Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bankruptcy Prediction via Hybrid Resampling and Stacking Ensemble Techniques with Explainable Artificial Intelligence (XAI)-Driven Analysis

Obu-Amoah Ampomah
Western Michigan University
Edmund Fosu Agyemang
Tulane University
Kofi Acheampong
Western Michigan University
Louis Agyekum
University of Ottawa
Enock Adu Bonsu
University of Arizona
Eric Nyarko
ericnyarko@ug.edu.gh
University of Ghana

Research Article

Keywords: Bankruptcy, Class imbalance, Explainable artificial intelligence, Financial distress

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11545501/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bankruptcy Prediction via Hybrid Resampling and Stacking Ensemble Techniques with Explainable Artificial Intelligence (XAI)-Driven Analysis

Abstract

Reliable bankruptcy prediction is essential for creditors, investors, auditors, and regulators because delayed identification of financial distress can result in substantial economic losses. However, bankruptcy is a rare event, and severe class imbalance can produce high overall accuracy while concealing poor detection of distressed firms. This study develops an integrated bankruptcy prediction framework using the Taiwanese Bankruptcy Prediction dataset. Five boosting classifiers and five deep learning models were evaluated as standalone models. A two-stage stacking procedure subsequently used five-fold out-of-fold predictions from the boosting models as inputs to the deep learning meta-learners. The results show that resampling strategy materially shaped model behavior. SVM-SMOTE and SMOTE-Tomek favored accuracy and specificity, whereas SMOTE-ENN delivered stronger minority-class detection. Among standalone models, the RNN with SVM-SMOTE achieved the best overall predictive balance, with recall of 0.8235, G-mean of 0.8639, and ROC-AUC of 0.9417. Among stacking ensembles, SMOTE-ENN with (GB+XGB+HGB+LGBM+AB) + LSTM provided the best results with recall of 0.7255, G-mean of 0.8254, and ROC-AUC of 0.9270. SHAP analysis further showed that predicted bankruptcy risk was driven primarily by profitability, solvency and capital structure, debt-servicing burden, and operational efficiency and cost pressure. The findings demonstrate that no single resampling-model configuration universally dominates; therefore, model selection should reflect whether the operational priority is overall classification performance or the detection of bankrupt firms. The proposed framework provides an interpretable and leakage-controlled approach for aligning bankruptcy prediction with different financial risk-management objectives.

Keywords: Bankruptcy, Class imbalance, Explainable artificial intelligence, Financial distress

1 Introduction

The expanding availability of financial statements and narrative disclosures has intensified the need for accurate bankruptcy predictions for creditors, investors, auditors, and regulators [1, 2]. In applied settings, bankruptcy prediction models estimate firm-level failure risk over future horizons, yet they must operate under severe class imbalance, heterogeneous firm populations, limited observations of actual failure, and instability across industries and macroeconomic conditions. These features make bankruptcy prediction materially different from ordinary static classification, because bankruptcy is a rare event, predictor distributions shift over time, and the evolution of a firm's condition often contains information that one-period cross-sectional models do not capture [3–5]. Accordingly, the recent works treat bankruptcy prediction as a cost-sensitive decision problem in which failing to identify a genuinely distressed firm may be more consequential than incorrectly flagging a healthy one; therefore, model evaluation should emphasize recall, Geometric (G)-mean, area under the curve (AUC), and class-specific error rates rather than raw accuracy alone [6–8]. Recent evidence further shows that predictive performance can be improved by moving beyond traditional ratio-based inputs to incorporate annual-report language, risk-related disclosures, and other unstructured textual signals, while maintaining interpretability for practical decision use [9, 10].

Bankruptcy events are rare relative to non-bankruptcy events, which can lead to trivial but misleading performance if evaluation relies on accuracy alone. For highly imbalanced settings, the precision recall (PR) curve and PR AUC are frequently recommended because they focus on performance on the positive class and can better reflect the practical tradeoff between catching financial distress and limiting false alarms [11, 12]. This evaluation perspective has direct methodological consequences. For instance, decision thresholds should be tuned to operational capacity, and training procedures often incorporate class weighting or sampling strategies to increase sensitivity to the minority class without inflating false positives. Synthetic oversampling methods such as Synthetic Minority Oversampling Technique (SMOTE) are widely used in imbalanced learning to augment minority class examples, although careful validation design is required to avoid leakage in time-dependent financial distress streams [13, 14].

Bankruptcy risk signals evolve over time as firms move through changing liquidity, leverage, profitability, and cash flow conditions, while shifts in macroeconomic and industry environments can alter the relationship between predictors and failure outcomes. Consequently, the bankruptcy prediction literature gives considerable attention to temporal stability, dynamic or panel-based modeling, and out-of-time evaluation, since models estimated in one period may generalize poorly when applied to another period with different economic conditions [15, 16]. In practice, bankruptcy risk can be inferred from static accounting and market indicators, multi-period trajectories of financial deterioration, and narrative disclosures in annual reports. Although conventional tabular models remain useful, recent studies show that predictive performance improves when models incorporate textual and other multimodal information, because single-period financial ratios alone provide only a partial view of a firm's evolving condition [9, 10, 17]. This motivates hybrid modeling, where complementary learners target different data modalities and are fused to improve robustness [18, 19].

A host of machine learning (ML) models have been employed for bankruptcy detection. Tree-based ensembles remain a dominant baseline because they capture nonlinear feature interactions, handle mixed variable types, and offer strong performance with limited preprocessing. Gradient boosting machines formalize an additive ensemble framework that incrementally improves predictive performance by fitting weak learners to residual errors [20]. Practical implementations such as XGBoost (Extreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) introduced scalable training and engineering optimizations that enable strong performance on large datasets, supporting their widespread adoption in applied data analytics [21, 22]. Beyond single-family ensembles, stacked generalization (stacking) combines multiple base learners by training a meta learner on their outputs, aiming to leverage complementary inductive biases. This is particularly relevant for bankruptcy detection because different models can excel under different regimes, such as sparse high-cardinality tabular features versus temporal patterns. Stacking provides a principled framework for hybridization when diverse learners capture different aspects of the underlying signal [23].

Deep learning (DL) has shown promise in bankruptcy prediction when the objective is to learn from the temporal progression of a firm's financial condition rather than from isolated one-period snapshots. Recent evidence indicates that these sequence-aware deep learning approaches can outperform conventional baselines in some settings, especially when multiple periods of firm information are available and the prediction task is explicitly framed in temporal terms [4, 24]. More broadly, recent review evidence identifies deep, hybrid, and time-aware models as an important direction in contemporary bankruptcy prediction research, while also noting that their performance depends on data design, class imbalance treatment, and appropriate temporal validation procedures. This line of work supports a broader conclusion: in many contexts, predictive performance depends on how effectively the model captures short-term deviations from a customer's typical behavior and longer-term patterns of spending. Such dependencies are difficult to represent with single transaction features alone [25, 26].

A hybrid ensemble framework is motivated by the observation that no single modeling family consistently dominates across all bankruptcy

conditions. The survey on bankruptcy prediction emphasizes the diversity of challenges and methods, including handling imbalance, drift, and sequential properties, suggesting that integrated solutions may be more robust than single-model approaches [4, 7, 27, 28]. In a hybrid ensemble, model fusion can be implemented through stacking, weighted averaging, or two-stage pipelines where deep models produce representations or risk scores that become inputs to an ensemble meta-learner or otherwise. Stacking is particularly suitable for combining heterogeneous learners because it learns how to weight base models under the validation distribution rather than assuming a fixed combination rule [29]. However, hybridization increases governance complexity. As model pipelines become more layered, institutions face challenges in calibration, monitoring under drift, and communicating rationales to investigators. This directly motivates the integration of explainable artificial intelligence (XAI)-driven analysis as a first-class component of the framework rather than an afterthought.

In bankruptcy detection, model output often triggers consequential actions such as declines, step-up authentication, or account restrictions. This makes transparency and interpretability important for auditability, internal governance, and effective human investigation. A prominent perspective argues that relying on black box models with post hoc explanations can create fragile decision pathways in high-stakes contexts, motivating careful design choices regarding interpretability and explanation use [30]. Systematic reviews of explainable AI in finance document rapid growth in XAI adoption across financial tasks, reflecting the need to align predictive models with transparency expectations and stakeholder trust [31]. Two widely adopted post hoc explanation approaches are LIME and SHAP. LIME explains individual predictions by fitting a local surrogate model around a specific instance, providing an interpretable approximation of local behavior [32]. SHAP provides a unified additive feature attribution framework grounded in Shapley values and is frequently used to explain tree ensembles and other complex predictors through consistent feature contribution estimates [33]. In bankruptcy analytics, these tools can support several practical needs: case-level narratives for investigators, global feature importance to validate domain plausibility, and monitoring of explanation stability over time. At the same time, the literature cautions that explanation methods should be assessed for faithfulness and stability, particularly under dataset shift.

Prior studies often emphasize predictive performance on static datasets or focus on a single modeling family, while comparatively fewer works present an end-to-end framework that simultaneously fuses complementary machine learning and deep learning signals [34], evaluates performance using imbalance-appropriate metrics aligned with operational constraints, and produces explanations that remain consistent and actionable as bankruptcy behavior changes over time. Recent bankruptcy prediction research has increasingly combined ensemble learning, imbalance correction, deep learning, and explainable artificial intelligence. However, several limitations remain. [35] found that hybrid sampling and ensemble methods generally performed well across multiple bankruptcy datasets, although model rankings varied considerably across countries, predictor structures, and imbalance levels. Recent SHAP-based and deep learning studies have improved predictive interpretation but have frequently evaluated a limited range of resampling strategies or modeling families. Studies published in 2026 including [36–38] have extended the field through cost-sensitive threshold optimization, knowledge distillation, and class-rebalanced semi-supervised learning. These studies also show that predictive rankings depend on the validation protocol, evaluation metric, computational requirements, and dataset characteristics. Comparatively less attention has been given to the controlled integration of consensus feature selection, multiple hybrid resampling techniques, leakage-controlled stacking, imbalance-sensitive evaluation, and model interpretation within a common experimental framework. Moreover, limited external and out-of-time validation continues to constrain conclusions regarding generalizability.

Based on these gaps, the study addresses three scientific questions. First, how do SVM-SMOTE, SMOTE-Tomek, and SMOTE-ENN affect minority-class detection and the trade-off

between recall and specificity? Second, does out-of-fold stacking of boosting-based probability estimates with nonlinear meta-learners improve performance relative to the corresponding standalone models? Third, which financial indicators consistently influence the predictions of the final selected machine learning and deep learning models?

Unlike studies that examine a single resampling strategy, modeling family, or explanation technique, the present study evaluates these components jointly under a common data partition and evaluation framework. It compares three hybrid resampling techniques across five boosting and five deep learning models and further examines five two-stage stacking configurations based on out-of-fold base-learner probabilities. Consensus feature selection is used to identify predictors supported by multiple selection algorithms, while SHAP is applied to the held-out test observations for the final selected models. The study therefore contributes an integrated empirical framework for examining the interactions among feature selection, class-imbalance correction, model combination, and interpretability. The contribution concerns the integration and controlled comparison of established methods rather than the introduction of a new learning algorithm.

The remainder of the article is structured as follows. Section 2 describes the dataset, preprocessing procedures, feature-selection criterion, resampling methods, prediction models, stacking procedure, SHAP analysis, and evaluation metrics. Section 3 presents the empirical findings for the standalone and stacking models. Section 4 discusses the findings in relation to recent research, practical implications, and study limitations. Section 5 provides the conclusions and recommendations for future research.

2 Data and Methods

2.1 Data Description

The Taiwanese Bankruptcy Prediction dataset is a multivariate business dataset designed for binary classification of firm bankruptcy status. The data were collected from the Taiwan Economic Journal and span the period 1999 to 2009, with bankruptcy defined according to the business regulations of the Taiwan Stock Exchange. The dataset contains 6,819 firm records and 95 predictive features, with the response variable labeled "Bankrupt". The predictors are primarily financial ratio type variables (for example, profitability, leverage, liquidity, turnover, and growth-related indicators), as reflected in the dataset's variable list. The UC Irvine (UCI) documentation reports no missing values. A salient characteristic of this dataset is substantial class imbalance: published analyses report approximately 3.226% bankrupt and 96.774% non-bankrupt firms, indicating that evaluation should emphasize imbalance-aware metrics (for example, Receiver Operating Characteristic curve (ROC)-AUC, recall, and F-measures) rather than accuracy alone. The dataset is distributed via the UCI Machine Learning Repository with DOI 10.24432/C5004D under a CC BY 4.0 license as provided in and may be assessed via https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction. No missing values were observed in the dataset.

Table 1 presents an overview of recent studies which employed ML and DL models for bankruptcy prediction tasks.

2.2 Data Preprocessing and Method of Analysis

Data preprocessing is a critical stage in ML workflows. To ensure data quality, duplicate observations were identified and removed. Preprocessing comprised the encoding of categorical target, the scaling of continuous variables, and the application of class-balancing procedures. The target variable, Bankrupt, was transformed into a numerical format using *LabelEncoder* from *sklearn.preprocessing*. The predictor variables were normalized using *MinMaxScaler* to rescale the feature values into the range of 0 to 1. The *scaler* was fitted on the training data and subsequently applied to both the training and test sets to ensure that all variables were measured on a comparable scale. This preprocessing step is particularly important for ML and DL algorithms that are sensitive to differences in feature magnitude, as it helps improve numerical stability and supports more reliable model training and evaluation.

Five feature selection techniques presented in Table 1 were applied to identify the most informative predictors for subsequent modelling.

Table 1 Overview of recent studies on bankruptcy prediction using ML and DL models

Author, Year	Methods	Number of Variables	Sample Size	Interpretable Methods	Major Findings
Du Jardin, [39], 2021	Biclustering, Neural Network-based Ensembles	44	700 bankrupt, 34,300 non-bankrupt firms	None	The ensemble approach that integrates biclustering with neural networks enhanced prediction accuracy.
Kim et al., [28], 2022	RNN, LSTM and Ensemble	8	454,752 firm-month observations, 2057 of which are a firm's bankruptcy	None	Compared with traditional techniques, RNN and LSTM methods showed superior performance in handling sequential financial data.
Jabeur & Serret, [40], 2023	Fuzzy set qualitative comparative analysis (fsQCA) and convolutional neural networks (CNN)	17	133 bankrupt, 133 non-bankrupt firms	None	Fuzzy CNN led to better performance than when using traditional methods.
Mattos et al., [41], 2024	Logistic Regression, SVM, AdaBoost, XGBoost, Bagging, and Random Forest	26	503 private firms	Tree-based feature importance; permutation importance	Machine learning models outperformed logistic regression in predicting reorganization outcomes. Financial ratios were less informative under low-quality accounting information, while institutional factors and financial-information-quality proxies were more influential.
Nguyen et al., [42], 2025	Logistic Regression, Discriminant Analysis, SVM, Neural Network, Random Forest, LightGBM, XGBoost, and NGBoost	24	4,054 bankrupt firms and 5,600 non-failed firms	SHAP	XGBoost achieved the best predictive performance across the 1-to-5-year horizons, with ensemble models generally outperforming traditional models. SHAP identified the tax ratio, interest level, debtor days as among the most influential predictors of bankruptcy.
Lin et al., [43], 2025	LASSO feature selection, CNN	35	2,915 bankrupt and 15,121 non-bankrupt firms	SHAP; LIME	The CNN model based on grayscale financial ratio images outperformed traditional bankruptcy prediction methods, while SHAP and LIME improved interpretability by identifying influential financial indicators.

Specifically, the study employed metaheuristic, nature-inspired optimization algorithms that iteratively explore the feature space and converge toward near-optimal subsets under a predefined objective function (for example, maximizing predictive performance while minimizing the number of selected variables). Feature selection is essential because removing irrelevant or redundant features can enhance classifier generalization, reduce overfitting, improve computational efficiency, and increase interpretability [44]. Randomized search cross validation was employed to identify the optimal hyperparameter combinations because it offers an efficient and robust alternative to exhaustive grid search, particularly when the search space is large and computationally demanding. This approach enables a broader exploration of candidate parameter settings within a limited computational budget, while cross validation helps ensure that the selected configuration generalizes well and reduces the risk of overfitting [45]. These optimal hyperparameters were used in training the ML models employed in the study.

Each component was selected to address a specific characteristic of the dataset. Consensus feature selection was used to reduce the instability and redundancy associated with the original 95 financial indicators. SVM-SMOTE was selected to generate minority observations near the class boundary, whereas SMOTE-Tomek and SMOTE-ENN combine synthetic oversampling with different levels of neighborhood cleaning. Boosting algorithms were selected because of their established effectiveness in capturing nonlinearities and interactions in structured financial data. The deep meta-learners were used to examine whether nonlinear combinations of the base-learner probabilities improved prediction beyond individual models. SHAP was incorporated to determine whether the resulting predictions were supported by economically meaningful financial indicators.

The bankruptcy dataset was partitioned into training and testing subsets using an 80:20 ratio. A severe class imbalance was observed, which may introduce bias in the predictions because the models can become predisposed toward the majority class [13, 14]. To address this issue, three hybrid resampling techniques were adopted: SVM-SMOTE [46], SMOTE-Tomek [14] and SMOTE-Edited Nearest Neighbors (ENN)[47]. In this study, five (5) ensemble learning models: gradient boosting (GB), extreme gradient boosting (XGB), Histogram-based gradient boosting (HGB), light extreme gradient machines (LGBM), and adaptive boosting (AB) were employed as base learners. Additionally, five (5) DL models, including recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent units (GRU), deep neural networks (DNN), and multi-layer perceptron (MLP), were employed as meta-learners.

The dataset adopted for the study is cross-sectional and its predictors have no natural temporal ordering. DNN and MLP were implemented as feed-forward networks, whereas RNN, LSTM, and GRU were retained only as exploratory nonlinear comparators, with no temporal interpretation assigned to the reshaped inputs.

The stacking procedure was implemented using five-fold stratified cross-validation. For each boosting model, out-of-fold positive-class probabilities were generated for the training observations, ensuring that the prediction for each observation came from a model that had not been trained on that observation. The five probability columns generated by GB, XGB, HGB, LGBM, and AB were combined to form the meta-training matrix. Each base learner was subsequently fitted to the complete training dataset and used to generate the corresponding probability columns for the held-out test set. The deep learning meta-learners were trained on the out-of-fold meta-training matrix and evaluated using the test meta-features. The test data were not used in either stage of model training.

SHAP analysis was conducted after model evaluation and was restricted to the final selected models. These models were selected based on their overall balance across recall, specificity, G-mean, and ROC-AUC. SHAP values were computed for observations in the held-out test set, while model training and resampling were restricted to the training data. Global feature importance was calculated as the mean absolute SHAP value across test observations, whereas the signed SHAP distributions indicated whether each feature increased or decreased the predicted likelihood of bankruptcy [48, 49].

2.3 Model Performance Evaluation Metrics

All ten (10) models (ML and DL) were evaluated mainly using the test accuracy, recall, specificity, area under the receiver operating curve (ROC-AUC) and G-mean.

The computational formulas are given respectively by (1)-(5):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{1}$$

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{3}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}\,(\text{FPR})\,d\,(\text{FPR}) \tag{4}$$

$$\text{G-mean} = (\text{Recall} \times \text{Specificity})^{\frac{1}{2}} \tag{5}$$

The empirical workflow adopted for the study is given in Figure 1

Algorithm 1 summarizes the complete hybrid-resampling and two-stage stacking procedure. In Stage 1, five boosting base learners generate five-fold out-of-fold probabilities for the resampled training data and full-fit probabilities for the untouched test data. In Stage 2, the resulting five-dimensional probability vectors are used to train and evaluate the RNN, LSTM, GRU, DNN, and MLP meta-learners.

Note: The procedure is executed separately for SVM-SMOTE, SMOTE-Tomek, and SMOTE-ENN. Only the training data are resampled; the test data remain unchanged throughout model selection, meta-feature generation, and final evaluation.

For a given resampling method, each base learner contributes one positive-class probability to the meta-feature space. Thus, the meta-training matrix Z_train has five columns, corresponding to GB, XGB, HGB, LGBM, and AB. Its entries are generated out of fold by five-fold stratified cross-validation, whereas Z_test is obtained from the same base learners after fitting them on the complete resampled training set. Each deep-learning meta-learner is then fitted independently on Z_train, producing five alternative 5+1 stacking configurations for each resampling method. This construction prevents direct in-sample base-learner probabilities from being used as meta-training inputs and preserves the original test set for final performance estimation.

2.4 Feature Selection

Due to the large number of features (95), feature selection techniques were employed to select the best features for the analysis. Table 2 lists the number of features selected by the 5 different feature selection algorithms.

2.4.1 Consensus-Based Feature Retention Criterion

To promote a predictor set that is both robust and not tied to any single algorithm, a consensus criterion was applied across the five feature selection approaches (Figure 2). The premise is that variables identified repeatedly by multiple, conceptually distinct methods are less likely to reflect idiosyncrasies of a particular scoring rule or search strategy and more likely to capture a stable underlying signal. Accordingly, only predictors exhibiting substantial cross method agreement were retained for subsequent modelling and substantive interpretation

A predictor was retained only when it was identified by at least two of the five feature selection algorithms. In doing so, the procedure supports both parsimony and robustness by reducing the feature space while emphasizing predictors that remain consistently informative across diverse selection frameworks.

From Figure 2, using the consensus-based feature retention criterion, 23 features (frequency $\geq$ 2) namely Net worth/Assets; ROA(C) before interest and depreciation before interest; Retained Earnings to Total Assets; Net Income to Total Assets; Net profit before tax/Paid-in capital; Persistent EPS in the Last Four Seasons; Current Asset Turnover Rate; Equity to Liability; Gross Profit to Sales; Operating Gross Margin; Operating Profit Growth Rate; Operating profit/Paid-in capital; Total expense/Assets; Average Collection Days; Cash Flow to Total Assets; Debt ratio %; Interest Coverage Ratio (Interest expense to EBIT); Inventory/Working Capital; Net Income to Stockholder's Equity; Operating Expense Rate; Per Share Net profit before tax (Yuan ¥); Total

Table 2 Feature Selection Results using the 5 feature selection algorithms

Feature selection method used	Number of features chosen	Features chosen	Validation ROC-AUC
Whale Optimization (WOA) [50]	7	ROA(C) before interest and depreciation before interest, Operating Gross Margin, Persistent EPS in the Last Four Seasons, Quick Ratio, Net worth/Assets, Retained Earnings to Total Assets, Current Asset Turnover Rate	0.9112
Particle Swarm Optimization (PSO) [51]	20	ROA(C) before interest and depreciation before interest, Operating Gross Margin, Operating Expense Rate, Interest-bearing debt interest rate, Persistent EPS in the Last Four Seasons, Operating Profit Growth Rate, Total debt/Total net worth, Net worth/Assets, Operating profit/Paid-in capital, Net profit before tax/Paid-in capital, Average Collection Days, Retained Earnings to Total Assets, Total expense/Assets, Current Asset Turnover Rate, Cash Turnover Rate, Cash Flow to Total Assets, Net Income to Total Assets, Total assets to GNP price, Gross Profit to Sales, Equity to Liability	0.9140
Salp Swarm Optimization (SSO) [52]	20	ROA(C) before interest and depreciation before interest, Realized Sales Gross Margin, Operating Expense Rate, Per Share Net profit before tax (Yuan ¥), Operating Profit Growth Rate, Current Ratio, Debt ratio %, Net worth/Assets, Operating profit/Paid-in capital, Net profit before tax/Paid-in capital, Total Asset Turnover, Inventory/Working Capital, Retained Earnings to Total Assets, Total expense/Assets, Net Income to Total Assets, Total assets to GNP price, Gross Profit to Sales, Net Income to Stockholder's Equity, Interest Coverage Ratio (Interest expense to EBIT), Equity to Liability	0.9004
Bat Algorithm (BA) [53]	20	ROA(C) before interest and depreciation before interest, Operating Gross Margin, Net Value Per Share (A), Persistent EPS in the Last Four Seasons, Operating Profit Growth Rate, Net worth/Assets, Operating profit/Paid-in capital, Net profit before tax/Paid-in capital, Total Asset Turnover, Accounts Receivable Turnover, Average Collection Days, Current Liability to Assets, Inventory/Working Capital, Retained Earnings to Total Assets, Total expense/Assets, Current Asset Turnover Rate, Cash Flow to Total Assets, Net Income to Total Assets, Gross Profit to Sales, Net Income Flag	0.9022
Mutual Information (MI) [54]	20	ROA(C) before interest and depreciation before interest, ROA(A) before interest and % after tax, Pre-tax net Interest Rate, Continuous interest rate (after tax), Net Value Per Share (B), Persistent EPS in the Last Four Seasons, Per Share Net profit before tax (Yuan ¥), Interest Expense Ratio, Debt ratio %, Net worth/Assets, Borrowing dependency, Net profit before tax/Paid-in capital, Retained Earnings to Total Assets, Total income/Total expense, Net Income to Total Assets, Net Income to Stockholder's Equity, Liability to Equity, Degree of Financial Leverage (DFL), Interest Coverage Ratio (Interest expense to EBIT), Equity to Liability	0.5417

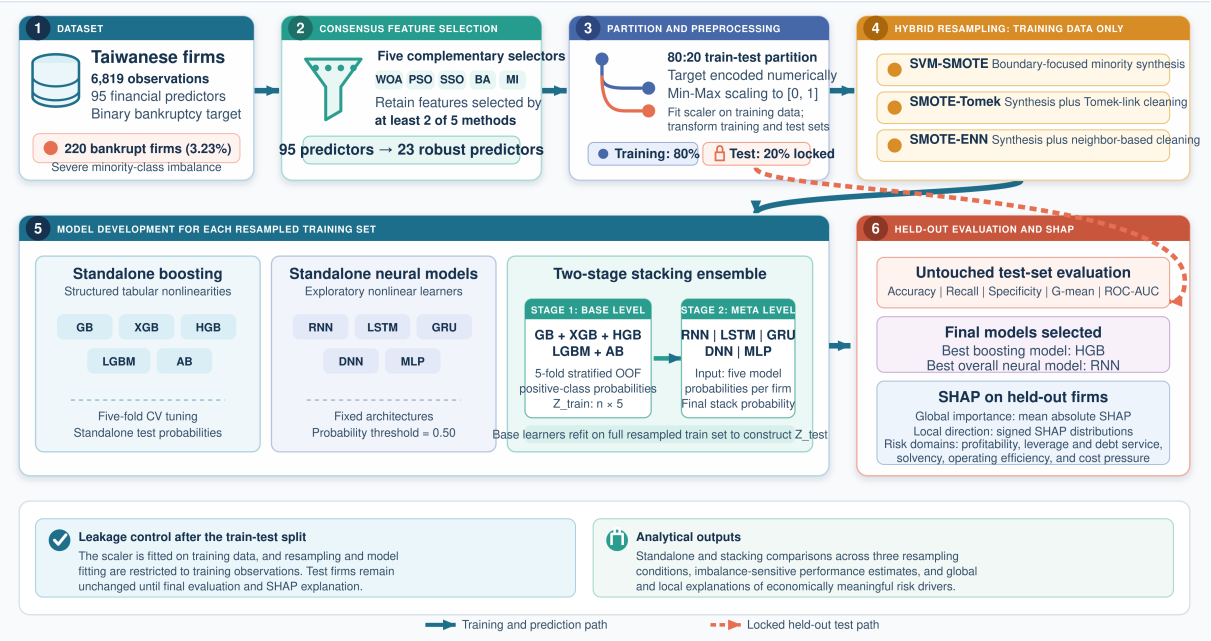


Fig. 1 Empirical workflow adopted for the study

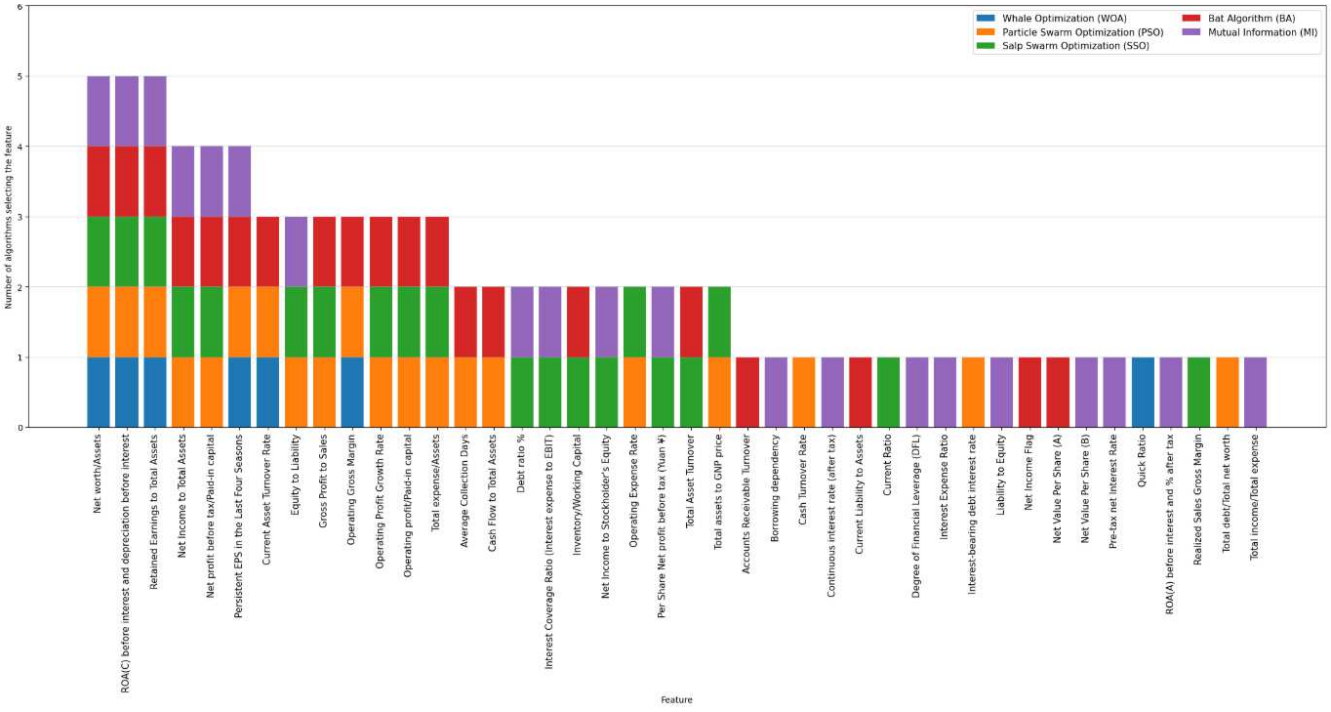


Fig. 2 Feature selection frequency across algorithms (stacked contributions)

Asset Turnover and Total assets to GNP price were retained for the analysis.

Algorithm 1 Hybrid-Resampling-Based Two-Stage Stacking Ensemble Training and Evaluation

Require: $X_{\text{train_scaled}}, X_{\text{test_scaled}}, y_{\text{train}}, y_{\text{test}}$; $\mathcal{R} = \{\text{SVM-SMOTE}, \text{SMOTE-Tomek}, \text{SMOTE-ENN}\}$; $\mathcal{B} = \{\text{GB}, \text{XGB}, \text{HGB}, \text{LGBM}, \text{AB}\}$; $\mathcal{H} = \{\text{RNN}, \text{LSTM}, \text{GRU}, \text{DNN}, \text{MLP}\}$; $K = 5$

Ensure: Fitted 5+1 stacking models and a test-set performance table

Initialize *Results* as an empty collection
for each resampling method $r \in \mathcal{R}$ **do**
  Apply r only to the scaled training data
  $(X_r, y_r) \leftarrow r.\text{fit_resample}(X_{\text{train_scaled}}, y_{\text{train}})$
  Create a five-fold stratified splitter with shuffling and `random_state = 42`
  Initialize Z_{train} and Z_{test} as empty meta-feature matrices
  Stage 1: Base-level learning and meta-feature construction
  for each base learner $b_m \in \mathcal{B}$ **do**
    Load the selected base-learner configuration obtained from the model-tuning phase
    $z_{\text{train},m} \leftarrow \text{cross_val_predict}(b_m, X_r, y_r, cv = K, method = \text{predict_proba})[:, 1]$
    Fit b_m on the complete resampled training set (X_r, y_r)
    $z_{\text{test},m} \leftarrow b_m.\text{predict_proba}(X_{\text{test_scaled}})[:, 1]$
    Append $z_{\text{train},m}$ to Z_{train} and $z_{\text{test},m}$ to Z_{test}
    Optionally retain the standalone test metrics for b_m for comparison
  end for
  $Z_{\text{train}} \leftarrow \text{column_stack}(z_{\text{train},1}, \ldots, z_{\text{train},5})$
  $Z_{\text{test}} \leftarrow \text{column_stack}(z_{\text{test},1}, \ldots, z_{\text{test},5})$
  Stage 2: Deep-learning meta-level training and evaluation
  for each meta-learner $h_l \in \mathcal{H}$ **do**
    Initialize h_l with `input_dim = 5`, `epochs = 30`, `batch_size = 32`, and `random_state = 42`
    For RNN, LSTM, and GRU, reshape Z to $(\text{samples}, 5, 1)$; retain two-dimensional Z for DNN and MLP
    Fit h_l on (Z_{train}, y_r)
    $p_l \leftarrow h_l.\text{predict_proba}(Z_{\text{test}})[:, 1]$
    $\widehat{y}_l \leftarrow 1$ if $p_l \geq 0.50$; otherwise $\widehat{y}_l \leftarrow 0$
    Compute Accuracy, Precision, Recall, Specificity, F1-score, G-mean, and ROC-AUC
    Store $(\text{GB} + \text{XGB} + \text{HGB} + \text{LGBM} + \text{AB}) + h_l$ and its test metrics in *Results*
  end for
  Rank the stacked models by ROC-AUC, G-mean, F1-score, and Accuracy
end for
return *Results*

3 Results and Findings

Figure 3 presents the bankruptcy class distribution before and after the application of hybrid resampling techniques. The original dataset was highly imbalanced, with 6,599 non-bankruptcy cases (96.77%) and only 220 bankruptcy cases (3.23%), as shown in Figure 3(a). Under such conditions, pure undersampling would likely be unsuitable because it would discard a large portion of the majority class, potentially resulting in information loss and reduced generalizability. To address this issue, hybrid oversampling methods were applied to improve minority-class representation while retaining as much useful information as possible. Figures 3(b) and 3(c) show that SVM-SMOTE and SMOTE-Tomek achieved balanced class distributions, with 5,286 and 5,279 instances per class, respectively. By contrast, SMOTE-ENN produced a slightly uneven distribution of 4,662 non-bankruptcy and 5,234 bankruptcy cases (as shown in Figure 3(d)), reflecting the effect of ENN in removing noisy or misclassified samples after resampling. These findings indicate that the adopted hybrid resampling methods effectively mitigated the severe class imbalance and produced training distributions that are more suitable for robust bankruptcy prediction.

Although the dataset contains 6,819 firms, only 220 observations represent bankrupt firms. The effective sample size for learning the minority class is therefore limited. Resampling was applied only to the training data to reduce majority-class bias, while evaluation was conducted on the original test data.

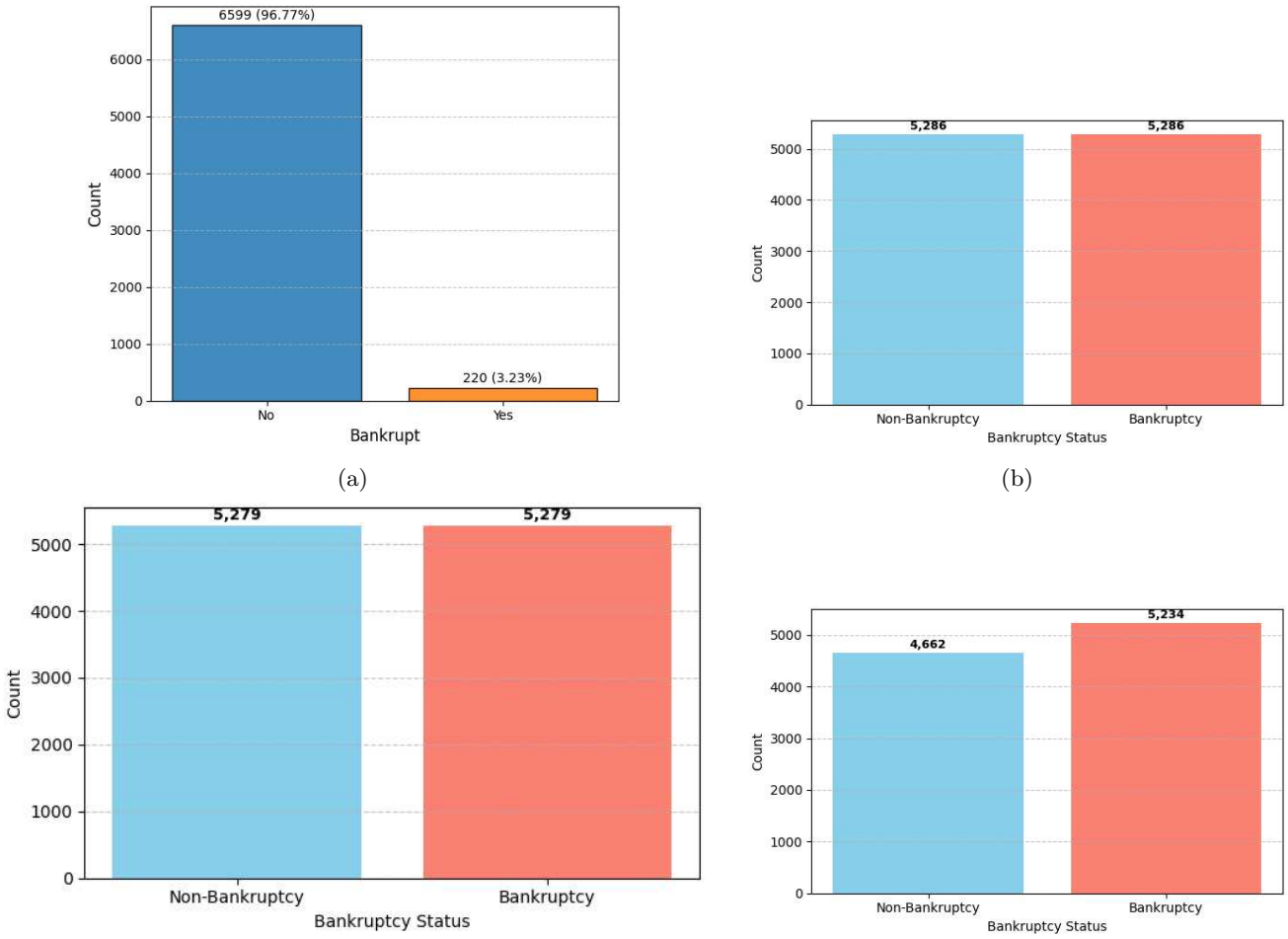


Fig. 3 Bar charts showing bankruptcy data counts before and after application of hybrid resampling techniques. (a) Bankruptcy class distribution of original imbalanced dataset (b) Bankruptcy class distribution after SVM-SMOTE (c) Bankruptcy class distribution after SMOTE–Tomek (d) Bankruptcy class distribution after SMOTE–ENN.

3.1 Rationale for ML and DL hyperparameter tuning

In Table 3, the optimized hyperparameter settings were derived from a targeted search informed by previous studies, best practices in model development, and the need to balance predictive accuracy with generalization. Across the boosting-based classifiers, the tuning process focused primarily on parameters that influence learning pace, ensemble strength, and tree complexity. The number of estimators and learning rate were varied to identify combinations that allowed the models to learn effectively without becoming overly sensitive to the training data. Similarly, depth- and split-related parameters were adjusted to reduce the likelihood of overfitting while preserving the ability to capture meaningful patterns in the data. The results demonstrate that all tuned ensemble models performed strongly, although Extreme Gradient Boosting produced the highest cross-validation ROC-AUC during hyperparameter tuning, with a CV ROC-AUC of 0.9349. This was followed closely by Histogram-based Gradient Boosting (0.9336) and Gradient Boosting (0.9324), which also showed competitive performance. These findings indicate that ensemble boosting approaches were particularly well suited to the classification task addressed in this study. The deep learning architectures were fixed a priori to maintain consistent model capacity across

the three resampling conditions and to isolate the effects of resampling and stacking. Parsimonious configurations were used because the dataset contained relatively few bankruptcy observations and extensive architecture tuning would substantially increase computational cost and the risk of selection instability. All models therefore used 30 epochs, a batch size of 32, the Adam optimizer, and consistent dropout regularization. The objective was to provide controlled comparisons using representative DL architectures.

3.2 Results and findings of ML and DL models under the three resampling techniques

In Table 4, the SVM-SMOTE resampling technique produced distinct performance differences between machine learning and deep learning models. The machine learning models generally achieved higher accuracy and specificity, with LGBM recording the best accuracy (0.9545) and specificity (0.9726), indicating stronger classification of non-bankrupt firms. However, the deep learning models showed better sensitivity to the minority class. RNN achieved the highest recall (0.8235), G-mean (0.8639) and ROC-AUC (0.9417), while HGB recorded the highest ROC-AUC (0.9408) and highest recall (0.7451) among the ML models. The MCC values of all ML and DL models under the SVM-SMOTE framework have moderate positive correlation; acceptable but room for improvement. These results indicate that, although some ML models were more effective in overall classification, the DL models provided better balance in detecting bankrupt firms, which is more important in imbalanced bankruptcy prediction.

Figure 4 presents the ROC and AUC of ML and DL models with SVM-SMOTE technique.

The results reported in Table 5 show that the SMOTE-Tomek resampling method produced noticeable differences in the predictive behavior of the ML and DL models used for bankruptcy prediction. The ML models generally achieved higher accuracy and specificity, with LGBM producing the best accuracy (0.9494) and specificity (0.9673), indicating strong performance in identifying non-bankrupt firms. HGB recorded the highest ROC-AUC (0.9274) and highest recall (0.7843) among the ML models. However, the DL models showed stronger recall, G-mean, and ROC-AUC, suggesting better capability in detecting bankrupt firms and maintaining balanced predictive performance. Notably, MLP achieved the highest recall (0.8627), while RNN recorded the highest G-mean (0.8570) and ROC-AUC (0.9382). The MCC values of all ML and DL models under the SMOTE-Tomek framework have moderate positive correlation; acceptable but room for improvement. These findings imply that, although the ML models were more effective in overall classification accuracy, the DL models were more suitable for bankruptcy prediction under SMOTE-Tomek because they better captured the minority bankrupt class, which is more critical in imbalanced financial datasets.

Figure 5 presents the resulting ROC and AUC of ML and DL models after resampling with SMOTE-Tomek technique.

The findings in Table 6 show that the SMOTE-ENN resampling technique affected ML and DL models differently in bankruptcy prediction. The ML models generally achieved higher accuracy and specificity, with LGBM producing the best overall accuracy (0.9289) and specificity (0.9383), indicating strong performance in identifying non-bankrupt firms. HGB recorded the highest ROC-AUC (0.9260) and highest recall (0.8039) among the ML models. However, the DL models outperformed the ML models in recall, G-mean, and ROC-AUC, suggesting greater effectiveness in detecting bankrupt firms and maintaining balanced predictive performance. Notably, LSTM achieved the highest recall (0.9216), MLP produced the highest G-mean (0.8563), and DNN recorded the best ROC-AUC (0.9397). The MCC values of all ML and DL models under the SMOTE-ENN framework have moderate positive correlation; acceptable but room for improvement. These results suggest that, although ML models were stronger in overall classification accuracy, DL models were more suitable for bankruptcy prediction under SMOTE-ENN because they provided better minority-class detection, which is more critical in imbalanced financial distress datasets.

Figure 6 presents the resulting ROC and AUC of ML and DL models after resampling with SMOTE-ENN technique.

Table 3 Optimized configuration and optimal value of hyperparameters for ML models

Models	Hyperparameter	Testing Range	Optimal Value	Best CV ROC-AUC
Gradient Boosting	n_estimators	100, 200, 400	200	0.9324
	learning_rate	0.01, 0.05, 0.1	0.1	
	max_depth	2, 3, 4	4	
	min_samples_split	2, 5, 10	10	
	min_samples_leaf	1, 2, 4	2	
Extreme Gradient Boosting	n_estimators	200, 400, 800	400	0.9349
	learning_rate	0.01, 0.05, 0.1	0.1	
	max_depth	3, 5, 7	5	
	min_child_weight	1, 5, 10	10	
	colsample_bytree	0.6, 0.8, 1.0	1.0	
Histogram-based Gradient Boosting	learning_rate	0.01, 0.05, 0.1	0.1	0.9336
	max_depth	2, 3, 5	2	
	l2_regularization	0.0, 0.1, 1.0	1.0	
	early_stopping	True	True	
	validation_fraction	0.1, 0.2	0.1	
Light Extreme Gradient Machines	n_estimators	200, 400, 800	200	0.9323
	learning_rate	0.01, 0.05, 0.1	0.05	
	max_depth	-1, 3, 5, 7	-1	
	min_child_samples	10, 20, 50	10	
	subsample	0.7, 0.85, 1.0	0.7	
Adaptive Boosting	n_estimators	50, 100, 200, 400	400	0.9262
	learning_rate	0.01, 0.1, 1.0	1.0	
	algorithm	SAMME, SAMME.R	SAMME	

Table 4 Performance evaluation metrics of ML and DL models for bankruptcy prediction with SVM-SMOTE resampling technique

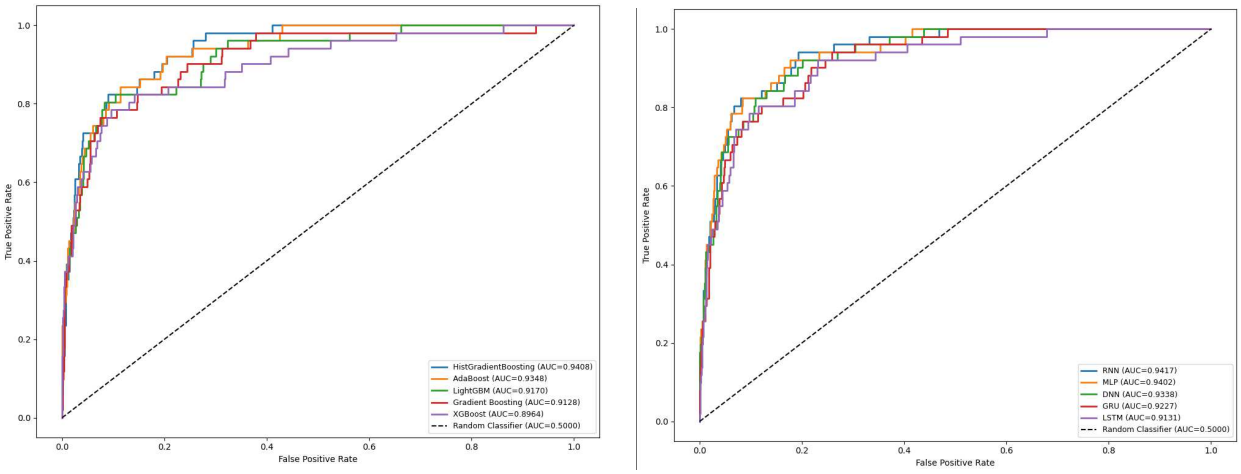
Models	Accuracy	Precision	Recall	F1-Score	Specificity	G-mean	MCC	ROC-AUC
GB	0.9428	0.3448	0.5882	0.4348	0.9566	0.7501	0.4230	0.9128
XGB	0.9538	0.4167	0.5882	0.4878	0.9680	0.7546	0.4719	0.8964
HGB	0.9172	0.2754	0.7451	0.4021	0.9238	0.8297	0.4208	0.9408
LGBM	0.9545	0.4098	0.4902	0.4464	0.9726	0.6905	0.4248	0.9170
AB	0.9326	0.3217	0.7255	0.4458	0.9406	0.8261	0.4548	0.9348
RNN	0.9032	0.2545	0.8235	0.3889	0.9063	0.8639	0.4246	0.9417
LSTM	0.8966	0.2353	0.7843	0.3620	0.9010	0.8406	0.3936	0.9131
GRU	0.9091	0.2583	0.7647	0.3861	0.9147	0.8363	0.4108	0.9227
DNN	0.9157	0.2714	0.7451	0.3979	0.9223	0.8290	0.4172	0.9338
MLP	0.8864	0.2234	0.8235	0.3515	0.8888	0.8555	0.3920	0.9402

Table 5 Performance evaluation metrics of ML and DL models for bankruptcy prediction with SMOTE-Tomek resampling technique

Models	Accuracy	Precision	Recall	F1-Score	Specificity	G-mean	MCC	ROC-AUC
GB	0.9289	0.2909	0.6275	0.3975	0.9406	0.7682	0.3958	0.9044
XGB	0.9472	0.3600	0.5294	0.4286	0.9634	0.7142	0.4102	0.9055
HGB	0.8871	0.2186	0.7843	0.3419	0.8911	0.8360	0.3759	0.9274
LGBM	0.9494	0.3676	0.4902	0.4202	0.9673	0.6886	0.3988	0.9177
AB	0.8952	0.2262	0.7451	0.3470	0.9010	0.8193	0.3730	0.9122
RNN	0.8893	0.2283	0.8235	0.3574	0.8919	0.8570	0.3973	0.9382
LSTM	0.8666	0.1925	0.8039	0.3106	0.8690	0.8358	0.3517	0.9140
GRU	0.8534	0.1775	0.8039	0.2908	0.8553	0.8292	0.3334	0.9280
DNN	0.8651	0.1878	0.7843	0.3030	0.8682	0.8252	0.3410	0.9189
MLP	0.8416	0.1739	0.8627	0.2895	0.8408	0.8517	0.3434	0.9361

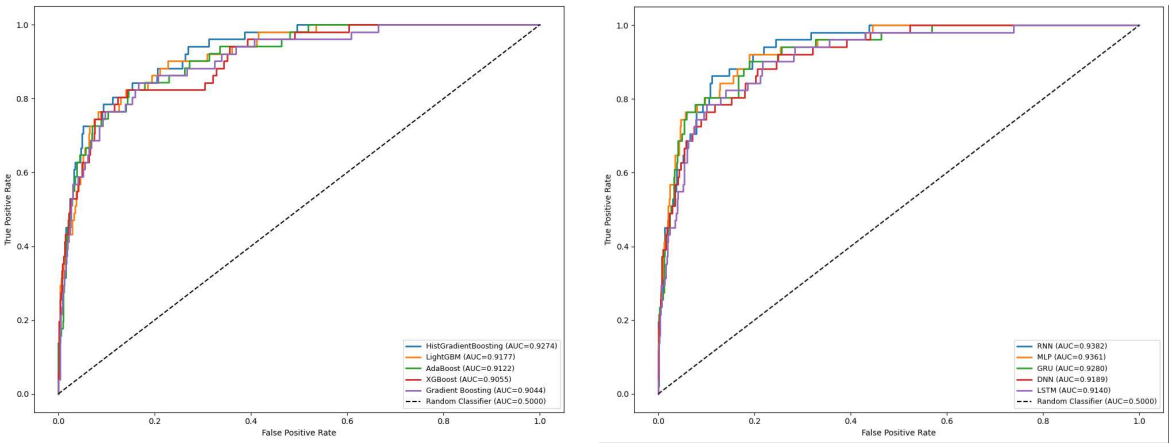
Table 6 Performance evaluation metrics of ML and DL models for bankruptcy prediction with SMOTE-ENN resampling technique

Models	Accuracy	Precision	Recall	F1-Score	Specificity	G-mean	MCC	ROC-AUC
GB	0.9091	0.2550	0.7451	0.3800	0.9155	0.8259	0.4017	0.9034
XGB	0.9164	0.2701	0.7255	0.3936	0.9238	0.8187	0.4098	0.9039
HGB	0.8732	0.2010	0.8039	0.3216	0.8759	0.8391	0.3616	0.9260
LGBM	0.9289	0.3017	0.6863	0.4192	0.9383	0.8025	0.4248	0.9145
AB	0.8761	0.1990	0.7647	0.3158	0.8804	0.8205	0.3489	0.9095
RNN	0.8446	0.1687	0.8039	0.2789	0.8462	0.8248	0.3223	0.9344
LSTM	0.7691	0.1313	0.9216	0.2298	0.7631	0.8386	0.2952	0.9264
GRU	0.8365	0.1692	0.8627	0.2830	0.8355	0.8490	0.3372	0.9371
DNN	0.8328	0.1710	0.9020	0.2875	0.8302	0.8653	0.3491	0.9397
MLP	0.8328	0.1685	0.8824	0.2830	0.8309	0.8563	0.3410	0.9386



(a) (b)

Fig. 4 ROC and AUC of ML and DL models for bankruptcy prediction with SVM-SMOTE resampling technique.



(a) (b)

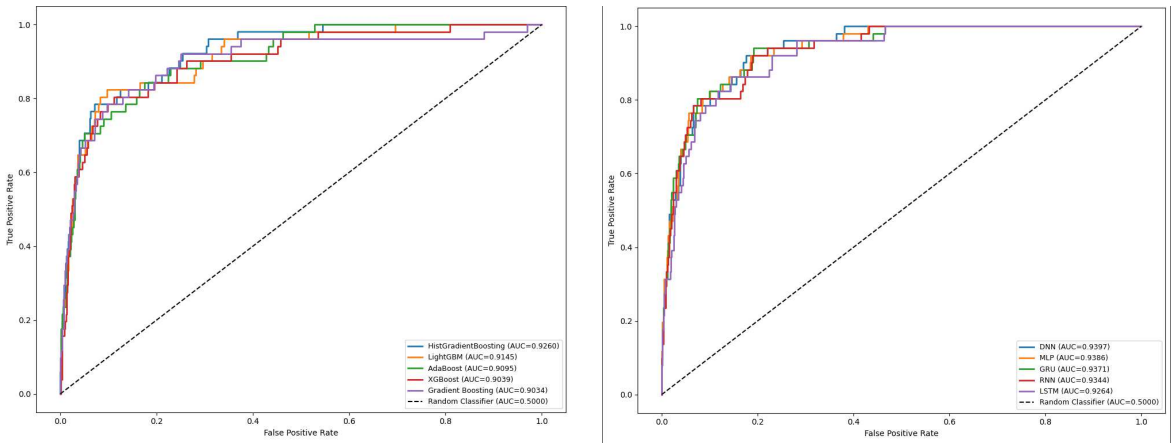
Fig. 5 ROC and AUC of ML and DL models for bankruptcy prediction with SMOTE-Tomek resampling technique.

Figures 7, 8 and 9 present the resulting ROC and PR of ML and DL models after resampling with SVM-SMOTE, SMOTE-Tomek and SMOTE-ENN techniques.

3.3 Results and findings of ML and DL stacking models under the three resampling techniques

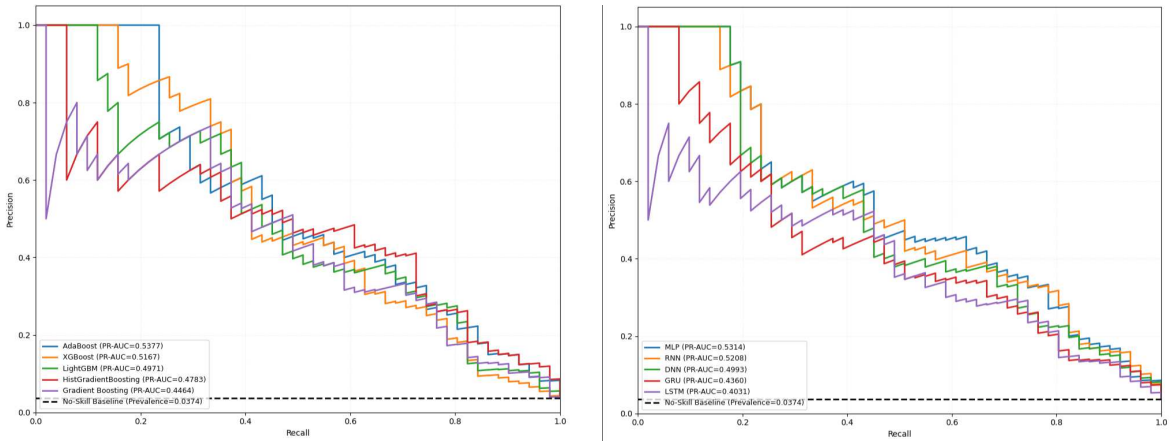
A descriptive statistical analysis of **Table 7** shows that the hybrid stacking ensembles exhibit a clear trade-off between overall classification accuracy and minority-class detection, depending on the resampling strategy used. Across the three resampling methods, the SMOTE-SVM variants produced the strongest average accuracy (mean = 0.9576) and specificity (mean = 0.9785), followed closely by SMOTE-Tomek (accuracy mean = 0.9559; specificity mean = 0.9758). In contrast, SMOTE-ENN yielded the lowest average accuracy (mean = 0.9361) and specificity (mean = 0.9461), but it achieved the best recall (mean = 0.6784) and the highest average G-mean (mean = 0.8006). This pattern indicates that SMOTE-ENN substantially improved the detection of bankrupt firms, although this gain came at the cost of more false positives among non-bankrupt firms.

The results in Table 7 reveal that model performance varied systematically across resampling



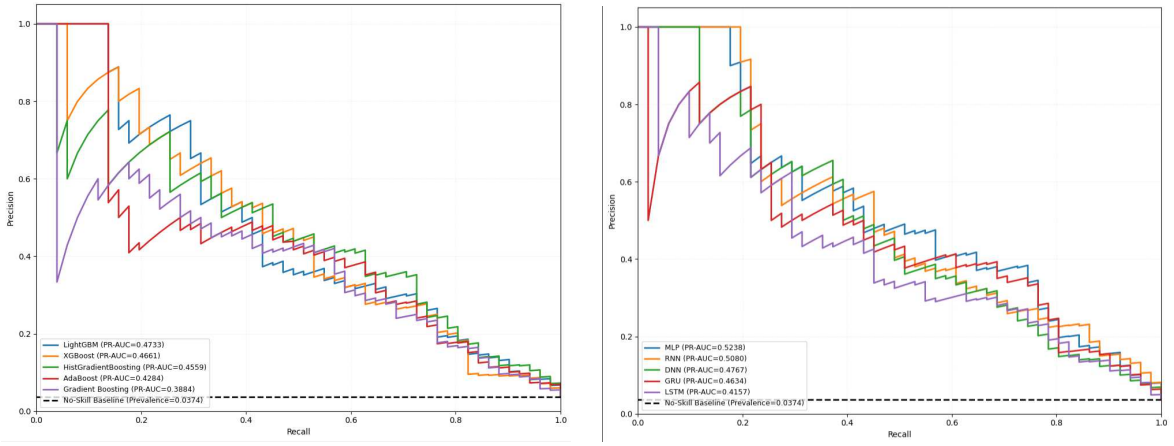
(a) (b)

Fig. 6 ROC and AUC of ML and DL models for bankruptcy prediction with SMOTE-ENN resampling technique.



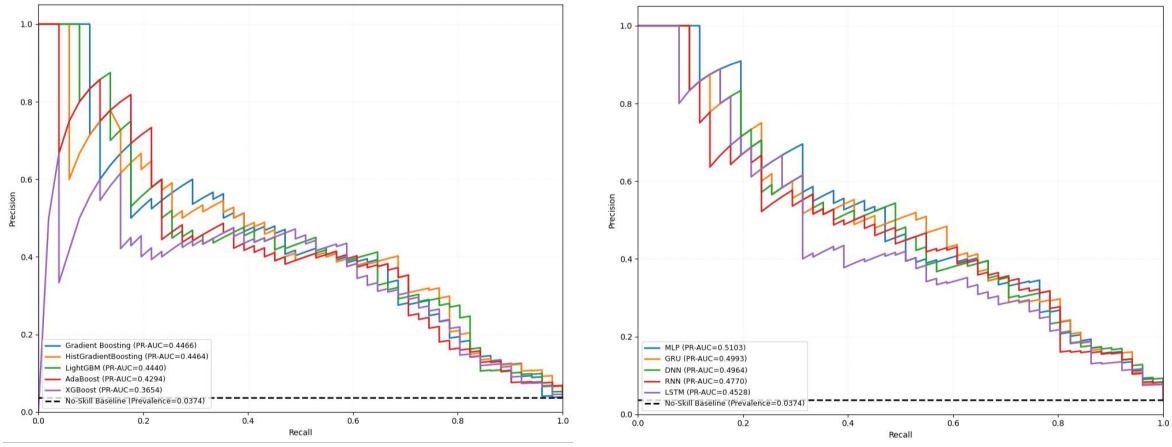
(a) (b)

Fig. 7 ROC and PR of ML and DL models for bankruptcy prediction with SVM-SMOTE resampling technique.



(a) (b)

Fig. 8 ROC and PR of ML and DL models for bankruptcy prediction with SMOTE-Tomek resampling technique.



(a) (b)

Fig. 9 ROC and PR of ML and DL models for bankruptcy prediction with SMOTE-ENN resampling technique.

strategies. SMOTE-SVM produced the highest average accuracy and specificity, indicating superior majority-class classification, whereas SMOTE-ENN achieved the highest recall and G-mean, reflecting better minority-class detection and more balanced predictive performance. Among all hybrid stacking ensembles, SMOTE-ENN+(GB+XGB+HGB+LGBM+AB)+LSTM emerged as the most practically effective model, achieving recall of 0.7255, G-mean of 0.8254, and an AUC value of 0.9270. These findings suggest that, although SMOTE-SVM favors overall correctness, SMOTE-ENN is more appropriate for bankruptcy prediction because it improves the identification of financially distressed firms, which is the more critical classification objective in imbalanced datasets.

3.4 Explainable Artificial Intelligence Driven Analysis using SHAP

The SHAP framework was applied exclusively after model evaluation and was restricted to the final selected ML/DL models: HGB and RNN using the three resampling strategies adopted in the study. Larger SHAP values indicating relatively higher importance in their feature contribution. All reported SHAP values were calculated using the held-out test observations; therefore, the resulting explanations reflect model behavior on previously unseen firms rather than on

Table 7 Performance metrics of hybrid stacking ensemble models (5+1) bankruptcy prediction with SMOTE-SVM, SMOTE-Tomek, and SMOTE-ENN resampling techniques with ML models as base learners and DL models as meta learners

Models	Accuracy	Recall	Specificity	G-mean	ROC-AUC
Testing Performance-SMOTE-SVM (5+1)					
(GB+XGB+HGB+LGBM+AB)+RNN	0.9619	0.3529	0.9855	0.5898	0.8841
(GB+XGB+HGB+LGBM+AB)+LSTM	0.9523	0.4706	0.9711	0.6760	0.9324
(GB+XGB+HGB+LGBM+AB)+GRU	0.9538	0.4902	0.9718	0.6902	0.9325
(GB+XGB+HGB+LGBM+AB)+DNN	0.9626	0.3529	0.9863	0.5900	0.9264
(GB+XGB+HGB+LGBM+AB)+MLP	0.9575	0.4314	0.9779	0.6495	0.7371
Testing Performance-SMOTE-Tomek (5+1)					
(GB+XGB+HGB+LGBM+AB)+RNN	0.9604	0.3333	0.9848	0.5729	0.6599
(GB+XGB+HGB+LGBM+AB)+LSTM	0.9457	0.5490	0.9612	0.7264	0.9256
(GB+XGB+HGB+LGBM+AB)+GRU	0.9604	0.4510	0.9802	0.6649	0.8173
(GB+XGB+HGB+LGBM+AB)+DNN	0.9575	0.3922	0.9794	0.6198	0.7671
(GB+XGB+HGB+LGBM+AB)+MLP	0.9553	0.4902	0.9733	0.6907	0.6650
Testing Performance-SMOTE-ENN (5+1)					
(GB+XGB+HGB+LGBM+AB)+RNN	0.9450	0.6078	0.9581	0.7631	0.7903
(GB+XGB+HGB+LGBM+AB)+LSTM	0.9311	0.7255	0.9391	0.8254	0.9270
(GB+XGB+HGB+LGBM+AB)+GRU	0.9304	0.7255	0.9383	0.8251	0.9268
(GB+XGB+HGB+LGBM+AB)+DNN	0.9413	0.6667	0.9520	0.7967	0.8145
(GB+XGB+HGB+LGBM+AB)+MLP	0.9326	0.6667	0.9429	0.7928	0.8079

synthetic or training observations. Mean absolute SHAP values were used to rank global feature importance, while the signed SHAP distributions showed the direction and variability of each feature's contribution. Because the explainers may operate on model-specific output scales, feature rankings were interpreted within each model rather than by directly comparing SHAP magnitudes across models.

SHAP analysis was conducted Among the boosting classifiers, HGB provided the strongest minority-sensitive performance. It consistently achieved the highest recall, G-mean, and test ROC-AUC across the three resampling conditions, indicating greater effectiveness in identifying bankrupt firms while maintaining balanced class-wise performance. HGB with SVM-SMOTE was selected as the final boosting configuration because it achieved the highest test ROC-AUC of 0.9408, together with a recall of 0.7451 and a G-mean of 0.8297. Although LGBM achieved higher accuracy and specificity and XGB produced the highest cross-validation ROC-AUC during hyperparameter tuning, these measures placed comparatively greater emphasis on overall or majority-class performance. HGB was therefore considered the most appropriate boosting model for the minority-sensitive objective of bankruptcy prediction.

Among the deep learning models, RNN provided the strongest overall performance across the evaluated resampling conditions. It achieved the highest average accuracy, precision, F1-score, specificity, and MCC while maintaining competitive recall, G-mean, and ROC-AUC. RNN with SVM-SMOTE represented the most balanced individual configuration, achieving a recall of 0.8235, G-mean of 0.8639, and ROC-AUC of 0.9417. Although MLP recorded marginally higher average recall, G-mean, and ROC-AUC, these gains were accompanied by lower precision, F1-score, specificity, and MCC. RNN was therefore selected as the most appropriate overall DL model because it provided a stronger balance between minority-class detection and false-positive control.

Thus, we assess the SHAP feature contribution analysis of the best ML model (HGB) and the best ML model (RNN) under the three resampling strategies.

3.5 Explainable Artificial Intelligence Driven Analysis using SHAP under SVM-SMOTE

For the RNN classifier under SVM-SMOTE resampling technique, it took a total of 1,365 iterations, a time of 2 hours, 46 minutes, and 43 seconds at 7.33 seconds per iteration to generate the feature importance and the SHAP values. However, it took less than 35 seconds for the HGB model to generate the feature importance and the SHAP values.

Figure 10 presents the global and local SHAP results under SVM-SMOTE. Based on mean absolute SHAP values, the HGB model was primarily influenced by Retained Earnings to Total Assets, Debt ratio, ROA(C), the interest expense-to-EBIT measure, and Total Asset Turnover. The RNN relied most strongly on Debt ratio, followed by ROA(C), Total Asset Turnover, Net Income to Total Assets, and Net worth/Assets. The signed SHAP summary distributions, which aggregate local explanations for individual test observations, show that higher debt ratios and greater interest expenses relative to EBIT generally shifted predictions toward bankruptcy. Conversely, higher retained earnings, ROA(C), asset turnover, net income, and net worth generally reduced predicted bankruptcy risk. Higher Total expense/Assets also tended to increase the RNN's predicted risk.

3.6 Explainable Artificial Intelligence Driven Analysis using SHAP under SMOTE-Tomek

For the RNN classifier under SMOTE-Tomek resampling technique, it took a total of 1,365 iterations, a time of 2 hours, 43 minutes, and 16 seconds at 7.19 seconds per iteration to generate the feature importance and the SHAP values. However, it took less than 50 seconds for the HGB model to generate the feature importance and the SHAP values.

Figure 11 shows the global and local SHAP results under SMOTE-Tomek. Globally, Net Income to Total Assets, the interest expense-to-EBIT measure, Retained Earnings to Total Assets, Net worth/Assets, and Persistent EPS were the most important HGB predictors. For the

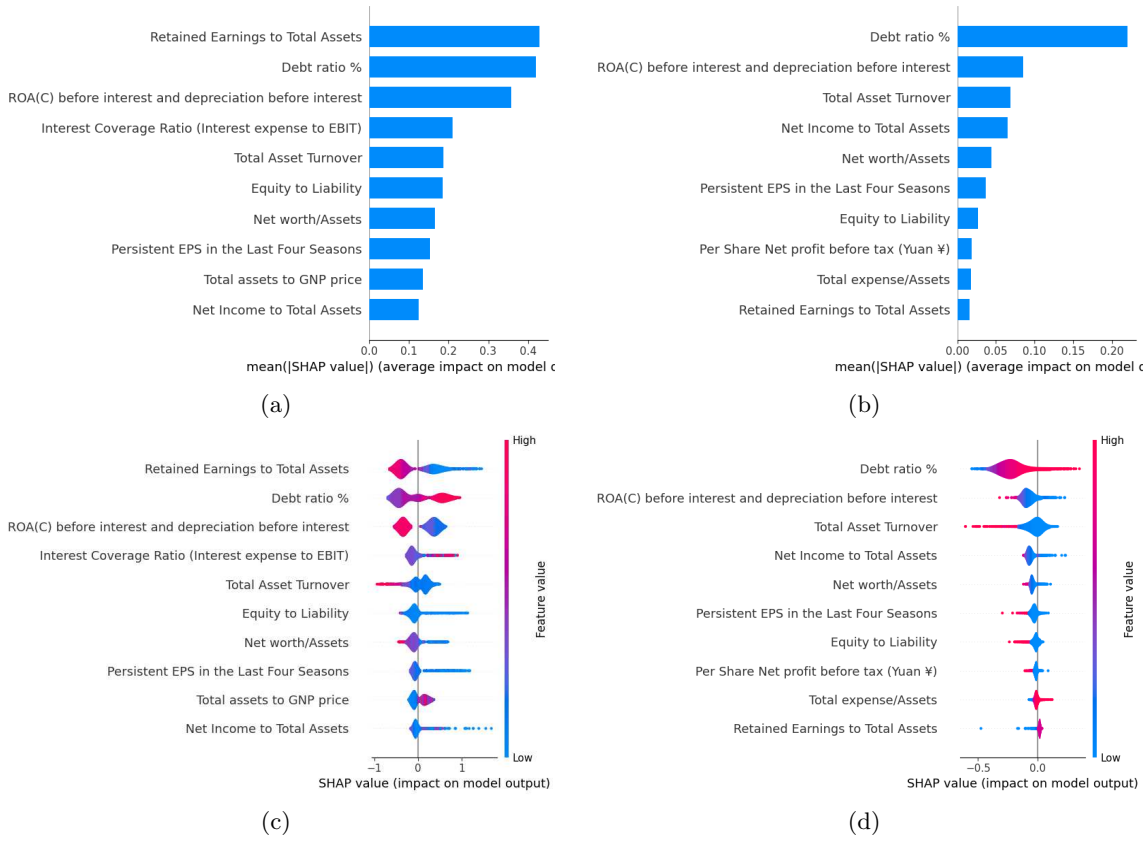


Fig. 10 SHAP explanations computed on the held-out test set for the final selected HGB and RNN models with SVM-SMOTE. Panels (a) and (b) present global feature importance based on mean absolute SHAP values for HGB and RNN, respectively. Panels (c) and (d) show the direction and distribution of feature contributions to the predictions of the respective models. Larger absolute SHAP values indicate stronger contributions to the model output.

RNN, Net worth/Assets and ROA(C) were dominant, followed by Persistent EPS, Debt ratio, and Total Asset Turnover. The local SHAP distributions indicate that higher net income, retained earnings, net worth, ROA(C), persistent EPS, cash flow, and asset turnover generally produced negative SHAP values and reduced predicted bankruptcy risk. In contrast, higher debt ratios, greater interest expenses relative to EBIT, and longer average collection periods produced more positive SHAP values, shifting predictions toward bankruptcy.

3.7 Explainable Artificial Intelligence Driven Analysis using SHAP under SMOTE-ENN

For the RNN classifier under SMOTE-ENN resampling technique, it took a total of 1,365 iterations, a time of 2 hours, 32 minutes, and 35 seconds at 6.71 seconds per iteration to generate the feature importance and the SHAP values. However, it took less than 60 seconds for the HGB model to generate the feature importance and the SHAP values.

Figure 12 presents the corresponding SHAP interpretations under SMOTE-ENN. The global HGB results identify Net Income to Total Assets, Net worth/Assets, the interest expense-to-EBIT measure, Retained Earnings to Total Assets, and Per Share Net Profit Before Tax as the leading

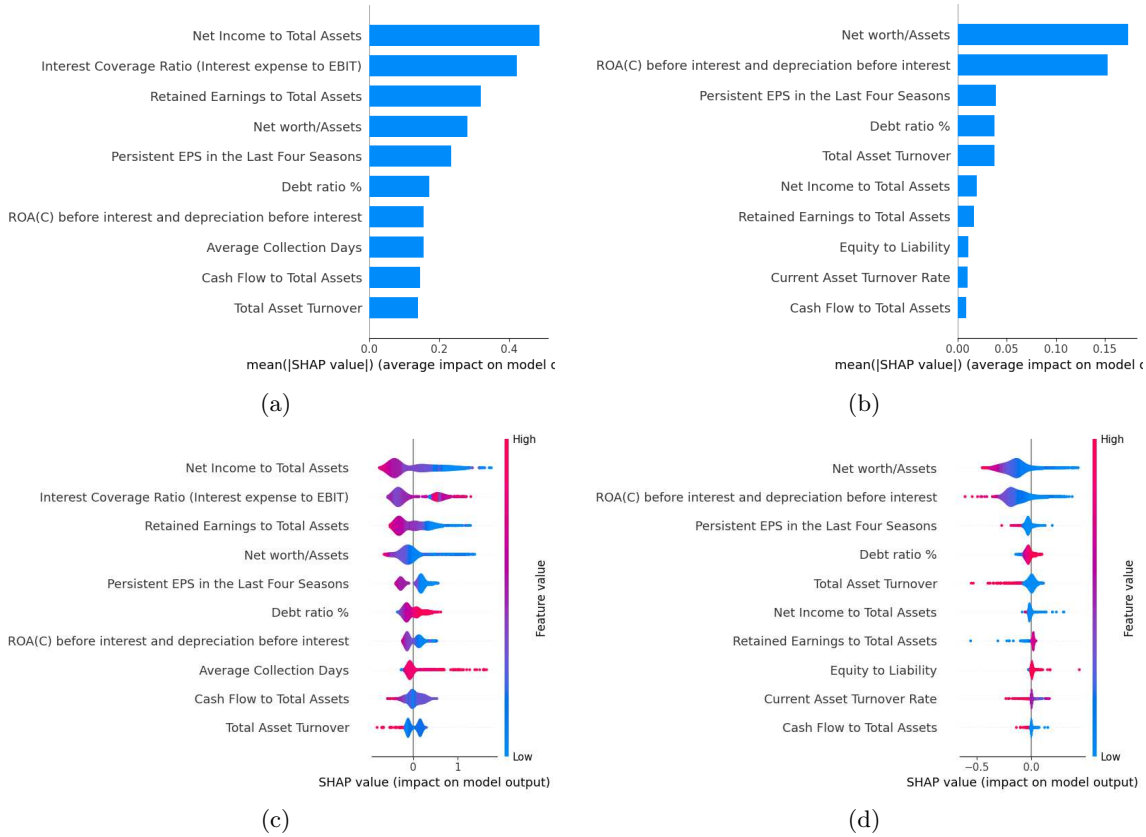


Fig. 11 SHAP explanations computed on the held-out test set for the final selected HGB and RNN models with SMOTE-Tomek. Panels (a) and (b) present global feature importance based on mean absolute SHAP values for HGB and RNN, respectively. Panels (c) and (d) show the direction and distribution of feature contributions to the predictions of the respective models. Larger absolute SHAP values indicate stronger contributions to the model output.

predictors. In the RNN, Debt ratio was the dominant feature, followed by ROA(C), Total Asset Turnover, Persistent EPS, and Operating Expense Rate. At the local level, higher net income, net worth, retained earnings, per-share profit, cash flow, ROA(C), and asset turnover generally reduced the model's predicted bankruptcy risk. Conversely, higher debt ratios, interest expenses relative to EBIT, average collection periods, operating expense rates, and total expenses relative to assets generally increased predicted bankruptcy risk. These patterns consistently associate financial leverage and expense burden with greater risk, while profitability, solvency, and operational efficiency are associated with lower risk.

3.8 Comparative Analysis with other studies within the last ten (10) years

Table 8 indicates a clear methodological progression in financial credit default and bankruptcy prediction, moving from conventional ML models such as logistic regression, decision trees, random forests, and SVMs toward more complex ensemble and deep learning architectures, including RNNs, LSTMs, and hybrid GRU-based systems. In broad terms, the more recent studies tend to report stronger discriminatory performance, particularly when evaluated with AUC, F1, ROC-AUC, recall, or G-mean rather than accuracy alone. Within this pattern, the current study appears highly competitive, reporting the AUC (HGB: 0.9408, RNN: 0.9417, stacking ensemble: 0.9270) together with recall (HGB: 74.51%, RNN: 82.35%, stacking ensemble: 72.55%) and G-mean (HGB: 82.97%,

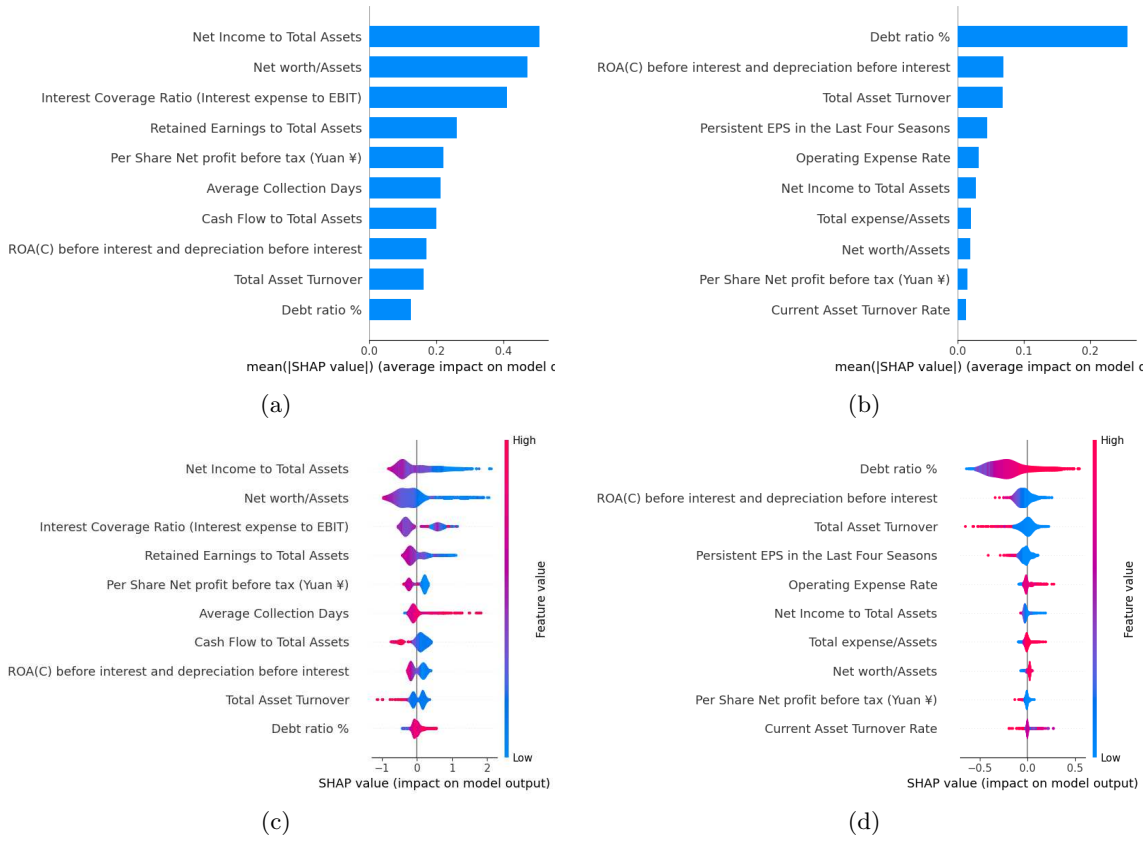


Fig. 12 SHAP explanations computed on the held-out test set for the final selected HGB and RNN models with SMOTE-ENN. Panels (a) and (b) present global feature importance based on mean absolute SHAP values for HGB and RNN, respectively. Panels (c) and (d) show the direction and distribution of feature contributions to the predictions of the respective models. Larger absolute SHAP values indicate stronger contributions to the model output.

RNN: 86.39%, stacking ensemble: 82.54%), which suggests good balance between sensitivity and class-wise performance.

Table 8 also highlights an important comparability problem. The studies differ substantially in dataset type, sample size, feature dimensionality, and prediction task, ranging from credit default to bankruptcy, with instances varying from a few hundred to tens of millions.

Recent evidence confirms that the effectiveness of bankruptcy prediction methods is strongly dataset-dependent. [35] evaluated 45 imbalanced-learning methods across 15 bankruptcy datasets and found that hybrid ensemble approaches, including SMOTE-ENN with XGBoost, Balanced Random Forest, and EasyEnsemble, were among the most consistently effective methods. [36] subsequently found that boosting models performed strongly on the Taiwanese dataset, although the additional benefit of SMOTE-ENN diminished when probability calibration and cost-sensitive threshold optimization were introduced. [37] evaluated a knowledge-distillation framework on Taiwanese and Polish data, reporting AUC values of 0.9317 and 0.8874, respectively, while substantially reducing computational cost. [38] similarly demonstrated the importance of validating class-rebalanced models across multiple UCI datasets and a separate sample of Chinese listed companies. These studies support the use of hybrid imbalance-learning frameworks but also show that performance can vary across datasets, validation protocols, and decision criteria. Consequently, numerical differences between studies should be interpreted cautiously rather than treated as direct evidence of superiority.

Table 8 Comparison of previous studies for financial credit default and bankruptcy prediction

Author	Year	Dataset	Features Count	Total Instances	Method	Best Results
[55]	2016	Six major USA financial institutions	186	54553000	Random forest	Average F1=81.50%
[56]	2018	German Credit Dataset	24	1000	Ensemble fusion techniques based on bstacking method	Accuracy=78.66%
[56]	2018	Australian credit dataset	14	600	Ensemble fusion techniques based on bstacking method	Accuracy=88.28%
[56]	2018	Dataset of We.com	17	1421	Ensemble fusion techniques based on bstacking method	Accuracy=84.07%
[56]	2018	Lending Club dataset	11	2642	Ensemble fusion techniques based on bstacking method	Accuracy=66.75%
[57]	2018	Taiwan credit client's dataset	24	30000	Boosting	Accuracy=71.06%
[58]	2018	Taiwan credit client's dataset	24	30000	SPR-RIPPER	F1=53.4%, F2=88.5, AUC=70.7%
[59]	2019	Chinese consumer finance company	490	44000	Default prediction with transfer learning	AUC=0.7170, Specificity=0.6142, Sensitivity=0.7039
[60]	2019	Taiwan credit client's dataset	24	30000	RNN-RF	AUC =0.782, Lift Index=0.659
[61]	2019	Chinese P2P lending company dataset	1138	15000	Decision tree-based heterogeneous ensemble model	AUC=71.85%
[62]	2019	Taiwan credit client's dataset	24	30000	Ensemble model	Accuracy=83.83%, Recall=94.50%, F1=90.31%
[63]	2020	Lending Club dataset	15	64139	Logistic regression	Accuracy=79.20%
[64]	2020	Taiwan credit client's dataset	24	30000	Gradient Boosted Decision Tree Model	Accuracy=88.70%
[64]	2020	South German credit client's dataset	21	1000	Gradient Boosted Decision Tree Model	Accuracy=83.50%
[64]	2020	Belgium credit client's dataset	28	285299	Gradient Boosted Decision Tree Model	Accuracy=86.30%
[28]	2022	Compustat North America dataset and Center for Research in Security Prices (CRSP) dataset	8	454752	RNN, LSTM and Ensemble	AUC=0.7305
[40]	2023	Orbis database bankruptcy dataset	17	266	FCNN	Accuracy=78.56%, AUC=0.8240
[41]	2024	Brazil financial ratios dataset	26	503	SVM	AUC=0.7776
[65]	2025	IND and SUM dataset	14	18722059	LSTM	Accuracy=92%
[66]	2025	Taiwan credit client's dataset	24	30000	Boruta+DBSCAN+SMOTE-Tomek+GBM	F1-score: 82.56%, G-mean: 82.98%, AUC: 0.9090, PR-AUC: 0.9185
This Study	2026	Taiwan bankruptcy dataset	95	6819	Best ML model (HGB)	G-mean: 82.97%, Recall: 74.51%, AUC: 0.9408
This Study	2026	Taiwan bankruptcy dataset	95	6819	Best DL model (RNN)	G-mean: 86.39%, Recall: 82.35%, AUC: 0.9417
This Study	2026	Taiwan bankruptcy dataset	95	6819	Best stacking ensemble model (SMOTE-ENN + (GB + XGB + HGB + LGBM + AB) + LSTM)	G-mean: 82.54%, Recall: 72.55%, AUC: 0.9270

4 Discussions

The present study adds to the recent bankruptcy prediction literature by showing that performance in a severely imbalanced corporate-failure setting depends more on how well a model identifies the minority bankrupt class than on overall accuracy alone. First, in this study data, models such as LGBM achieved very high accuracy and specificity under SVM-SMOTE and SMOTE-Tomek, but these gains were accompanied by materially lower recall. By contrast, several deep models, especially MLP, GRU, LSTM, and RNN under SMOTE-ENN, identified a much larger share of bankrupt firms, even at the cost of more false positives. This divergence is not trivial. Recent research has shown that the ranking of bankruptcy prediction models changes when the evaluation criterion shifts from purely statistical accuracy to economically meaningful error trade-offs, because missed bankruptcies and false alarms carry asymmetric costs for creditors, investors, auditors, and regulators [67]. The current findings fit that logic closely. If the practical objective is early warning and intervention, the superior recall and G-mean of the best deep learning models may be more valuable than the higher specificity of the best boosting models.

Furthermore, hybrid resampling was not merely a preprocessing convenience but a central determinant of model behavior. SVM-SMOTE and SMOTE-Tomek tended to preserve higher specificity and accuracy, whereas SMOTE-ENN consistently produced the strongest recall and the best average G-mean across the stacked models. This pattern is consistent with the broader bankruptcy studies on imbalanced learning, which shows that aggressive editing and cleaning procedures can improve the separability of rare failure cases, even though they often do so by sacrificing some majority-class precision [7]. In this study, the practical consequence was clear: SMOTE-ENN created the conditions under which the classifiers became more responsive to financially distressed firms. This result is also in line with more recent work emphasizing that bankruptcy prediction in imbalanced settings should not rely on a single balancing recipe, because the preferred method depends on the degree of imbalance, the sample size, and the forecast horizon [68, 69]. The evidence here suggests that, for the Taiwanese dataset, the edited neighborhood structure induced by SMOTE-ENN was more compatible with the study's operational objective of minimizing missed bankruptcies.

At the level of individual model families, the results show a meaningful division of labor between ML and DL models. Among the ML models, HGB was the most consistently competitive classifier, yielding the best ROC-AUC under all three resampling strategies and comparatively strong recall relative to GB, XGB, LGBM, and AdaBoost. This suggests that histogram-based gradient boosting was particularly effective at extracting nonlinear structure from the selected ratio set while remaining relatively robust to the altered class geometry created by resampling. However, LGBM repeatedly produced the highest accuracy and specificity, indicating that it was especially conservative in assigning firms to the bankrupt class. That contrast mirrors the broader literature, where tree-ensemble methods often provide strong overall discrimination but not always the highest bankrupt-class sensitivity, especially when the training signal is sparse or heavily skewed [42, 70]. The fact that HGB rather than XGBoost emerged as the strongest ML model in this study also usefully qualifies recent evidence from other samples in which XGBoost has been dominant, suggesting that there is no universally best boosting architecture across datasets, horizons, and preprocessing pipelines [4, 42].

The DL results are equally notable. Under all three resampling strategies, the neural models generally outperformed the ML models on recall and often on G-mean, with the strongest gains emerging under SMOTE-ENN. RNN achieved the highest average accuracy, precision, F1-score, specificity, and MCC while maintaining competitive recall, G-mean, and ROC-AUC. RNN with SVM-SMOTE represented the most balanced individual configuration, achieving a recall of 0.8235, G-mean of 0.8639, and ROC-AUC of 0.9417. These results are broadly consistent with studies showing that neural architecture can be highly effective when failure risk depends on nonlinear interactions that are difficult to represent through linear or shallow tree boundaries alone [40, 43]. These findings support the usefulness of DL for bankruptcy prediction, but they also suggest that the full potential of RNN, LSTM and

GRU models would probably be better assessed on multi-period firm histories rather than on one-record-per-firm data.

Because RNN, LSTM, and GRU are order-sensitive, their application to unordered cross-sectional predictors and base-learner probabilities is not architecturally natural. Their results may be influenced by the artificially imposed feature order and should therefore be interpreted cautiously. These findings are not considered evidence that recurrent modeling is intrinsically appropriate for this dataset. The boosting, DNN, and MLP models provide more structurally aligned benchmarks for the tabular data.

A particularly important result is that stacking did not automatically dominate the best standalone models. Within the stacking framework, the (GB+XGB+HGB+LGBM+AB) +LSTM coupled with SMOTE-ENN model provided the strongest overall balance, with recall of 0.7255, G-mean of 0.8254, and ROC-AUC of 0.9270. This is a credible result and confirms that heterogeneous meta-learning can still yield a practically useful compromise between sensitivity and specificity. However, several standalone deep models under SMOTE-ENN achieved substantially higher recall and higher or comparable ROC-AUC. This indicates that, in the present studwy, additional ensemble depth did not translate into a better early-warning frontier. One plausible explanation is limited diversity among the base learners: all five base models were boosting ensembles operating on the same tabular feature space, so the meta-learners may have received highly correlated signals. Another explanation is that the stacked architecture may have been constrained by the absence of tuned deep meta-learner hyperparameters. These are aligned with earlier evidence that stacking is most beneficial when learner diversity and feature complementarity are sufficiently strong, and that feature selection or alternative information sets can be decisive in unlocking ensemble gains [71, 72].

Any improvement obtained through stacking should be attributed to the nonlinear combination of the base learners' probability estimates rather than to the modeling of temporal dependencies. In particular, the recurrent meta-learners are treated as exploratory nonlinear aggregators because their inputs are unordered model probabilities.

The convergence between the consensus feature-selection procedure and the SHAP results in Figures 10–12 strengthens the substantive validity of the predictive framework. Although the precise rankings varied across models and resampling strategies, the most recurrent predictors were Retained Earnings to Total Assets, Net Income to Total Assets, ROA(C), Net worth/Assets, Debt ratio, Total Asset Turnover, Persistent EPS, the interest expense-to-EBIT measure, Operating Expense Rate, Total expense/Assets, and Average Collection Days. The HGB models generally emphasized cumulative profitability and solvency, whereas the RNN models assigned greater importance to leverage and operating efficiency. The signed SHAP distributions further showed that higher debt ratios, interest expenses relative to EBIT, operating and total expense burdens, and longer collection periods shifted predictions toward bankruptcy. Conversely, stronger retained earnings, profitability, net worth, cash flow, EPS, and asset turnover reduced predicted bankruptcy risk. These findings indicate that bankruptcy risk reflects the combined deterioration of earnings capacity, solvency, debt-servicing capacity, working-capital efficiency, and cost control. This interpretation is consistent with recent SHAP-based bankruptcy research, although cross-market and network-based evidence indicates that the relative importance of individual predictors may vary with institutional context and the available feature set [42, 43, 73, 74]. The agreement between the present SHAP results and these studies increases confidence that the models are capturing economically plausible signals rather than spurious statistical artifacts.

The SHAP analysis has several limitations. First, SHAP is a post hoc explanation method and does not establish that an influential predictor has a causal effect on bankruptcy. Second, when financial indicators are strongly correlated, their contributions may be divided or redistributed among related variables, making individual feature rankings sensitive to the included predictor set. This issue was partially mitigated by the consensus feature-selection procedure, which reduced the original 95 predictors to 23 consistently selected variables, and by interpreting related financial indicators collectively rather than as isolated determinants. Similarities between the model explanations were also examined to identify

broader patterns that remained consistent across model classes. Nevertheless, residual correlation may remain, and small differences in individual SHAP rankings should therefore be interpreted cautiously.

5 Conclusion and Recommendations

This study supports three broader conclusions. First, class imbalance is not a peripheral nuisance in bankruptcy prediction; it shapes model rankings, error distributions, and practical usefulness. Second, deep learning can offer substantial gains in bankrupt-class detection even in structured tabular settings, although its advantage would likely be even clearer with true sequential data. Third, explainability can meaningfully strengthen bankruptcy prediction research when it is used to validate whether influential predictors are economically interpretable and consistent across model classes. By combining consensus feature selection, hybrid resampling, heterogeneous learning architectures, and SHAP-based interpretation, the present study advances the bankruptcy studies toward a more balanced view of performance, one that values not only discrimination but also interpretability and operational relevance. That combination makes the study well aligned with the current direction of bankruptcy research, while also indicating clear next steps, especially the integration of multi-period and textual disclosures into the present framework. Accordingly, one of the most important implications of this study is not that a single algorithm universally dominates, but that different resampling-model combinations serve different institutional objectives.

A major contribution of this study is that it clarifies what can still be achieved using structured financial-ratio data alone. Recent bankruptcy research increasingly integrates textual and other non-financial inputs, including annual-report language, risk-related disclosures, sustainability reports, and multimodal corporate information, and many of these studies report incremental gains over purely numeric models [9, 10, 75, 76]. The current study does not challenge that trajectory. Rather, it shows that when feature selection, imbalance correction, ensemble learning, and XAI are carefully combined, structured accounting data remain highly informative and can still deliver strong discriminatory performance. In that sense, the present models should be seen as a robust tabular benchmark rather than as a claim that bankruptcy prediction has reached its methodological ceiling. Indeed, recent surveys emphasize that the frontier of the field is moving toward hybrid, multimodal, and more context-aware systems, not away from financial ratios but beyond them [3, 4].

This study is not without limitations. First, the analysis was based on a single Taiwanese bankruptcy dataset, which may limit the generalizability of the results to firms in other countries, industries, or regulatory environments. Second, the models relied mainly on structured financial ratios and did not incorporate potentially useful information such as macroeconomic indicators, market-based variables, or textual disclosures from annual reports.

Furthermore, the study did not include out-of-time validation because the dataset contains no reliable temporal identifiers for constructing chronological training and test cohorts. Consequently, the reported results demonstrate internal predictive performance but do not establish that the models will remain accurate under future economic conditions, regulatory changes, or concept drift. Prospective validation using rolling or forward-chaining evaluation on later firm cohorts is therefore required before practical deployment. Likewise, future studies should validate the proposed framework on larger and more diverse multi-country datasets, use longitudinal and out-of-time testing designs, and integrate additional financial, market, and textual information to improve robustness, practical relevance, and generalizability.

Data Availability. The data that support the findings of this study are openly available in Github Repository at https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction.

References

[1] Correia, M., Accounting and corporate failure: the evolving role of accounting information in bankruptcy prediction. Accounting

and Business Research, 2025. 55(5): p. 510-537.

[2] Gavious, I., O. Milo, and H. Weihs, Financial regulation and bankruptcy prediction: The informational value of auditor opinions. Finance research letters, 2025: p. 108432.

[3] Zhao, J., J. Ouenniche, and J. De Smedt, Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction. Machine learning with applications, 2024. 15: p. 100527.

[4] Dasilas, A. and A. Rigani, Machine learning techniques in bankruptcy prediction: A systematic literature review. Expert systems with applications, 2024. 255: p. 124761.

[5] du Jardin, P., Forecasting corporate failure using ensemble of self-organizing neural networks. European Journal of Operational Research, 2021. 288(3): p. 869-885.

[6] Zoričák, M., et al., Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets. Economic Modelling, 2020. 84: p. 165-176.

[7] Gnip, P., et al., An experimental survey of imbalanced learning algorithms for bankruptcy prediction. Artificial Intelligence Review, 2025. 58(4): p. 104.

[8] Fasano, F., et al., The dilemma of accuracy in bankruptcy prediction: a new approach using explainable AI techniques to predict corporate crises. European Journal of Innovation Management, 2025. 28(11): p. 1-22.

[9] Chen, T.-K., et al., Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports. Expert Systems with Applications, 2023. 233: p. 120714.

[10] Hajek, P. and M. Munk, Corporate financial distress prediction using the risk-related information content of annual reports. Information Processing & Management, 2024. 61(5): p. 103820.

[11] Saito, T. and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 2015. 10(3): p. e0118432.

[12] Davis, J. and M. Goadrich. The relationship between Precision-Recall and ROC curves. in Proceedings of the 23rd international conference on Machine learning. 2006.

[13] Chawla, N.V., et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002. 16: p. 321-357.

[14] Agyemang, E.F., et al., Addressing class imbalance problem in health data classification: Practical application from an oversampling viewpoint. Applied Computational Intelligence and Soft Computing, 2025. 2025(1): p. 1013769.

[15] Alam, N., J. Gao, and S. Jones, Corporate failure prediction: An evaluation of deep learning vs discrete hazard models. Journal of International Financial Markets, Institutions and Money, 2021. 75: p. 101455.

[16] Du Jardin, P. and E. Séverin, Forecasting financial failure using a Kohonen map: A comparative study to improve model stability over time. European Journal of Operational Research, 2012. 221(2): p. 378-396.

[17] Che, W., et al., Predicting financial distress using multimodal data: An attentive and regularized deep learning method. Information Processing & Management, 2024. 61(4): p. 103703.

[18] Jurgovsky, J., et al., Sequence classification for credit-card fraud detection. Expert systems with applications, 2018. 100: p. 234-245.

[19] Motie, S. and B. Raahemi, Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications, 2024. 240: p. 122156.

[20] Friedman, J.H., Greedy function approximation: a gradient boosting machine. Annals of

statistics, 2001: p. 1189-1232.

[21] Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[22] Ke, G., et al., Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017. 30.

[23] Xu, K., Y. Gan, and A. Wilson, Stacked Generalization for Robust Prediction of Trust and Private Equity on Financial Performances. Innovations in Applied Engineering and Technology, 2024: p. 1-12.

[24] Pellegrino, M., et al., A multi-head LSTM architecture for bankruptcy prediction with time series accounting data. Future Internet, 2024. 16(3): p. 79.

[25] Dheepa, V. and R. Dhanapal, Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft computing, 2012. 2(4): p. 391-397.

[26] Xie, Y., et al., A transactional-behavior-based hierarchical gated network for credit card fraud detection. IEEE/CAA Journal of Automatica Sinica, 2025.

[27] Sun, J., M. Zhao, and C. Lei, Class-imbalanced dynamic financial distress prediction based on random forest from the perspective of concept drift. Risk Management, 2024. 26(4): p. 19.

[28] Kim, H., H. Cho, and D. Ryu, Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. Computational Economics, 2022. 59(3): p. 1231-1249.

[29] Wolpert, D.H., Stacked generalization. Neural networks, 1992. 5(2): p. 241-259.

[30] Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 2019. 1(5): p. 206-215.

[31] Černevičienė, J. and A. Kabašinskas, Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 2024. 57(8): p. 216.

[32] Ribeiro, M.T., S. Singh, and C. Guestrin. " 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. 2016.

[33] Mosca, E., et al. SHAP-based explanation methods: a review for NLP interpretability. in Proceedings of the 29th international conference on computational linguistics. 2022.

[34] Nortey, E.N., et al., AI meets economics: Can deep learning surpass machine learning and traditional statistical models in inflation time series forecasting? Data Science in Finance and Economics, 2025.

[35] Gnip, P., et al., An experimental survey of imbalanced learning algorithms for bankruptcy prediction. Artificial Intelligence Review, 2025. 58(4): p. 104.

[36] Dam, T.T. and X.T. Dang, Cost-sensitive ensemble learning for bankruptcy prediction under extreme class imbalance. Discover Artificial Intelligence, 2026. 6(1): p. 473.

[37] Azad, M.A.K., et al., Predicting corporate bankruptcy through a scalable and interpretable knowledge distillation framework. Discover Artificial Intelligence, 2026.

[38] Chen, Z., et al., Forecasting corporate bankruptcy through class-rebalanced self-training semi-constrained matrix factorization. Journal of Forecasting, 2026. 45(2): p. 530-546.

[39] Du Jardin, P., Forecasting bankruptcy using biclustering and neural network-based ensembles. Annals of Operations Research, 2021. 299(1): p. 531-566.

[40] Jabeur, S.B. and V. Serret, Bankruptcy prediction using fuzzy convolutional neural networks. Research in International Business and Finance, 2023. 64: p. 101844.

[41] da Silva Mattos, E. and D. Shasha, Bankruptcy prediction with low-quality financial information. Expert Systems with Applications, 2024. 237: p. 121418.

[42] Nguyen, H.H., J.-L. Viviani, and S. Ben Jabeur, Bankruptcy prediction using machine learning and Shapley additive explanations: H.-H. Nguyen et al. Review of Quantitative Finance and Accounting, 2025. 65(1): p. 107-148.

[43] Lin, Y.-C., et al., Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques. International Journal of Accounting Information Systems, 2025. 56: p. 100744.

[44] Kaewtrakulpong, K., A. Chantiratikul, and O. Srinawakorn, Feature Selection Technique for Model Development, in IoT and AI in Agriculture: Strategies to Develop Smart Agricultural Space for In-field and In-house Crop Production. 2026, Springer. p. 119-132.

[45] Ali, Y.A., et al., Hyperparameter search for machine learning algorithms for optimizing the computational complexity. Processes, 2023. 11(2): p. 349.

[46] Miftahushudur, T., et al. Enhanced svm-smote with cluster consistency for imbalanced data classification. in International Conference on Intelligent Data Engineering and Automated Learning. 2023. Springer.

[47] Muntasir Nishat, M., et al., A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. Scientific Programming, 2022. 2022(1): p. 3649406.

[48] Wenzheng, H., et al., Artificial Intelligence–Enhanced Multi-Algorithm R Shiny Application for Predictive Modeling and Analytics: Case Study of Alzheimer Disease Diagnostics. JMIR aging, 2025. 8(1): p. e70272.

[49] Wenzheng, H., et al., AI-enhanced multi-algorithm R Shiny app for predictive modeling and analytics: a case study of Alzheimer's disease diagnostics. JMIR aging, 2025.

[50] Al Ghamri, M., et al., Whale optimization algorithm for feature selection enhances classification in malware datasets. Journal of Computational and Cognitive Engineering, 2025. 4(3): p. 387-396.

[51] Dwivedi, B., S. Gupta, and V. Kumar, Diversified and enhanced learning-based evolutionary particle swarm optimization and its applications in feature selection. Cluster Computing, 2026. 29(2): p. 109.

[52] El-Shorbagy, M. and I. Nasar, Salp Swarm Optimization: A Comprehensive Review of Recent Advances, Variants, Applications, and Future Research Directions. European Journal of Pure and Applied Mathematics, 2025. 18(4): p. 7030-7030.

[53] Qasim, O.S. and F.A.Y. Al-Taie, CHAOTIC BINARY BAT ALGORITHM WITH MUTUAL INFORMATION FOR FEATURE SUBSET SELECTION. Applicationes Mathematicae, 2026. 53(1).

[54] Li, R., et al., MIARS: Mutual Information-Guided Feature Selection With Angle Reconstruction and Semantic Alignment for Multi-Label Learning. Knowledge-Based Systems, 2026: p. 115424.

[55] Butaru, F., et al., Risk and risk management in the credit card industry. Journal of Banking & Finance, 2016. 72: p. 218-239.

[56] Xia, Y., et al., A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Systems with Applications, 2018. 93: p. 182-199.

[57] Hamori, S., et al., Ensemble learning or deep learning? Application to default risk analysis. Journal of Risk and Financial Management, 2018. 11(1): p. 12.

[58] Xu, P., Z. Ding, and M. Pan, A hybrid interpretable credit card users default prediction model based on RIPPER. Concurrency and Computation: Practice and Experience, 2018. 30(23): p. e4445.

[59] Li, W., et al., Transfer learning-based default prediction model for consumer credit in China: W. Li et al. The Journal of Supercomputing, 2019. 75(2): p. 862-884.

[60] Hsu, T.-C., et al. Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. IEEE.

[61] Zhou, J., et al., Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications, 2019. 534: p. 122370.

[62] Prusti, D. and S.K. Rath. Fraudulent transaction detection in credit card by applying ensemble machine learning techniques. in 2019 10th international conference on computing, communication and networking technologies (ICCCNT). 2019. IEEE.

[63] Xia, Y., et al., Predicting loan default in peer-to-peer lending using narrative data. Journal of Forecasting, 2020. 39(2): p. 260-280.

[64] Alam, T.M., et al., An investigation of credit card default prediction in the imbalanced datasets. Ieee Access, 2020. 8: p. 201173-201198.

[65] Polytarchos, E., Credit Card Fraud Detection Through Deep Learning and Real-Time Data Streams: A Comparison and New Directions. FinTech and Sustainable Innovation, 2025: p. 1-9.

[66] Ampomah, O.-A., et al., Enhancing Credit Default Prediction Using Boruta Feature Selection and DBSCAN Algorithm with Different Resampling Techniques. arXiv preprint arXiv:2509.19408, 2025.

[67] Radovanovic, J. and C. Haas, The evaluation of bankruptcy prediction models based on socio-economic costs. Expert systems with applications, 2023. 227: p. 120275.

[68] Matsumaru, M. and H. Katagiri, A Two-Stage Machine Learning Approach to Bankruptcy Prediction: Integrating Full-Feature Modeling and Optimized Feature Selection. Journal of Risk and Financial Management, 2025. 18(12): p. 662.

[69] Veganzones, D., E. Séverin, and S.B. Jabeur, Forecasting corporate bankruptcy in imbalanced datasets using a new hybrid machine learning approach. Research in International Business and Finance, 2025: p. 103200.

[70] Barboza, F., H. Kimura, and E. Altman, Machine learning models and bankruptcy prediction. Expert systems with applications, 2017. 83: p. 405-417.

[71] Liang, D., et al., Combining corporate governance indicators with stacking ensembles for financial distress prediction. Journal of Business Research, 2020. 120: p. 137-146.

[72] Tsai, C.-F., et al., Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. Journal of Business Research, 2021. 130: p. 200-209.

[73] Hussain, A.B., et al., Cross-market bankruptcy prediction: An interpretable ensemble learning framework using SHAP analysis. Computational Economics, 2026: p. 1-46.

[74] Papíková, L. and M. Papík, Similarity failure proximity score: A network-based metric for bankruptcy prediction. Finance Research Letters, 2026: p. 109631.

[75] Zhu, X., et al., Predicting financial distress using textual risk disclosures in annual reports: How and what risks are disclosed? The British Accounting Review, 2026: p. 101836.

[76] Qin, S., M. Bakoush, and F. McGroarty, Using Machine Learning to Detect Financial Distress From Sustainability Reports. Business Strategy and the Environment, 2026.