The Economics and Finance Letters
2026 Vol. 13, No. 3, pp. 94-116
ISSN(e): 2312-430X
ISSN(p): 2312-6310
DOI: 10.18488/29.v13i3.5122

Predicting credit rating downgrades in emerging markets: A comparative analysis of artificial neural networks and traditional binary classification models

Jiroj Buranasiri[1]
Prajya Ngamjan[2+]
Nuttawaree Ratchpiboon[3]

[1,2]*College of Innovation, Thammasat University, Bangkok, Thailand.*
[1]*Email: jirojb@gmail.com*
[3]*Email: nattawaree7719@gmail.com*
[2]*Faculty of Liberal Arts and Management Science, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus, Sakon Nakhon, Thailand.*
[2]*Email: prajyan15@gmail.com*


(+ Corresponding author)

Article History
Received: 20 April 2026
Revised: 2 June 2026
Accepted: 30 June 2026
Published: 25 August 2026

Keywords
Credit risk
Downgrade prediction
Emerging markets
Financial forecasting
Machine learning.

JEL Classification:
G24; G32; C45; C53.

ABSTRACT

This research compares the effectiveness of machine-learning and traditional statistical techniques in predicting annual credit rating downgrades for Thai non-financial firms listed on the Stock Exchange of Thailand during 2018–2023, using a time-ordered train-validation-test framework for predictive model evaluation. The machine-learning models are the artificial neural network (ANN), Random Forest, and XGBoost. The traditional techniques are the linear probability model, logistic regression, and probit regression. Model performance is evaluated using threshold-free PR-AUC and threshold-based metrics because downgrade events occur infrequently. The findings show that logistic regression is more useful for risk ranking, while Random Forest is more useful for downgrade screening. Under threshold-free metrics, logistic regression has the highest PR-AUC of 0.214. Under threshold-based metrics, Random Forest performs best, with a recall of 0.455 and an F1 score of 0.233, while the ANN is second-ranked. Overall, the results show that the preferred model depends on the purpose of use. Logistic regression is more suitable when the objective is to rank firms by downgrade risk, while Random Forest is more suitable when the objective is to screen firms for possible downgrade. Given the limited number of downgrade events in the test sample, the model comparisons should be interpreted as indicative rather than definitive.

Contribution/Originality: This study contributes to the existing literature by comparing ANN, Random Forest, XGBoost, logistic regression, probit regression, and the linear probability model in a small Thai credit-rating sample, documenting that the preferred model depends on the purpose of use: Random Forest for downgrade screening and logistic regression for risk ranking.

1. INTRODUCTION

Rating agencies, such as Standard & Poor's, Moody's, and Fitch, evaluate a firm's financial condition, business risk, and debt-servicing capacity before assigning letter-grade ratings (Treacy & Carey, 2000; White, 2010). Credit ratings summarize creditworthiness and provide signals that investors and creditors can use in financial markets (Partnoy, 1999; White, 2010). Rating changes are not merely reporting events; prior studies show that they can affect bond and stock prices as well as firms' financing decisions (Hand, Holthausen, & Leftwich, 1992; Kisgen, 2006). Accordingly, downgrades are important to firms because they are associated with negative market reactions and tighter financing conditions (Dichev & Piotroski, 2001; Hand et al., 1992; Kisgen, 2006).

A downgrade may be especially costly in emerging markets because firms' access to external capital depends partly on governance quality, legal institutions, financial-market development, and capital-structure conditions

(Claessens & Yurtoglu, 2013; Fan, Titman, & Twite, 2012). In this paper, corporate rating data are taken from TRIS Rating, a major domestic credit rating agency in Thailand (TRIS Rating Co Ltd, 2024). Predicting downgrades among Thai listed firms is difficult because the number of rated firms is limited and rating changes are infrequent. In this study, downgrades occur in only 77 of 558 firm-year observations, and this situation creates a rare-event classification problem.

Altman's Z-score framework is widely used for bankruptcy prediction. It combines several accounting ratios that reflect a firm's financial condition into one discriminant score, and many researchers use it as a benchmark in financial-distress studies (Altman, 1968; Radovanovic & Haas, 2023; Wu, Gaunt, & Gray, 2010; Zoričák, Gnip, Drotár, & Gazda, 2020). Changes in the Z-score can be read as signals of changes in a firm's financial condition. Accordingly, rather than using the Z-score level itself, this paper uses year-on-year changes in its determinants, since rating changes may reflect agencies' reassessment of firms' underlying creditworthiness over time (Altman & Rijken, 2006; Löffler, 2004).

For a binary outcome such as downgrade versus no downgrade, logit, probit, and the linear probability model provide the natural parametric benchmarks (Hosmer, Lemeshow, & Sturdivant, 2013; Long, 1997). These models provide transparent parametric benchmarks against which the ANN and tree-based models can be compared (Greene, 2012). Their limitation is that they may not capture nonlinear relationships or interactions among the financial-ratio changes (Hastie, Tibshirani, & Friedman, 2009).

Earlier credit-risk studies have used neural networks for bankruptcy prediction, credit scoring, and related classification problems (Shin & Han, 1999; Tam & Kiang, 1992). The reason for using an ANN in this study is that changes in financial ratios may be associated with downgrade risk in a nonlinear way. ANNs can learn such patterns through hidden layers and activation functions (Goodfellow, Bengio, & Courville, 2016; Rumelhart, Hinton, & Williams, 1986).

This paper applies a time-ordered out-of-sample prediction design, with model development and evaluation conducted year by year to predict annual credit rating downgrades. The empirical setting is deliberately compact, covering 93 firms from 2018 to 2023. Because only 77 observations are downgrade cases, focusing only on models' accuracy would result in an incomplete performance overview. The analysis, thus, employs recall, F1, MCC, and PR-AUC to evaluate model performance under rare-event conditions.

To avoid a narrow benchmark comparing the ANN only with traditional classifiers and to make the benchmark comparison more comprehensive, the study's comparison places the ANN alongside three parametric models, logit, probit, and the linear probability model, and two nonlinear tree-based models, Random Forest and XGBoost (Breiman, 2001; Chen & Guestrin, 2016). Random Forest and XGBoost provide a useful check on the ANN results because they can capture nonlinearities and interactions through tree-based ensemble structures rather than neural-network layers.

Our objective is not only to rank models by predictive performance but also to clarify how performance depends on the operational definition of success. Because the decision threshold that maps predicted probabilities into downgrade flags is itself a modeling choice, we optimize thresholds on the validation set under two regimes, an F1-optimized threshold and an MCC-optimized threshold, before applying the selected rules to the 2023 test set. We also use PR-AUC to complement threshold-based results with a threshold-free view of ranking quality.

The central research question is how different classification models perform in annual credit rating downgrade prediction when the sample is small and downgrade events are rare. This paper, hence, compares the created models under three criteria: screening downgrade cases, balancing errors across downgrade and non-downgrade classes, and ranking firms by downgrade risk.

This study contributes to the literature in three aspects: comparing ANNs with traditional parametric models and modern ensemble methods in a small, imbalanced emerging-market setting, applying a time-ordered train-validation-test design with rare-event-focused metrics, namely, recall, F1, MCC, and PR-AUC, so that the evaluation

reflects forecasting performance rather than in-sample fit, and showing that model rankings vary with the evaluation metric and decision regime.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 explains the data and variables. Section 4 describes the methodology. Section 5 reports the results, and Section 6 discusses the findings before the conclusion.

2. LITERATURE REVIEW

2.1. Credit Ratings and Their Economic Consequences

Credit ratings are important financial information because they summarize a firm's creditworthiness in a form that investors and creditors can use. Rating agencies, such as Standard & Poor's, Moody's, and Fitch, review a firm's financial condition, business risk, and debt-servicing capacity before assigning a letter-grade rating (Treacy & Carey, 2000; White, 2010). Rating changes also give information to the market and may lead investors and creditors to reassess a firm's credit risk (Partnoy, 1999; White, 2010).

Rating changes affect financing and investment conditions for firms, especially downgrades. Prior research identifies three adverse consequences of a credit rating downgrade: negative stock-market reactions, higher financing costs, and more restrictive investment conditions (Dichev & Piotroski, 2001; Hand et al., 1992; Kisgen, 2006). Therefore, from a credit-surveillance view, detecting financial deterioration associated with rating actions is useful for monitoring credit risk (Thomas, Crook, & Edelman, 2017).

2.2. Financial Distress Prediction and the Altman Z-Score

Financial distress prediction assumes that weakening financial ratios can warn of a firm's failure before it really occurs. Several studies, for example, the studies of Beaver (1966), Altman (1968), Ohlson (1980), and Shumway (2001) use financial ratios in their distress prediction. The revised Altman Z-score is still useful in credit-risk measurement as it is straightforward to implement and excludes the sales-to-assets ratio, which is usually sensitive to industry structure (Altman, 2013). The score's components cover liquidity, accumulated profitability, operating performance, and market valuation. Accordingly, deteriorated movements in these ratios may indicate financial weakening. The rating-behavior studies report the supportive evidence that the movements in financial fundamentals may precede rating transitions (Altman & Rijken, 2006; Löffler, 2004). The information in these financial data is valuable for financial distress prediction.

2.3. Traditional Binary Classification Models

The prediction of rating change for this study is under the two possible outcomes: a company will be downgraded or not be downgraded. Consequently, the suitable traditional models for this study are logistic regression, probit regression, and the linear probability model. These models are often used when the outcome has only two choices, such as downgrade or no downgrade (Hosmer et al., 2013; Long, 1997). They use different functional forms, but each model links the explanatory variables to the estimated probability of downgrade. They are useful benchmarks because of their simplicity in the estimation process and relatively direct interpretation of coefficient signs and magnitudes (Greene, 2012). The explainability of the model is significant in credit analysis because users often need to understand why a firm is classified as risky. However, these linear-index models may miss nonlinearities or interactions among the financial-ratio changes (Hastie et al., 2009).

2.4. Artificial Neural Networks in Credit Risk

Artificial neural networks provide an alternative complement to traditional binary-response models. Through hidden layers and activation functions, ANNs learn relationships that are difficult to impose through logit, probit, or other generalized linear specifications (Goodfellow et al., 2016; Rumelhart et al., 1986). Thus, earlier neural-network

and related computational credit-risk studies have applied these methods to bankruptcy prediction, credit scoring, corporate bond rating, and related classification problems (Shin & Han, 1999; Tam & Kiang, 1992).

For credit-risk problems, the motivation for using ANNs relies on their practicality. They can discover nonlinear interactions among predictors by themselves. There is no necessity for the researcher to specify those interactions in advance (Tam & Kiang, 1992). Studies comparing credit-risk models report that the ANNs demonstrate good performance, even though their advantage varies by dataset and evaluation criterion (Lessmann, Baesens, Seow, & Thomas, 2015). Recent hybrid deep-learning work confirms that model capacity and architecture may matter when the predictive signal is complex (Lin, Padliansyah, Lu, & Liu, 2025).

However, ANNs also require careful tuning and may be less transparent than simpler statistical models (Rudin, 2019).

2.5. Class Imbalance and Evaluation Metrics

Credit rating downgrades are less frequent than rating affirmations. This turns downgrade prediction into an imbalanced classification problem (He & Garcia, 2009). A model may correctly classify many non-downgrade firms, yet still fail to detect many downgraded firms.

Accuracy is therefore a weak guide in this setting. It can look high even when minority-class detection is poor. He and Garcia (2009) note that imbalanced data need careful treatment in model training and evaluation. Saito and Rehmsmeier (2015) also show that precision-recall curves are often more useful than ROC curves when the class distribution is uneven.

The choice of metric also has a practical meaning. Recall is useful for screening because it measures how many downgrade cases are found. When false positives are also a concern, F1 and MCC add further information. F1 combines precision and recall, while MCC uses all four cells of the confusion matrix and is less affected by class imbalance than accuracy (Chicco & Jurman, 2020).

2.6. Research Gaps

Three gaps guide this study. First, much recent machine-learning work on corporate credit ratings uses broad or multi-country datasets. These datasets are useful for model development, but they may not reflect the conditions of a small domestic rating market. Country-specific evidence is therefore still needed, especially in emerging markets where rating coverage is limited and downgrade events are relatively infrequent. Second, the data split matters for forecasting. Random splits can mix observations from different years, so future information may enter the evaluation (Qian et al., 2022). Third, when downgrade cases are rare, common performance measures may make a model look stronger than it really is. A model may rank observations reasonably well while still doing poorly on the minority class. Precision-recall analysis is therefore more informative in this setting because it focuses directly on the trade-off between detecting downgrades and avoiding false alarms (Saito & Rehmsmeier, 2015). This study addresses these three issues by focusing on TRIS-rated Thai firms, using a strict chronological data partition, and reporting both threshold-based metrics, namely F1 and MCC, and threshold-free PR-AUC.

In the case of emerging markets, several researchers often use accounting-based indicators similar to those used in this study but with a variety of predictor constructions. Their conclusions about model performance are varied. Rahman and Zhu (2024) examine Chinese listed companies in the construction sector, and they compare several Z-score variants, including the original Altman model and alternative techniques, with machine-learning classifiers. They find that machine-learning methods perform better than conventional Z-score benchmarks when evaluated by AUC and precision-recall measures. De Oliveira and Basso (2025) study corporate credit ratings across several countries and find that ANN-based models can outperform traditional techniques when firms' risk profiles contain nonlinear patterns. Evidence from ASEAN industrial firms also shows that accounting-ratio-based distress classifications differ across countries and periods, which supports the need for market-specific model design (Yunisa

& Santi, 2023). Taken together, these studies suggest that accounting variables remain useful in emerging-market credit analysis, but their performance depends on the model, the market, and the outcome being predicted. The Thai TRIS-rated sample brings these issues into one setting.

Accordingly, the empirical analysis compares an ANN with five benchmark models: logistic regression, probit regression, the linear probability model, Random Forest, and XGBoost. The focus is on annual credit rating downgrade prediction for Thai listed non-financial firms using a time-ordered train-validation-test design. Because the sample is small and downgrade cases are limited, the study separates training, validation, and testing by year instead of using a random split. Model performance is evaluated with F1, MCC, and PR-AUC. The analysis then checks whether the model ranking changes when the evaluation is tied to different credit-monitoring tasks.

3. DATA AND SAMPLE DESCRIPTION

3.1. Data Sources and Sample Construction

The analysis is based on annual firm-year data for Thai listed non-financial firms during 2018–2023. Financial statement data are taken from publicly available filings submitted to the Stock Exchange of Thailand (SET). Credit rating data are collected from TRIS Rating. A firm-year observation is included only when both the rating data and the required financial variables are available.

The final sample has 558 firm-year observations from 93 firms over six years. Banks, insurance companies, securities companies, and other financial firms are excluded because their regulation, leverage, and balance-sheet structure are different from those of non-financial firms. For this reason, their financial ratios are less comparable (Rajan & Zingales, 1995). If a firm-year observation lacks any required variable, it is excluded for that year. If the data are available in a later year, the firm re-enters the sample for that year.

3.2. Variable Construction

3.2.1. Dependent Variable: Credit Rating Downgrade

Credit rating downgrade, ΔY, is coded at the firm-year level. It equals 1 if the rating in year t is lower than the rating in year $t-1$, and 0 otherwise. Affirmed and upgraded ratings are coded as 0. Accordingly, the empirical task is to predict the annual downgrade indicator ΔY_t using year-on-year changes in the financial-ratio determinants measured over the same annual interval. In this study, a downgrade means any move to a lower long-term issuer rating on the domestic rating scale. For example, a move from A to A− or from BBB+ to BBB is counted as a downgrade.

3.2.2. Independent Variables: Changes in Altman Z-Score Determinants

Table 1 exhibits the definition of four explanatory variables that measure year-on-year changes in the Z-score determinants, based on the revised Altman Z-score framework (Altman, 2013).

In all model specifications, the explanatory variable vector consists of Δx_1, Δx_2, Δx_3, and Δx_4, corresponding to changes in working capital to total assets, retained earnings to total assets, EBIT to total assets, and market value of equity to book value of liabilities, as exhibited in Table 1.

All four predictors are constructed as year-on-year changes rather than levels because a firm whose profitability or liquidity is deteriorating may carry a different downgrade risk from another firm with the same current ratio but a more stable trajectory. This difference may not be captured when the model uses only ratio levels (Campbell, Hilscher, & Szilagyi, 2008; Shumway, 2001). The change specification keeps the focus on how each firm's financial ratios move from one year to the next.

Table 1. Variable Definitions.

Variables	Definition	Formula	Interpretation
Δx_1	Change in working capital to total assets ratio	$\left(\frac{\text{WC}}{\text{TA}}\right)_t - \left(\frac{\text{WC}}{\text{TA}}\right)_{t-1}$	This variable represents changes in short-term liquidity and operating efficiency. Positive values suggest an improving liquidity position.
Δx_2	Change in retained earnings to total assets ratio	$\left(\frac{\text{RE}}{\text{TA}}\right)_t - \left(\frac{\text{RE}}{\text{TA}}\right)_{t-1}$	This variable reflects changes in cumulative profitability and financial stability. Positive values mean an increase in retained earnings.
Δx_3	Change in EBIT to total assets ratio	$\left(\frac{\text{EBIT}}{\text{TA}}\right)_t - \left(\frac{\text{EBIT}}{\text{TA}}\right)_{t-1}$	This variable measures changes in operating profitability and earnings-generating capacity. Positive values mean stronger operating performance.
Δx_4	Change in market value of equity to the book value of total liabilities ratio	$\left(\frac{\text{MVE}}{\text{BVL}}\right)_t - \left(\frac{\text{MVE}}{\text{BVL}}\right)_{t-1}$	This variable captures changes in market valuation and investors' forward-looking assessments. Positive values indicate improved market perception relative to debt obligations.
ΔY	Binary downgrade indicator (Dependent variable)	1 if rating downgrade in year t relative to year $t-1$; 0 otherwise	This variable equals 1 if the firm experiences a reduction in long-term issuer credit rating; 0 for affirmations and upgrades.

Note: WC = Working Capital; TA = Total Assets; RE = Retained Earnings; EBIT = Earnings Before Interest and Taxes; MVE = Market Value of Equity; BVL = Book Value of Total Liabilities. All change variables are calculated as the year-on-year first difference of the respective ratio.

3.3. Sample Characteristics

Table 2 reports the annual number of observations and downgrade events from 2018 to 2023.

Table 2. Sample Description by Year.

Year	Observations	Downgrades	Downgrade Rate
2018	93	19	20.4%
2019	93	8	8.6%
2020	93	17	18.3%
2021	93	12	12.9%
2022	93	10	10.8%
2023	93	11	11.8%
Total	558	77	13.8%

Note: The downgrade rate is the annual number of downgrades divided by annual observations.

These figures show why accuracy alone can be misleading: downgrades account for only 13.8% of the full sample. The annual downgrade rate ranges from 8.6% in 2019 to 20.4% in 2018. The rate is also high in 2020, possibly because of pandemic-period stress. During 2021–2023, the rate is narrower, between 10.8% and 12.9%.

3.4. Temporal Data Splits

The sample is grouped by calendar year. The training period covers 2018–2021, with 372 firm-year observations and 56 downgrades, or a downgrade rate of 15.1%. The 2022 observations are used for validation and include 93 observations with 10 downgrades. The final test set is from 2023, which contains 93 observations with 11 downgrades.

Given the relatively small sample size and the limited number of downgrade events, overfitting is a particular concern for flexible models such as the ANN and XGBoost. For this reason, the ANN is specified parsimoniously and regularized through dropout, L2 regularization, and early stopping, while XGBoost is constrained using shallow trees, a low learning rate, subsampling, and L2 regularization.

This split follows the timing that a credit analyst would face in practice. The model is fitted on past data, tuning decisions are made on a later year, and the final check is made on a still later year that was not used in model selection. In this paper, the 2022 validation set is used for two tasks: early stopping for the ANN and threshold selection for all models.

3.5. Descriptive Statistics

The four predictors, Δx_1, Δx_2, Δx_3, and Δx_4, vary widely across firm-years and include some extreme observations. Some firms improve from one year to the next, while others weaken, so the change variables naturally take both positive and negative values. Correlation analysis, not reported for brevity, indicates weak to moderate associations among the four variables. This indicates that these variables are associated. Nevertheless, each represents a different dimension of firms' financial health. Because extreme observations are a common characteristic of this type of data, their influence is examined separately through winsorization-based robustness checks later in the paper.

4. METHODOLOGY

4.1. Model Specifications

The study compares six binary classifiers: ANN, logistic regression, probit regression, the linear probability model, Random Forest, and XGBoost. Each model estimates the probability that firm *i* experiences a credit rating downgrade in year *t*, conditional on year-on-year changes in the Altman Z-score determinants over the same annual interval. Random Forest and XGBoost are included to assess whether the main findings are specific to the ANN or instead reflect a broader advantage of flexible nonlinear classifiers.

4.1.1. Artificial Neural Network (ANN)

Because the sample is limited and downgrade events are relatively rare, the ANN is specified as a parsimonious and regularized model. This approach is consistent with the literature treating architecture and hyperparameter choice as a model-selection problem in which model capacity should be matched to the prediction setting to support generalization (Baum & Haussler, 1989; Goodfellow et al., 2016; Murata, Yoshizawa, & Amari, 1994). Overfitting is addressed through dropout and early stopping (Goodfellow et al., 2016; Prechelt, 1998; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The final specification was selected through iterative tuning based on validation-set performance in a small-sample, class-imbalanced setting.

The final model is a feedforward multilayer perceptron with four standardized inputs, two hidden layers (10 and 6 neurons, tanh activation), L2 regularization, and dropout. A single sigmoid output neuron is trained using binary cross-entropy loss. Because downgrade events represent only 13.8% of the sample, we assign a training weight of 2.5 to the downgrade class and 1.0 to the non-downgrade class, which is a standard classifier-level adjustment for imbalanced learning (He & Garcia, 2009). Optimization is carried out using the Adam optimizer (Kingma & Ba, 2014) with the learning rate set to 0.0007 in the final specification. Training is conducted for up to 150 epochs with a batch size of 16, and early stopping based on validation loss with a patience of 15 epochs is used to improve out-of-sample generalization.

4.1.2. Logistic Regression

Logistic regression models the log-odds of a downgrade as a linear function of the predictors.

$$\log\left(\frac{P(\Delta Y_{i,t}=1|\Delta X_{i,t})}{1-P(\Delta Y_{i,t}=1|\Delta X_{i,t})}\right)=\beta_0+\beta_1\Delta x_1+\beta_2\Delta x_2+\beta_3\Delta x_3+\beta_4\Delta x_4 \quad (1)$$

Let $\Delta X_{i,t}=(\Delta x_1,\Delta x_2,\Delta x_3,\Delta x_4)$ denote the vector of explanatory variables, where each term represents the year-on-year change in a revised Altman Z-score determinant for the firm i in year t. The model is estimated with L2

regularization and C = 1.0. No direct class weighting is applied in the baseline specification. Instead, class imbalance is handled primarily through validation-based threshold optimization.

4.1.3. Probit Regression

Probit regression models the probability of a downgrade using the cumulative standard normal distribution.

$$P(\Delta Y_{i,t} = 1 | \Delta X_{i,t}) = \Phi(\beta_0 + \beta_1 \Delta x_1 + \beta_2 \Delta x_2 + \beta_3 \Delta x_3 + \beta_4 \Delta x_4) \quad (2)$$

Where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The model is estimated by maximum likelihood. Unlike logistic regression and the linear probability model, this implementation does not directly support class weighting. Class imbalance is therefore addressed through the threshold optimization procedure described in Section 4.3.

4.1.4. Linear Probability Model

The linear-probability model applies ordinary least squares directly to the binary outcome.

$$P(\Delta Y_{i,t} = 1 | \Delta X_{i,t}) = \beta_0 + \beta_1 \Delta x_1 + \beta_2 \Delta x_2 + \beta_3 \Delta x_3 + \beta_4 \Delta x_4 \quad (3)$$

The model is estimated using scikit-learn's LinearRegression class. Although this specification is simple and easy to interpret, fitted values are not constrained to the [0,1] interval. Predicted values are therefore clipped to that interval before threshold application.

For purposes of coefficient interpretation, we re-estimated the logistic and probit models on the training sample using maximum-likelihood estimation. This additional estimation is included to support coefficient interpretation, while the classification results reported later are based on the predictive model implementations described in this section.

4.1.5. Additional Benchmark Models: Random Forest and XGBoost

To assess whether the main findings are specific to the ANN or instead reflect a broader advantage of flexible nonlinear classifiers, two additional benchmark models are estimated: Random Forest and XGBoost. Both are tree-based ensemble methods that can capture nonlinear relationships and interactions without requiring explicit functional-form specification. Their inclusion is also consistent with the broader benchmarking literature in credit scoring and tabular prediction, which treats tree-based ensembles as strong comparators to both traditional statistical models and neural networks (Breiman, 2001; Chen & Guestrin, 2016; Grinsztajn, Oyallon, & Varoquaux, 2022; Lessmann et al., 2015). In the final specification, Random Forest is estimated with 500 trees, maximum depth of 4, minimum leaf size of 3, and square-root feature subsampling, while XGBoost is estimated with 300 trees, maximum depth of 3, learning rate of 0.05, subsample and column-subsample ratios of 0.90, minimum child weight of 1, and an L2 regularization parameter of 1.0. These values are treated as final implementation settings selected on the basis of validation performance. In the baseline specification, neither model uses direct class weighting; as with the traditional benchmark models, class imbalance is handled primarily through validation-based threshold selection.

4.2. Data Preprocessing

Before estimation, the data are processed in a way that preserves a realistic forecasting setting. In the baseline specification, extreme observations are not trimmed or winsorized, so the models are evaluated under conditions closer to practical use. The influence of extreme values is examined separately through the winsorization-based robustness checks reported in Section 4.5.3.

For the ANN, logistic regression, and probit models, the input features are standardized to zero mean and unit variance using the mean and standard deviation of the training set. The same transformation is then applied to the validation and test sets. The linear probability model is estimated using unstandardized features in order to preserve straightforward coefficient interpretation.

To address class imbalance, the ANN applies a moderate class-weighting scheme in which the non-downgrade class receives a weight of 1.0 and the downgrade class receives a weight of 2.5. This choice is more moderate than a fully inverse-frequency rule and is intended to improve minority-class detection without generating an excessive number of false positives in a small-sample setting. For logistic regression, class imbalance is addressed primarily through validation-based threshold optimization rather than direct class weighting.

4.3. Threshold Optimization

Binary classifiers produce continuous scores or probabilities that must be converted into class predictions through a decision threshold. In imbalanced settings, the conventional threshold of 0.5 is often inappropriate because it can lead to poor detection of the minority class. We therefore use validation-based threshold optimization rather than relying on a fixed cutoff.

The two threshold rules are used because they answer different practical questions. The F1-optimized threshold fits a screening use case: it gives more weight to finding downgrade cases, while still penalizing false alarms through precision. The MCC-optimized threshold is more conservative because it considers all four cells of the confusion matrix and gives a better sense of balance between downgrade and non-downgrade classifications. Reporting both rules avoids making the results depend on one cutoff choice and shows how each model behaves under two plausible credit-monitoring settings.

For each model, we evaluate a grid of 990 candidate thresholds ranging from 0.010 to 0.999 in increments of 0.001. At each threshold, the F1 score and the Matthews Correlation Coefficient (MCC) are computed on the validation set. For each regime, the threshold is selected from the validation set. The F1-based regime uses the threshold with the highest F1 score, while the MCC-based regime uses the threshold with the highest MCC. These thresholds are then applied unchanged to the 2023 test set, keeping the test evaluation out of sample.

F1 combines precision and recall.

$$F1 = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (4)$$

This measure is useful for screening because it gives credit for detecting downgrades while still accounting for false alarms.

The MCC is defined as:

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC ranges from -1 to +1. Higher values indicate better classification. A value near 0 means that the model adds little beyond random classification. Since MCC uses TP, TN, FP, and FN together, it helps assess performance for both downgrade and non-downgrade cases.

Recall is important for imbalanced data, but this study does not use recall as the only threshold rule. A very low threshold can raise recall by flagging more firms as possible downgrades. This reduces missed downgrades, but it can also create too many false positives. In practice, analysts need a screening rule that finds downgrade candidates without sending too many non-downgrade firms for review. Thus, the F1 score can keep the recall–precision trade-off explicit. MCC is applied as a second threshold rule. It considers all four cells of the confusion matrix in performance evaluation and is less vulnerable than accuracy to class imbalance (Chicco & Jurman, 2020).

4.4. Performance Metrics

This paper reports the test-set results across multiple classification measures because no single metric is adequate for providing a complete answer for the models' performance. Accuracy measures the share of correctly classified observations, (TP + TN) / (TP + TN + FP + FN). Precision measures how many firms identified as downgrades

were actually downgraded. It is calculated as TP / (TP + FP). Recall, also called sensitivity, measures how many downgraded firms were detected by the model among the firms that were actually downgraded. It is calculated as TP / (TP + FN). Specificity measures the proportion of actual non-downgrade cases and is calculated as TN / (TN + FP).

F1 and MCC are reported because they summarize different aspects of threshold-based performance. F1 emphasizes the precision-recall trade-off, while MCC provides a broader balance measure based on all four cells of the confusion matrix.

PR-AUC is used because it does not rely on one cutoff. It summarizes the precision-recall trade-off across many possible thresholds. For imbalanced data, this is more informative than measures mainly influenced by the majority class (Saito & Rehmsmeier, 2015).

4.5. Robustness Checks

Robustness checks examine whether the main findings depend on the baseline design. The baseline results come from one train-validation-test split. Five additional checks are conducted.

4.5.1. Bootstrap Confidence Intervals

The 2023 test set is small, so the point estimates may look more precise than they are. To show this uncertainty, the paper uses bootstrap resampling. For each model and threshold regime, F1-optimized and MCC-optimized, 2,000 samples are drawn with replacement from the test set. The metrics are recalculated for each sample. The bootstrap results report the median and the 95% confidence interval, using the 2.5th and 97.5th percentiles.

4.5.2. Rolling-Origin Re-Estimation

The study also checks whether the results are stable across nearby forecasting windows. In the additional rolling-origin window, the models are trained on 2018–2020 data, validated on 2021 data, and tested on 2022 data. The baseline split is then stated in the same format, with training on observations from 2018 to 2021, validation on observations in 2022, and testing on observations in 2023. The first window gives an additional temporal check, while the baseline window provides the reference case.

All models are re-estimated within each window. Thresholds are re-selected using the relevant validation year. The final performance is measured on the corresponding test year. As a result, it is possible to see whether the model rankings hold across nearby forecast periods, rather than only in the 2023 test year.

4.5.3. Outlier Sensitivity via Winsorization

Financial ratios can rapidly change from year to year, especially when the denominator is small or when firms encounter unusual shocks. For that reason, the paper compares the baseline results, which use the raw predictors, with two winsorized versions of the data. In the robustness checks, the predictors are winsorized at the 1.0% and 2.5% tails. For each split, winsorization cutoffs are calculated from the training sample and then applied unchanged to the validation and test samples. This allows us to see whether the main model rankings change when the largest year-on-year movements in the predictors are pulled back to less extreme values.

4.5.4. Precision-Recall Area Under the Curve (PR-AUC)

In addition to threshold-specific metrics, we evaluate model performance using precision-recall area under the curve (PR-AUC) on both the validation and test sets. PR-AUC provides a threshold-independent summary of performance and is particularly informative in imbalanced settings (Saito & Rehmsmeier, 2015). We report PR-AUC for the baseline specification and the main robustness checks to complement the threshold-optimized F1 and MCC results.

4.5.5. COVID-19 Exclusion Robustness Check

As a supplementary robustness check, we examine whether the main results are sensitive to excluding COVID-affected years from the training sample. Three scenarios are considered: (1) the baseline specification, which trains on 2018–2021 and evaluates on 2023 after validation on 2022; (2) a specification that excludes 2020 from the training sample; and (3) a more restrictive specification that excludes both 2020 and 2021. For each scenario, we report results under the F1-optimized and MCC-optimized threshold regimes together with test-set PR-AUC. This check shows whether the model rankings change when pandemic-period observations are removed from training.

4.6. Implementation

All analyses are conducted in Python. NumPy and pandas are used for data handling, scikit-learn for preprocessing and the logistic, linear, and Random Forest models, statsmodels for probit estimation, TensorFlow/Keras for the ANN, XGBoost for gradient-boosted tree estimation, and SciPy for selected auxiliary procedures, including winsorization-based robustness analysis.

5. RESULTS

This section reports the out-of-sample results for the 2023 test set and then presents the main robustness checks, including threshold-free PR-AUC comparisons, rolling-origin re-estimation, winsorization, and COVID-period exclusion.

5.1. Coefficient Estimates for Benchmark Models

Table 3 reports the estimated coefficients for the logistic, probit, and linear probability models, providing insight into the relationship between the revised Altman Z-score determinants and downgrade probability.

Table 3. Coefficient Estimates for Benchmark Models.

Variables	Logistic Regression	Probit Regression	Linear Probability Model
Constant	-1.8003*** (0.1530)	-1.0722*** (0.0827)	0.1498*** (0.0185)
Δx_1	0.0176 (0.1489)	0.0137 (0.0801)	0.0464 (0.2259)
Δx_2	-0.0137 (0.1576)	-0.0071 (0.0925)	-0.0004 (0.0015)
Δx_3	-0.4308*** (0.1567)	-0.2545*** (0.0903)	-1.2536*** (0.4011)
Δx_4	0.2287 (0.1465)	0.1372 (0.0870)	0.0420 (0.0270)
Observations	372	372	372
Pseudo R^2/R^2	0.0327	0.0348	0.0280

Note: Standard errors are reported in parentheses. The logistic and probit models are estimated using standardized predictors. The linear probability model is estimated using the same raw predictors; standard errors for the linear probability model are heteroskedasticity-robust HC1 standard errors estimated using OLS. The reported R^2 for the linear probability model is the conventional coefficient of determination, while the logistic and probit models report pseudo R^2 values, which are not directly comparable. Δx_1 = Change in working capital to total assets; Δx_2 = Change in retained earnings to total assets; Δx_3 = change in EBIT to total assets; Δx_4 = Change in market value of equity to book value of liabilities. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In all three benchmark parametric models, Δx_3, the change in EBIT to total assets is negative and statistically significant at the 1% level. This suggests that deterioration in operating performance is associated with a higher probability of an annual credit rating downgrade. However, Δx_1, Δx_2, and Δx_4 are not statistically significant. The relatively low pseudo R^2 and R^2 values are consistent with the difficulty of predicting rare downgrade events in this small emerging-market sample.

5.2. Baseline Results

The 2023 out-of-sample test set contains 93 firm-year observations, including 11 downgrade events. Tables 4 and 5 report the baseline test-set results under the two validation-based threshold-selection regimes, namely the F1-optimized threshold and the MCC-optimized threshold.

Table 4. Baseline Test-Set Performance (F1-Optimized Threshold).

Model	Threshold	Accuracy	Precision	Recall	Specificity	F1	MCC	TP	TN	FP	FN
ANN	0.209	0.677	0.148	0.364	0.720	0.211	0.059	4	59	23	7
Logistic Regression	0.187	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
Probit Regression	0.192	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
Linear Probability Model	0.173	0.699	0.095	0.182	0.768	0.125	-0.039	2	63	19	9
Random Forest	0.158	0.645	0.156	0.455	0.671	0.233	0.085	5	55	27	6
XGBoost	0.146	0.634	0.129	0.364	0.671	0.190	0.024	4	55	27	7

Note: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. Recall = TP/(TP + FN). Specificity = TN/(TN + FP). F1 = 2 × (Precision × Recall)/(Precision + Recall). MCC = Matthews Correlation Coefficient. Test set contains 93 observations (11 downgrades, 82 non-downgrades). Thresholds are selected by maximizing the F1 score on the 2022 validation set and are then applied unchanged to the 2023 test set.

Table 5. Baseline Test-Set Performance (MCC-Optimized Threshold).

Model	Threshold	Accuracy	Precision	Recall	Specificity	F1	MCC	TP	TN	FP	FN
ANN	0.183	0.538	0.119	0.455	0.549	0.189	0.002	5	45	37	6
Logistic Regression	0.191	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
Probit Regression	0.194	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
Linear Probability Model	0.173	0.699	0.095	0.182	0.768	0.125	-0.039	2	63	19	9
Random Forest	0.161	0.667	0.167	0.455	0.695	0.244	0.103	5	57	25	6
XGBoost	0.160	0.699	0.160	0.364	0.744	0.222	0.078	4	61	21	7

Note: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. Recall = TP/(TP + FN). Specificity = TN/(TN + FP). F1 = 2 × (Precision × Recall)/(Precision + Recall). MCC = Matthews Correlation Coefficient. Test set contains 93 observations (11 downgrades, 82 non-downgrades). Thresholds are selected by maximizing MCC on the 2022 validation set and are then applied unchanged to the 2023 test set.

As reported in Table 4, Random Forest records recall of 0.455, F1 of 0.233, and MCC of 0.085 at a threshold of 0.158, identifying 5 of the 11 downgrade cases. The ANN ranks second, with a recall of 0.364 and an F1 of 0.211 at a threshold of 0.209, correctly identifying 4 downgrade cases. XGBoost is also competitive, matching the ANN's recall of 0.364, although with a lower F1 score of 0.190. The parametric benchmark models flag far fewer downgrade cases. Logistic regression and probit regression each identify only 1 downgrade case, giving recall of 0.091 and F1 of 0.100. The linear probability model detects two downgrade cases, giving a recall of 0.182 and an F1 score of 0.125.

The MCC-optimized results in Table 5 lead to much the same conclusion. Random Forest remains the strongest model, with a recall of 0.455, an F1 score of 0.244, and an MCC of 0.103 at the selected threshold of 0.161. XGBoost is next, with a recall of 0.364, an F1 score of 0.222, and an MCC of 0.078. The ANN also captures five downgrade cases, the same number as Random Forest, but its lower F1 score of 0.189 and very small MCC of 0.002 indicate that this additional downgrade capture is achieved with many more false positives. The parametric models remain much more cautious. Logistic regression and probit regression each identify only one downgrade case, while the linear probability model identifies two.

5.3. Threshold-Free Evaluation

Table 6 reports PR-AUC for the validation and test sets. Unlike the threshold-based measures in Tables 4 and 5, PR-AUC does not rely on a single cutoff. It evaluates how well each model orders firms by downgrade risk over the full range of possible thresholds. This gives a useful second view of performance, especially because a model may rank firms well even if its selected classification threshold produces low recall.

For the 2023 test set, logistic regression records the highest PR-AUC, 0.214. Random Forest follows at 0.173, almost the same as the ANN at 0.172. Probit regression and the linear probability model each record 0.167, while XGBoost has the lowest value, 0.146. The validation results show a similar pattern, with logistic regression again producing the highest PR-AUC.

Table 6. Baseline Precision-Recall Area Under the Curve.

Model	Validation PR-AUC	Test PR-AUC
ANN	0.225	0.172
Logistic	0.247	0.214
Probit	0.245	0.167
Linear	0.244	0.167
Random Forest	0.191	0.173
XGBoost	0.198	0.146

5.4. Bootstrap Confidence Intervals

Table 7 reports the bootstrap confidence intervals for F1 and MCC under the two threshold-selection rules. The results are based on 2,000 bootstrap resamples from the 2023 test set. This additional check is important because, in such a small test set, one or two correctly classified downgrade cases can noticeably change the reported performance measures. Panel A presents the bootstrap results for F1. Random Forest has the highest median F1 under both threshold rules. Its median F1 is 0.227 under the F1-optimized threshold and 0.238 under the MCC-optimized threshold. The ANN follows, with median F1 values of 0.200 and 0.182. XGBoost records median F1 values of 0.182 and 0.214. The three parametric benchmark models, logistic regression, probit regression, and the linear probability model, remain below the leading nonlinear models in median F1. Panel B reports the bootstrap results for MCC. Random Forest again gives the strongest median values, with a median MCC of 0.079 under the F1-optimized threshold and 0.098 under the MCC-optimized threshold. XGBoost follows, with median MCC of 0.018 and 0.073. The ANN has a positive median MCC of 0.052 under the F1-optimized threshold, but its median MCC falls to -0.002 under the MCC-optimized threshold. The parametric models remain close to zero or below zero. Importantly, the 95% confidence intervals for MCC include zero for every model. This indicates that balanced classification performance is still statistically uncertain, even though Random Forest has the strongest point estimates in the baseline results.

Table 7. Bootstrap Confidence Intervals for F1 and MCC.

Panel A. Bootstrap Confidence Intervals for F1 Score					
Model	**Regime**	**Threshold**	**Median F1**	**95% CI Lower**	**95% CI Upper**
ANN	F1	0.209	0.200	0.051	0.389
Logistic	F1	0.187	0.095	0.000	0.300
Probit	F1	0.192	0.095	0.000	0.300
Linear	F1	0.173	0.121	0.000	0.294
Random Forest	F1	0.158	0.227	0.059	0.415
XGBoost	F1	0.146	0.182	0.048	0.359
ANN	MCC	0.183	0.182	0.043	0.328
Logistic	MCC	0.191	0.095	0.000	0.300
Probit	MCC	0.194	0.095	0.000	0.300
Linear	MCC	0.173	0.121	0.000	0.294
Random Forest	MCC	0.161	0.238	0.061	0.426
XGBoost	MCC	0.160	0.214	0.056	0.400
Panel B. Bootstrap Confidence Intervals for MCC					
Model	Regime	Threshold	Median MCC	95% CI Lower	95% CI Upper
ANN	F1	0.209	0.052	-0.152	0.279
Logistic	F1	0.187	-0.017	-0.141	0.232
Probit	F1	0.192	-0.017	-0.141	0.232
Linear	F1	0.173	-0.042	-0.200	0.176
Random Forest	F1	0.158	0.079	-0.121	0.307
XGBoost	F1	0.146	0.018	-0.171	0.245
ANN	MCC	0.183	-0.002	-0.193	0.203
Logistic	MCC	0.191	-0.017	-0.141	0.232
Probit	MCC	0.194	-0.017	-0.141	0.232
Linear	MCC	0.173	-0.042	-0.200	0.176
Random Forest	MCC	0.161	0.098	-0.110	0.328
XGBoost	MCC	0.160	0.073	-0.133	0.314

Note: Bootstrap results are based on 2,000 resamples with replacement from the 2023 test set. The test set includes 93 observations, of which 11 are downgrade cases. Panel A reports the bootstrap median and 95% confidence interval for F1, while Panel B reports the corresponding statistics for MCC.

5.5. Rolling-Origin Re-Estimation

Tables 8 and 9 present the rolling-origin results. Fold 1 uses 2018–2020 for training, 2021 for validation, and 2022 for testing. Fold 2 follows the baseline split, with 2018–2021 for training, 2022 for validation, and 2023 for testing. The purpose of this exercise is to check whether the main results are specific to the 2023 test year or whether they also appear in the neighboring forecast period.

The threshold-based results in Table 8 show that model performance changes noticeably across the two folds. In Fold 1, the ANN gives the strongest balance of performance, with an F1 score of 0.333, an MCC of 0.248, and a recall of 0.600. The parametric models also perform better than in the 2023 baseline. Logistic regression, probit regression, and the linear probability model each identify 60.0% of downgrade cases, with F1 scores ranging from 0.245 to 0.273. The two tree-based models follow a different pattern. XGBoost captures all 10 downgrade cases in the 2022 test set, giving a recall of 1.000, while Random Forest captures 9 of the 10 cases, giving a recall of 0.900. However, these high recall values come with many false positives, so their F1 scores remain moderate, 0.250 for XGBoost and 0.265 for Random Forest.

In Fold 2, which is the baseline 2023 test split, the results are weaker and the ranking changes. Random Forest has the highest F1 score, 0.233, followed by the ANN at 0.211 and XGBoost at 0.190. The three parametric models perform much less well than they did in Fold 1, particularly in downgrade recall.

Table 9 reports the same rolling-origin comparison using PR-AUC. The ranking is different from the threshold-based results. In Fold 1, probit regression records the highest test PR-AUC, 0.242, followed closely by logistic regression at 0.240, the ANN at 0.236, and the linear probability model at 0.235. XGBoost records a test PR-AUC of 0.213. Random Forest is lower than the other models in the same fold, with a value of 0.185. In Fold 2, the ranking changes again. Logistic regression leads with a test PR-AUC of 0.214, followed by Random Forest at 0.173 and the ANN at 0.172, while XGBoost falls to 0.146.

Table 8. Rolling-Origin Re-Estimation (F1-Optimized Threshold).

Fold	Model	Threshold	Accuracy	Precision	Recall	Specificity	F1	MCC	TP	TN	FP	FN
1	ANN	0.230	0.742	0.231	0.600	0.759	0.333	0.248	6	63	20	4
1	Logistic	0.156	0.624	0.162	0.600	0.627	0.255	0.143	6	52	31	4
1	Probit	0.157	0.602	0.154	0.600	0.602	0.245	0.127	6	50	33	4
1	Linear	0.174	0.656	0.176	0.600	0.663	0.273	0.169	6	55	28	4
1	Random Forest	0.114	0.462	0.155	0.900	0.410	0.265	0.198	9	34	49	1
1	XGBoost	0.042	0.355	0.143	1.000	0.277	0.250	0.199	10	23	60	0
2	ANN	0.209	0.677	0.148	0.364	0.720	0.211	0.059	4	59	23	7
2	Logistic	0.187	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
2	Probit	0.192	0.806	0.111	0.091	0.902	0.100	-0.007	1	74	8	10
2	Linear	0.173	0.699	0.095	0.182	0.768	0.125	-0.039	2	63	19	9
2	Random Forest	0.158	0.645	0.156	0.455	0.671	0.233	0.085	5	55	27	6
2	XGBoost	0.146	0.634	0.129	0.364	0.671	0.190	0.024	4	55	27	7

Note: Fold 1 trains on 2018–2020, validates on 2021, and tests on 2022. Fold 2 follows the baseline design, with training on 2018–2021, validation on 2022, and testing on 2023. TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

Table 9. Rolling-Origin Precision-Recall Area Under the Curve.

Fold	Model	PR-AUC (Validation)	PR-AUC (Test)
1	ANN	0.159	0.236
1	Logistic	0.168	0.240
1	Probit	0.169	0.242
1	Linear	0.168	0.235
1	Random Forest	0.148	0.185
1	XGBoost	0.172	0.213
2	ANN	0.225	0.172
2	Logistic	0.247	0.214
2	Probit	0.245	0.167
2	Linear	0.244	0.167
2	Random Forest	0.191	0.173
2	XGBoost	0.198	0.146

Note: PR-AUC refers to the precision-recall area under the curve. Fold 1 trains on 2018–2020, validates on 2021, and tests on 2022. Fold 2 follows the baseline design, with training on 2018–2021, validation on 2022, and testing on 2023. Since PR-AUC does not depend on a selected threshold, Table 9 should be read alongside the threshold-based results in Table 8.

5.6. Winsorization Sensitivity

Table 10 compares the baseline results, where the predictors are not winsorized, with two alternatives that winsorize the predictors at the 1.0% and 2.5% tails. Because there are many model-by-specification results, the table reports only the main metric for each threshold regime together with recall. Recall is included because downgrade capture is central to the screening task. Under the F1 regime, the headline metric is F1, which already includes precision. Under the MCC regime, the headline metric is MCC, which already reflects all four cells of the confusion matrix.

Under the F1-optimized regime, the leading model changes as the treatment of outliers becomes more restrictive. In the baseline case, Random Forest has the highest F1 score, 0.233, and recall, 0.455. The ANN follows with an F1 of 0.211 and a recall of 0.364. With 1.0% winsorization, the ANN moves ahead on F1, reaching 0.235, while its recall stays at 0.364. With 2.5% winsorization, the ANN reaches an F1 score of 0.242, while recall remains 0.364. Random Forest stays close to the ANN in all three specifications. Its F1 score is 0.233 in the baseline case, 0.227 with 1.0% winsorization, and 0.235 with 2.5% winsorization. XGBoost performs below the ANN and Random Forest, while the three parametric models have lower F1 scores.

The MCC-optimized results show a different pattern. Random Forest is more stable across the three outlier treatments. Its MCC is 0.103 in the baseline case, 0.094 with 1.0% winsorization, and 0.099 with 2.5% winsorization. Its recall changes only slightly, from 0.455 in the baseline and 1.0% winsorized cases to 0.364 at the 2.5% level. The ANN is more affected by winsorization. With 1.0% winsorization, its MCC rises from 0.002 to 0.069, and recall rises from 0.455 to 0.545. With 2.5% winsorization, recall falls to 0.091, although MCC rises to 0.086. Thus, stronger winsorization makes the ANN look better on the balance measure, but worse on downgrade capture. The parametric models also become more conservative at the 2.5% level. Logistic regression, probit regression, and the linear probability model each report MCC values of about 0.060 and recall of 0.091.

PR-AUC gives one further check. In the baseline specification, logistic regression has the highest test PR-AUC, 0.214. Random Forest follows at 0.173, and the ANN is almost the same at 0.172. With 1.0% winsorization, the differences become small: the linear probability model and probit regression each reach 0.178, while logistic regression and the ANN each record 0.177. Random Forest remains slightly lower at 0.173. With 2.5% winsorization, the ANN records the highest PR-AUC, 0.183.

Table 10. Winsorization Sensitivity Analysis.

Winsorization	Model	F1 Regime		MCC Regime		PR-AUC (Test)
		F1	Recall	MCC	Recall	
None (Baseline)	ANN	0.211	0.364	0.002	0.455	0.172
None (Baseline)	Logistic	0.100	0.091	−0.007	0.091	0.214
None (Baseline)	Probit	0.100	0.091	-0.007	0.091	0.167
None (Baseline)	Linear	0.125	0.182	-0.039	0.182	0.167
None (Baseline)	Random Forest	0.233	0.455	0.103	0.455	0.173
None (Baseline)	XGBoost	0.190	0.364	0.078	0.364	0.146
1.0% Winsor	ANN	0.235	0.364	0.069	0.545	0.177
1.0% Winsor	Logistic	0.111	0.182	−0.072	0.182	0.177
1.0% Winsor	Probit	0.111	0.182	-0.072	0.182	0.178
1.0% Winsor	Linear	0.111	0.182	-0.072	0.182	0.178
1.0% Winsor	Random Forest	0.227	0.455	0.094	0.455	0.173
1.0% Winsor	XGBoost	0.200	0.364	0.069	0.364	0.153
2.5% Winsor	ANN	0.242	0.364	0.086	0.091	0.183
2.5% Winsor	Logistic	0.111	0.091	0.060	0.091	0.175
2.5% Winsor	Probit	0.111	0.182	0.060	0.091	0.175
2.5% Winsor	Linear	0.108	0.182	0.060	0.091	0.176
2.5% Winsor	Random Forest	0.235	0.364	0.099	0.364	0.175
2.5% Winsor	XGBoost	0.174	0.364	0.059	0.364	0.145

Note: "None" refers to the baseline specification without winsorization. "1.0% Winsor" and "2.5% Winsor" refer to robustness checks in which the predictors are winsorized at the 1.0% and 2.5% tails, respectively. To keep the table compact, only the main metric for each threshold regime and the corresponding recall value are reported. F1 and recall under the F1 regime are taken from the F1-optimized output, MCC and recall under the MCC regime are taken from the MCC-optimized output, and PR-AUC is reported separately as a threshold-free measure.

5.7. COVID-Period Exclusion

Tables 11 and 12 report the COVID-period exclusion robustness check. In addition to the baseline specification, which trains on 2018–2021, we examine two alternative training windows, one that excludes 2020 and another that excludes both 2020 and 2021. All scenarios continue to use 2022 for validation and 2023 for testing. This exercise is intended to assess whether the baseline findings depend materially on the inclusion of pandemic-period observations in the estimation sample.

Under the F1-optimized threshold regime, the results remain favorable to the stronger nonlinear classifiers, although the ranking changes across training windows. In the baseline specification, Random Forest performs best, with recall of 0.455, F1 of 0.233, and MCC of 0.085, followed by the ANN with recall of 0.364 and F1 of 0.211. When 2020 is removed from the training window, Random Forest performs better than in the baseline case. Its recall rises to 0.545, and its F1 score increases to 0.255. The ANN improves slightly, with F1 rising to 0.235 while recall stays at 0.364. XGBoost also moves closer to the leading models, with a recall of 0.455 and an F1 of 0.200. Logistic regression also improves, reaching a recall of 0.273 and an F1 score of 0.162.

When both 2020 and 2021 are removed, the ranking becomes less clear. The ANN records the highest F1 score, 0.229, but Random Forest is very close at 0.227. Logistic regression also improves, with an F1 score of 0.211. XGBoost records the highest recall in this specification, 0.636, but its F1 score remains lower at 0.192, and its MCC becomes negative, -0.024. This indicates that the additional downgrade capture is achieved by flagging many more non-downgrade firms.

Table 12 reports the MCC-optimized results for the same COVID-period exclusion checks. In the baseline case, Random Forest again performs best, with a recall of 0.455, an F1 score of 0.244, and an MCC of 0.103. XGBoost follows, with an F1 score of 0.222 and an MCC of 0.078. The ANN matches Random Forest in recall, at 0.455, but its MCC is only 0.002, which means that its downgrade capture is not accompanied by strong overall classification balance.

After excluding 2020, Random Forest improves further, with a recall of 0.545, an F1 score of 0.255, and an MCC of 0.119. The ANN still identifies 45.5% of downgrade cases, but its MCC falls to -0.014, suggesting a weaker balance between correct and incorrect classifications. When both 2020 and 2021 are excluded, Random Forest still has the

highest MCC, 0.076. In this case, logistic regression is closer to Random Forest. It has a recall of 0.364, an F1 of 0.211, and MCC of 0.059. The ANN performs less well under the MCC-optimized rule. It identifies only 9.1% of downgrade cases, with an F1 score of 0.118 and an MCC of 0.039.

The PR-AUC results add another point to consider. In the baseline scenario, logistic regression remains the top-ranked model on test PR-AUC (0.214). When 2020 is excluded, however, the threshold-free ranking shifts in favor of the tree-based models, with Random Forest rising to 0.212 and XGBoost to 0.203, both above the ANN (0.170) and logistic regression (0.165). When both 2020 and 2021 are excluded, PR-AUC falls back across the stronger nonlinear models, with Random Forest at 0.165, ANN at 0.162, and XGBoost at 0.156, while the traditional models cluster narrowly between 0.165 and 0.168.

Table 11. COVID-19 Exclusion Robustness Check (F1-Optimized Threshold).

Scenario	Model	Threshold	Recall	F1	MCC	PR-AUC (Test)
Baseline	ANN	0.209	0.364	0.211	0.059	0.172
Baseline	Logistic	0.187	0.091	0.100	−0.007	0.214
Baseline	Probit	0.192	0.091	0.100	−0.007	0.167
Baseline	Linear	0.173	0.182	0.125	−0.039	0.167
Baseline	Random Forest	0.158	0.455	0.233	0.085	0.173
Baseline	XGBoost	0.146	0.364	0.190	0.024	0.146
Excl. 2020	ANN	0.208	0.364	0.235	0.099	0.170
Excl. 2020	Logistic	0.146	0.273	0.162	−0.006	0.165
Excl. 2020	Probit	0.149	0.182	0.118	−0.056	0.164
Excl. 2020	Linear	0.154	0.182	0.114	−0.064	0.165
Excl. 2020	Random Forest	0.139	0.545	0.255	0.119	0.212
Excl. 2020	XGBoost	0.083	0.455	0.200	0.026	0.203
Excl. 2020–2021	ANN	0.218	0.364	0.229	0.088	0.162
Excl. 2020–2021	Logistic	0.163	0.364	0.211	0.059	0.167
Excl. 2020–2021	Probit	0.168	0.273	0.162	−0.006	0.165
Excl. 2020–2021	Linear	0.178	0.273	0.162	−0.006	0.168
Excl. 2020–2021	Random Forest	0.134	0.455	0.227	0.076	0.165
Excl. 2020–2021	XGBoost	0.033	0.636	0.192	-0.024	0.156

Note: All scenarios use 2022 for validation and 2023 for testing. The baseline model is trained on 2018–2021, "Excl. 2020" on 2018–2019 and 2021, and "Excl. 2020–2021" on 2018–2019 only. Recall, F1, and MCC are evaluated at the F1-optimized threshold, while PR-AUC is reported as a threshold-free measure.

Table 12. COVID-19 Exclusion Robustness Check (MCC-Optimized Threshold).

Scenario	Model	Threshold	Recall	F1	MCC	PR-AUC (Test)
Baseline	ANN	0.183	0.455	0.189	0.002	0.172
Baseline	Logistic	0.191	0.091	0.100	-0.007	0.214
Baseline	Probit	0.194	0.091	0.100	-0.007	0.167
Baseline	Linear	0.173	0.182	0.125	-0.039	0.167
Baseline	Random Forest	0.161	0.455	0.244	0.103	0.173
Baseline	XGBoost	0.160	0.364	0.222	0.078	0.146
Excl. 2020	ANN	0.172	0.455	0.182	-0.014	0.170
Excl. 2020	Logistic	0.209	0.091	0.125	0.060	0.165
Excl. 2020	Probit	0.213	0.091	0.118	0.039	0.164
Excl. 2020	Linear	0.212	0.091	0.125	0.060	0.165
Excl. 2020	Random Forest	0.140	0.545	0.255	0.119	0.212
Excl. 2020	XGBoost	0.085	0.455	0.204	0.034	0.203
Excl. 2020–2021	ANN	0.295	0.091	0.118	0.039	0.162
Excl. 2020–2021	Logistic	0.163	0.364	0.211	0.059	0.167
Excl. 2020–2021	Probit	0.169	0.273	0.162	-0.006	0.165
Excl. 2020–2021	Linear	0.178	0.273	0.162	-0.006	0.168
Excl. 2020–2021	Random Forest	0.134	0.455	0.227	0.076	0.165
Excl. 2020–2021	XGBoost	0.033	0.636	0.192	-0.024	0.156

Note: Table 12 reports the MCC-optimized results for the same training scenarios used in Table 11. All scenarios use 2022 for validation and 2023 for testing. Recall, F1, and MCC are evaluated at the MCC-optimized threshold. PR-AUC is threshold-free and is therefore identical to the corresponding values reported in Table 11.

6. DISCUSSION

6.1. Main Interpretation

The main implication of the results is that model usefulness is task-dependent. Random Forest performs best when the objective is threshold-based downgrade screening, while logistic regression performs best when performance is assessed in threshold-free ranking terms. The ANN remains competitive, especially in downgrade capture, but its relative advantage is less consistent once stronger nonlinear comparators are included.

More broadly, the results show that model rankings are sensitive to the evaluation criterion. In rare-event downgrade prediction, this is not a minor technical detail, because the choice of metric is directly tied to the practical role of the model. Screening and ranking are not the same task. A screening model is evaluated at a particular operating threshold: it matters whether the model can flag firms with elevated downgrade risk. PR-AUC speaks to the ranking task because it evaluates performance across possible thresholds. No model is best for every use. The better choice depends on the purpose of the forecast.

Logistic regression has the highest PR-AUC, 0.214, but it should not be treated as the best model for issuing downgrade alerts. What the PR-AUC result shows is that logistic regression orders firms fairly well by relative downgrade risk. After the validation-selected threshold is imposed, however, it flags only a small number of firms. This explains why logistic regression has low recall in the threshold-based results, even though it records the highest PR-AUC.

This is also why the PR-AUC result should not be read as evidence of good probability calibration. The study does not formally test whether the predicted probabilities are calibrated. In addition, the 2023 test set has only 11 downgrade events, so a calibration plot would be quite sensitive to one or two observations.

The coefficient estimates give an additional economic check on the results. Across the three parametric benchmark models, Δx_3, the change in EBIT to total assets is the only predictor that remains statistically significant. In this sample, the clearest accounting signal of downgrade risk is therefore weakening operating profitability.

6.2. Stability Across Alternative Specifications

The robustness checks generally support the baseline results, but they also make clear that the model rankings should be treated with caution. In the bootstrap analysis, Random Forest has the strongest median F1 and MCC. The intervals, however, are wide and overlap with those of other models. This is not surprising, given the limited number of downgrade events in the test set. A small change in the number of correctly classified downgrade cases can noticeably affect recall, F1, and MCC. The bootstrap results therefore support Random Forest as the strongest baseline model, but not as a model that clearly dominates in a statistical sense.

The rolling-origin exercise gives a similar warning. The ranking changes when the test year changes. In Fold 1, where 2022 is the test year, the ANN has the highest F1 and MCC. XGBoost and Random Forest capture more downgrade cases in that fold, but they do so with many false positives. In Fold 2, the baseline 2023 test split, Random Forest moves to the top under threshold-based evaluation. The PR-AUC results also vary between the two folds. In Fold 1, probit regression and logistic regression are among the stronger models. In Fold 2, logistic regression again records the highest PR-AUC. The change across folds shows that the model ordering is not fully stable from one forecast window to the next.

The winsorization checks point in the same direction. When extreme predictor values are treated more strongly, the ANN performs better under the F1 rule and records the highest F1 and PR-AUC at the 2.5% winsorization level. Random Forest, however, remains stronger under the MCC rule. The two models therefore respond differently to extreme financial-ratio changes. The ANN is more sensitive to outlier treatment, while Random Forest is steadier when balanced classification is emphasized.

The COVID-period exclusion checks also change some model rankings. Removing 2020, or removing both 2020 and 2021, affects the relative performance of several models. Still, the main interpretation is mostly the same. Random

Forest remains the most consistent model for threshold-based screening. The ANN still detects some downgrade cases, but its MCC-optimized results are less stable. Logistic regression improves in some exclusion checks and still performs well under PR-AUC, but it does not become the strongest screening model. Overall, the robustness checks suggest that the findings are not tied to one design choice. They also show that the forecast window, outlier treatment, and evaluation metric can change the results.

6.3. Implications and Limitations

The results have a practical implication for credit monitoring. Random Forest is more suitable when the user wants a downgrade review list. In the 2023 baseline test, it identifies five of the eleven downgrade cases under the F1-optimized threshold, with an F1 score of 0.233. Under the MCC-optimized threshold, it also gives the highest MCC, 0.103. These values are not high enough for automatic decisions, but they can help analysts decide which firms should be reviewed first.

Logistic regression serves a different purpose. Its test PR-AUC is 0.214, the highest among the models. This makes it more useful for ranking firms by relative downgrade risk than for issuing a binary downgrade alert. It also has the advantage that the coefficient signs can be linked to accounting ratios, which is useful when analysts need to explain a risk signal to committees or supervisors.

For financial institutions, the results support a two-step use of the models. Random Forest can help create a review list, while analyst judgment remains necessary after the alert is produced. Logistic regression can be used alongside it to rank firms across the portfolio. For regulators, the same idea applies. The model can help direct attention, but it should not replace rating analysis, supervisory judgment, or other monitoring information.

The evidence is still limited. The bootstrap results show that every MCC confidence interval includes zero, so balanced classification performance remains uncertain. The small number of downgrade events in the test set also means that the reported recall, F1, and MCC values are sensitive to a few classifications. The evidence comes from one market and one domestic rating agency. The predictors are also limited to changes in the revised Altman Z-score determinants. Other variables, such as cash flow, market measures, governance information, or macroeconomic conditions, may lead to different results. Finally, the model comparison does not cover every possible method. The findings should therefore be read as evidence from one Thai emerging-market sample, not as a general ranking of credit-risk classifiers.

6.4. Relation to Prior Work

The results are broadly consistent with studies showing that nonlinear models can be useful when financial relationships are not well captured by simple parametric forms (De Oliveira & Basso, 2025; Lessmann et al., 2015; Lin et al., 2025; Tam & Kiang, 1992). Random Forest and XGBoost support this view because they perform relatively well in downgrade screening. Their advantage is plausible because both models can use nonlinear movements and interactions in the financial-ratio changes.

The results do not mean that machine-learning models always outperform traditional models. The ANN detects downgrade cases reasonably well, but it is not the strongest model after Random Forest and XGBoost are included. This matters because the ANN may look more favorable when it is compared only with logit, probit, and the linear probability model. Once other nonlinear benchmarks are added, its advantage becomes less clear.

The paper therefore adds a narrower point to the literature. In this Thai sample, Random Forest fits threshold-based screening better, while logistic regression fits threshold-free risk ranking better. The preferred model depends on the benchmark set, threshold rule, and evaluation metric.

7. CONCLUSION

This study compares ANN, logistic regression, probit regression, the linear probability model, Random Forest, and XGBoost for annual credit rating downgrade prediction among Thai listed non-financial firms during 2018–2023. The sample contains 558 firm-year observations and 77 downgrade cases. The models are evaluated under a time-ordered train-validation-test design using threshold-based measures and PR-AUC.

The coefficient results show one main accounting signal. The coefficient estimates consistently identify Δx_3, the change in EBIT to total assets as the only statistically significant predictor. In this sample, weakening operating profitability is the clearest accounting-based warning sign of downgrade risk.

The model comparison gives a task-specific conclusion. Random Forest is more useful for flagging downgrade candidates at a selected threshold. Logistic regression is more useful for ranking firms by relative downgrade risk. The ANN still detects some downgrade cases, but its advantage is less stable after Random Forest and XGBoost are included.

Future research could use a larger sample, other markets, and additional predictors. Cash-flow variables, market-based indicators, governance measures, or macroeconomic variables may add information beyond the four Altman-based changes used here. Future work could also examine probability calibration and interpretable machine-learning methods for credit surveillance.

The findings should be read with caution because the test set is small and the evidence comes from one Thai emerging-market credit-rating sample. The results should therefore be interpreted as sample-specific evidence rather than final rankings of the competing classifiers.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Data Availability Statement: The corresponding author can provide the supporting data of this study upon a reasonable request.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Supervision, validation, and manuscript review and editing, Jiroj Buranasiri (J.B.); conceptualization, methodology, data analysis, and preparation of the original draft, Prajya Ngamjan (P.N.); data curation, formal analysis, and manuscript review and editing, Nuttawaree Ratchpiboon (N.R.). Final manuscript review and approval, all authors. All authors have read and agreed to the published version of the manuscript.
Disclosure of AI Use: The authors used OpenAI's ChatGPT to support language editing and manuscript polishing. All content was reviewed and validated by the authors.

REFERENCES

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589-609. https://doi.org/10.2307/2978933

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the Z-score and ZETA® models. In Handbook of research methods and applications in empirical finance. In (pp. 428-456). Cheltenham, UK; Northampton, MA: Edward Elgar Publishing.

Altman, E. I., & Rijken, H. A. (2006). A point-in-time perspective on through-the-cycle ratings. *Financial Analysts Journal*, *62*(1), 54-70. https://doi.org/10.2469/faj.v62.n1.4058

Baum, E. B., & Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, *1*(1), 151–160. https://doi.org/10.1162/neco.1989.1.1.151

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71-111. https://doi.org/10.2307/2490171

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899-2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Claessens, S., & Yurtoglu, B. B. (2013). Corporate governance in emerging markets: A survey. *Emerging Markets Review*, *15*, 1-33. https://doi.org/10.1016/j.ememar.2012.03.002

De Oliveira, N. A., & Basso, L. F. C. (2025). Advancing credit rating prediction: The role of machine learning in corporate credit rating assessment. *Risks*, *13*(6), 116. https://doi.org/10.3390/risks13060116

Dichev, I. D., & Piotroski, J. D. (2001). The long-run stock returns following bond ratings changes. *The Journal of Finance*, *56*(1), 173-203. https://doi.org/10.1111/0022-1082.00322

Fan, J. P. H., Titman, S., & Twite, G. (2012). An international comparison of capital structure and debt maturity choices. *Journal of Financial and Quantitative Analysis*, *47*(1), 23-56. https://doi.org/10.1017/S0022109011000597

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in neural information processing systems. In (Vol. 35, pp. 507–520). Red Hook, NY: Curran Associates, Inc.

Hand, J. R. M., Holthausen, R. W., & Leftwich, R. W. (1992). The effect of bond rating agency announcements on bond and stock prices. *The Journal of Finance*, *47*(2), 733-752. https://doi.org/10.1111/j.1540-6261.1992.tb04407.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W. J., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Kisgen, D. J. (2006). Credit ratings and capital structure. *The Journal of Finance*, *61*(3), 1035-1072. https://doi.org/10.1111/j.1540-6261.2006.00866.x

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

Lin, Y.-C., Padliansyah, R., Lu, Y.-H., & Liu, W.-R. (2025). Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques. *International Journal of Accounting Information Systems*, *56*, 100744. https://doi.org/10.1016/j.accinf.2025.100744

Löffler, G. (2004). An anatomy of rating through the cycle. *Journal of Banking & Finance*, *28*(3), 695-720. https://doi.org/10.1016/S0378-4266(03)00041-4

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, *5*(6), 865-872. https://doi.org/10.1109/72.329683

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

Partnoy, F. (1999). The Siskel and Ebert of financial markets?: Two thumbs down for the credit rating agencies. *Washington University Law Quarterly, 77*(3), 619–712.

Prechelt, L. (1998). Early stopping - but when? In G. B. Orr & K.-R. Müller (Eds.), Neural networks: Tricks of the trade (Lecture Notes in Computer Science). In (Vol. 1524, pp. 55–69). Berlin, Germany: Springer.

Qian, H., Wang, B., Ma, P., Peng, L., Gao, S., & Song, Y. (2022). *Managing dataset shift by adversarial validation for credit scoring.* Paper presented at the Pacific rim International Conference on Artificial Intelligence, Springer.

Radovanovic, J., & Haas, C. (2023). The evaluation of bankruptcy prediction models based on socio-economic costs. *Expert Systems with Applications, 227*, 120275. https://doi.org/10.1016/j.eswa.2023.120275

Rahman, M. J., & Zhu, H. (2024). Predicting financial distress using machine learning approaches: Evidence China. *Journal of Contemporary Accounting & Economics, 20*(1), 100403. https://doi.org/10.1016/j.jcae.2024.100403

Rajan, R. G., & Zingales, L. (1995). What do we know about capital structure? Some evidence from international data. *The Journal of Finance, 50*(5), 1421-1460. https://doi.org/10.1111/j.1540-6261.1995.tb05184.x

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536. https://doi.org/10.1038/323533a0

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shin, K.-S., & Han, I. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications, 16*(2), 85-95. https://doi.org/10.1016/S0957-4174(98)00063-3

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business, 74*(1), 101-124. https://doi.org/10.1086/209665

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38*(7), 926-947. https://doi.org/10.1287/mnsc.38.7.926

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). *Credit scoring and its applications* (2nd ed.). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Treacy, W. F., & Carey, M. (2000). Credit risk rating systems at large US banks. *Journal of Banking & Finance, 24*(1-2), 167-201. https://doi.org/10.1016/S0378-4266(99)00056-4

TRIS Rating Co Ltd. (2024). *Company overview.* Retrieved from https://www.trisrating.com/company-overview

White, L. J. (2010). Markets: The credit rating agencies. *Journal of Economic Perspectives, 24*(2), 211-226. https://doi.org/10.1257/jep.24.2.211

Wu, Y., Gaunt, C., & Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics, 6*(1), 34-45. https://doi.org/10.1016/j.jcae.2010.04.002

Yunisa, R., & Santi, F. (2023). Analysis financial distress potential in ASEAN industrial companies using Altman Z-score and Springate methods. *East Asian Journal of Multidisciplinary Research, 2*(12), 4993-5008. https://doi.org/10.55927/eajmr.v2i12.6902

Zoričák, M., Gnip, P., Drotár, P., & Gazda, V. (2020). Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. *Economic Modelling, 84*, 165-176. https://doi.org/10.1016/j.econmod.2019.04.003