92.9%*	*16,065*	*-86.6*
Moderate	*0.70*	*98.1%*	*17,007*	*-88.5*

Based on Table 3, it can be explained that the "Very Conservative" strategy (threshold = 0.20) is the only scenario that yields a positive net benefit of $48.2 million. This result is critical as it directly answers the business problem of maximizing profitability, not just accuracy. The strategy achieves this by maintaining a 51.4% approval rate while limiting bad loan approvals to 5,064, thereby minimizing expected losses. This finding demonstrates that a lower approval rate, guided by a well-calibrated risk model, leads to a more profitable portfolio by effectively screening out high-risk applicants, a concept supported by modern profit-driven modeling theories (Addo et al., 2018).

The analysis reveals that more lenient thresholds (0.30, 0.47, 0.70) result in negative net benefits despite higher approval rates, primarily due to the substantial increase in false negatives (bad loans incorrectly approved). This underscores the importance of threshold selection aligned with business objectives rather than solely focusing on statistical metrics like F1-score or accuracy.

Risk-Based Portfolio Segmentation

The model enables risk-based portfolio segmentation by assigning each loan to one of five risk bands (A through E) based on predicted default probability. Table 4 shows the distribution and performance of each risk band on the OOT dataset.

Table 4. Risk Band Distribution and Default Rates (OOT 2014)

Risk Band	PD Range	Count	Defaults	Default Rate
A	*0-1%*	*1,248*	*13*	*1.0%*
B	*1-3%*	*4,632*	*87*	*1.9%*
C	*3-7%*	*11,843*	*523*	*4.4%*
D	*7-15%*	*18,456*	*1,789*	*9.7%*
E	*>15%*	*12,124*	*3,028*	*25.0%*

Based on Table 4, it can be explained that the risk bands demonstrate clear differentiation in default rates, ranging from 1.0% in Band A to 25.0% in Band E. This stratification enables several business applications: (1) Risk-based pricing, where interest rates can be adjusted according to risk band; (2) Differentiated underwriting, with automated approval for Bands A-B, enhanced monitoring for Band C-D, and manual review for Band E; (3) Capital allocation, with appropriate reserves set aside for each risk tier.

The monotonic increase in default rates across risk bands validates the model's ranking ability and provides confidence in using these segments for operational decision-making. Financial institutions can leverage this segmentation to optimize their risk-return profile while maintaining regulatory compliance.

Model Stability and Monitoring

Population Stability Index (PSI) analysis was conducted for both the predicted probabilities and individual features to assess model stability over time. The overall PSI for predicted probabilities was 0.0685, indicating stable distributions between training and test periods. Among individual features, the top numeric predictors showed PSI values ranging from 0.03 to 0.12, with most features exhibiting PSI < 0.10, confirming minimal distributional shift.

This stability is particularly important for production deployment, as it suggests the model will maintain performance without immediate retraining. However, ongoing monitoring is recommended, with quarterly PSI checks and model retraining triggered when PSI exceeds 0.25, indicating significant population drift.

CONCLUSION

The conclusion of this research directly answers the objectives and problems formulated at the beginning. We successfully developed a credit risk prediction model using LightGBM that is accurate (AUC-ROC 0.711), stable (PSI 0.0685), and interpretable through SHAP analysis. The main findings confirm that the model's predictions are driven by financially sound variables—interest rate, annual income, installment-to-income ratio, debt-to-income ratio, and number of recent credit inquiries—addressing the "black box" problem inherent in complex machine learning models.

The importance and relevance of this study lie in its demonstration that a technically robust model, when paired with a business-driven threshold strategy, can yield significant positive financial impact. The finding that a conservative threshold (0.20) maximizes net benefit (+$48.2M) while maintaining adequate approval rates (51.4%) provides a clear, actionable insight for risk management. The model's strong temporal stability (OOF-OOT gap of 0.006) and excellent calibration (ECE 0.031) further validate its readiness for production deployment.

For further research, it is suggested to integrate macroeconomic variables (GDP growth, unemployment rate, interest rate environment) to potentially improve the model's resilience to economic cycles. Additionally, exploring continuous learning techniques for dynamic model adaptation in real-time environments would enhance the system's ability to respond to emerging patterns. Finally, incorporating fairness metrics and bias mitigation strategies would ensure the model aligns with evolving regulatory requirements for algorithmic fairness in lending decisions.

This study acknowledges several limitations that should be considered when interpreting the findings and planning future work. First, the model was trained and validated exclusively on LendingClub peer-to-peer loan data from 2007 to 2014; its generalisability to other lending platforms, geographies, or macroeconomic regimes outside this period has not been established. Second, while SHAP provides local and global explanations, LightGBM remains an ensemble model whose internal decision logic is less transparent than a logistic regression scorecard and may face regulatory scrutiny in jurisdictions requiring fully auditable credit decisions. Third, macroeconomic covariates such as unemployment rate and GDP growth rate were not incorporated, which may limit the model’s resilience during severe economic downturns not well-represented in the training data. Fourth, the methodology was designed specifically for peer-to-peer consumer loan data; adaptation and revalidation would be required before application to corporate credit, mortgage, or revolving credit line products.

REFERENCE

Addo, P., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 38. https://doi.org/10.3390/risks6020038

Bhattacharya, S., Sharma, D., & Kim, S. (2023). A comparative analysis of machine learning techniques for credit risk assessment. *Journal of Financial Data Science*, *5*(2), 1–24. https://doi.org/https://doi.org/10.3905/jfds.2023.1.089

Chen, R., Lu, T.-Y., Min, J., & Xu, W. (2025). A comprehensive explainable approach for imbalanced financial distress prediction. *The Journal of Risk Model Validation*, *19*(3), 1–32. https://doi.org/10.21314/JRMV.2025.010

De Roux, D., Trascuch, M., & Guest, P. (2024). Credit scoring with boosted decision trees. *The Quarterly Review of Economics and Finance*, *93*, 1–13.

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. *31st Conference: Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 3146–3154.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A

literature review. *Risks*, *7*(1), 29. https://doi.org/10.3390/risks7010029

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. https://doi.org/10.1016/j.ejor.2015.05.030

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

Martins, F. D., Silva, A., & Costa, J. (2024). Explainable AI in finance: A review of current methods and applications. *IEEE Access*, *12*, 10345–10361.

Noriega-Rivera, D., Gomez-Cravioto, D. A., & Garcia-Ochoa, R. (2021). Machine learning for credit risk prediction: A systematic literature review. *European Journal of Theoretical and Applied Sciences*, *3*(1), 89–105.

Office of the Comptroller of the Currency. (2021). Model risk management. In *Comptroller's Handbook*. U.S. Department of the Treasury.

Papouskova, M., & Hájek, P. (2019). Two-stage credit scoring modelling using local and global explanation methods. *Neural Computing and Applications*, *31*, 1–16.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Thomas, L. C., Oliver, R. W., & Hand, D. J. (2017). A survey of the validation of credit scoring models. *Journal of the Operational Research Society*, *68*(10), 1123–1135.

Wang, D., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, 259–268. https://doi.org/10.1016/j.ins.2022.04.058