2 (2026) 100186

Contents lists available at ScienceDirect

Finance Research Open

journal homepage: www.sciencedirect.com/journal/finance-research-open

A systematic literature review of bank failure prediction models

Japheth Jev [a,b], Lukman A. Akanbi [b,*], Muhammad Bilal [b], Arun Sukumar [b], Simeon Babatunde [b], Hakeem Owolabi [c], Vasileios Kolovos [b]

[a] Accounting and Finance Programme, Birmingham Newman University, Bartley Green, Birmingham, B32 3NT, United Kingdom
[b] Business School, Birmingham City University, The Curzon Building, 4 Cardigan Street, Birmingham, B4 7BD, United Kingdom
[c] Greater Manchester Business School, University of Greater Manchester, Bolton, BL3 5AB, United Kingdom

ARTICLE INFO

Keywords:
Banking
Bank failures
Artificial intelligence
Banking crisis
Bank regulation
Ensemble AI
Machine learning
Early Warning System

ABSTRACT

Following the Global Financial Crisis of 2008–2009, bank-failure prediction and Early Warning Systems (EWS) have received renewed attention from researchers and banking regulators seeking to identify vulnerabilities in financial institutions before failure occurs. However, the literature employs heterogeneous definitions of bank failure, predictor variables, modelling approaches, prediction horizons, validation designs and performance measures, making comparison and interpretation of reported predictive performance challenging.

This systematic literature review identifies, evaluates and synthesises a focused body of post-Global Financial Crisis research on bank-failure prediction models, with particular attention to the methodological characteristics, validation practices and comparative predictive performance of conventional econometric and artificial intelligence/machine learning (AI/ML) approaches. Following a PRISMA-informed study-selection process, a sequentially refined search of the Scopus database identified 23 publications meeting the review criteria, comprising 21 primary empirical studies and two review studies. The narrative synthesis examined model type, operational definition of bank failure, data and predictor variables, prediction horizon, validation design, performance metrics and comparative predictive performance.

The findings reveal substantial methodological heterogeneity across the empirical literature. Bank-level financial and accounting variables remain prominent predictors, while broader macroeconomic and non-financial risk factors are less consistently represented. Several studies reported competitive or higher predictive performance for AI/ML approaches relative to conventional econometric models; however, the magnitude and consistency of these differences varied across datasets, failure definitions, prediction horizons, class distributions, evaluation metrics and validation designs. Greater interpretive weight was therefore placed on within-study comparisons conducted under broadly comparable empirical conditions. Overall, the evidence indicates that AI/ML methods can provide incremental predictive value in particular settings but does not establish their universal superiority over conventional econometric approaches. The review further identifies limitations in the treatment of class imbalance, model validation, performance evaluation and model interpretability. These findings support the need for more transparent and methodologically comparable evaluation frameworks and provide directions for future research on robust and regulator-relevant bank-failure Early Warning Systems.

1. Introduction

There has been a renewed interest in the prediction of bank failures since after the Global Financial Crisis of 2008–2009, which resulted in the failures of big global banks such as Lehman Brothers and losses of over $6 trillion in global banking assets (Citterio, 2024; IMF Research Dept., 2010). Banking is a highly regulated business given the pivotal roles that banks play in an economy as custodians of wealth and financial intermediaries that mobilise savings from surplus saving units to deficit savings units while channelling credits to the most productive sectors of the economy to propel economic growth (Tarawneh et al., 2024). To effectively regulate banks, monetary authorities, and central banks adopt forward-looking regulatory frameworks and models that are designed to serve as Early Warning Systems (EWS) capable of providing a future view into any future bank failures. These regulatory frameworks focus on capital adequacy requirements, liquidity requirements, stress testing, macroprudential evidence, and supervisory

* Correspondence to: Department of Management, Business and Marketing, Birmingham City University, 4 Cardigan Street, Birmingham, B4 7BD, United Kingdom.
E-mail address: lukman.akanbi@bcu.ac.uk (L.A. Akanbi).

https://doi.org/10.1016/j.finr.2026.100186
Received 2 April 2026; Received in revised form 8 October 2026; Accepted 13 October 2026
Available online 19 October 2026
3050-7006/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

reviews as well as the capital adequacy, asset quality, management, earnings, liquidity, and systemic risk (CAMELS) rating system which has been utilised by central banks as a principal tool in assessing banks' risk profile and potential failures (Acharya & Mora, 2015; Aikman et al., 2018; Board, 2010).

For the purpose of this review, bank failure refers primarily to the cessation of a bank's independent operation through regulatory closure, licence withdrawal, liquidation, receivership, or transfer to a resolution or deposit-insurance authority. This is distinguished from financial distress, which refers to substantial deterioration in a bank's financial condition without necessarily resulting in closure, and from bailout or regulatory intervention, where an institution may continue operating following external support. However, because the operational definition of bank failure varies across the literature, studies employed clearly specified proxies for severe financial distress were retained where these were explicitly used to develop or evaluate bank-failure prediction models Federal Deposit Insurance Corporation (2025) and Citterio (2024).

Several studies have attempted to predict bank failures as well as develop Early Warning Systems (EWS) to predict them using a wide spectrum of models and tools. Some of these studies have adopted traditional economic models while others have utilised artificial intelligence and machine learning models to predict bank failures (Barajas et al., 2023; Behbood et al., 2015; Jing & Fang, 2018; Liu et al., 2021). The foundational work of Altman (1968) on corporate bankruptcy prediction paved the way for the rise of research interest in bankruptcy predictions and the adoption of economic models in developing EWS. Although the work was not based on a sample dataset of banking firms, the Z-score developed by Altman has been widely applied by studies that focused on predicting bank failures. Similar to the work of Altman, Cox and Wang (2014) adopted discriminant analysis to investigate bank failures in the United States from 2007 to 2010 and found that the main predictors of bank failures were illiquid loans and exposure to the inter-bank funding market. They, therefore, concluded that these predictors can be used as early warning signals for predicting bank failures. Another connected work that followed the approach of Altman was the work of Kumar and Sharma (2022) who adopted discriminant analysis to investigate the financial health and soundness of banks in India using financial ratios such as capital adequacy, asset quality, management efficiency, earnings, and liquidity. They posited that capital adequacy ratios, non-performing loan ratios, and return on equity acted as significant predictors for predicting bank failures. Furthermore, Li et al. (2022) employed discriminant analysis to predict bank failures in China. The work presented an improved version of the earlier works as the authors incorporated not only bank-specific financial variables but also macroeconomic variables such as interest rates, exchange rates, and GDP. The authors identified key variables such as liquidity ratios, profitability ratios, and macroeconomic indicators as the most significant predictors of bank failures. They also argued that the integration of macroeconomic variables in the model significantly improved the model's predictive accuracy. This result pointed to the flaws and weaknesses in other bankruptcy prediction models including Altman's Z-score which utilised only firm-specific financial data without incorporating macroeconomic variables in making bankruptcy predictions.

Other strands of studies predicting bank failures that fall within the purview of traditional economic models have adopted logistic regression and Hazard models to predict bank failures. Demirgüç-Kunt and Detragiache (1998) utilised a logistic regression model to investigate the causes of bank failures in developing countries. They argued that banking crises in developing countries were precipitated mainly by macroeconomic variables such as GDP, inflation, and real interest rates but not bank-specific financial ratios as claimed by other related studies. Kolari et al. (2002) undertook another study that applied logistic regression to predict bank failures in a developed country—the United States. Although they adopted a similar methodology, their result contrasted that of Demirgüç-Kunt and Detragiache (1998) as they concluded that bank-specific variables such as capital adequacy, asset quality, and earnings had significant influence in predicting bank failures in the United States.

Lane et al. (1986) adopted the proportional Hazard model or survival analysis, which attempted to estimate the period that exists before an event occurs to predict the timing of bank failures. They found that the timing of bank failures was predicted with a high degree of accuracy by bank-specific variables relating to asset quality and earnings. Betz (2014) conducted a related study that used the proportional Hazard model to predict bank failures in the economic zone of the European Union during the 2008 banking crisis. The researchers incorporated both bank-specific financial variables and macroeconomic variables in their model. They argued that the combination of the two variables made their model more robust and effective in predicting the timing of bank failures.

Recent advances in technology and breakthroughs in artificial intelligence have paved the way for the application of machine learning techniques to predict bank failures. Machine learning models have huge advantages over traditional economic models for predicting bank failures. First, machine learning models are non-parametric in nature, which implies that they do not assume the normal distribution of the dataset and can model complex non-linear relationships that are inherent in the financial data (Breiman, 2001). Second, machine learning models like recurrent neural networks, support vector machines, and ensemble models have proven to achieve predictive performance accuracy that was higher than that attained by traditional economic models due to their ability to capture complex interactions and relationships among variables (Xia et al., 2025). Third, machine learning models can process large volumes of data from diverse sources such as financial statements, financial time series, macroeconomic variables, and even unstructured data like social media posts, thus capturing market sentiments and enhancing the model's capability to undertake a more comprehensive analysis of variables affecting a bank's risk profile (Gogas et al., 2018).

Considering the advantages of machine learning models, recent empirical studies on bank failure predictions have increasingly adopted them as the main methodology to predict bank failures. The machine learning algorithms that are commonly used by bank failure prediction studies include support vector machines (SVM), decision trees and random forests, neural networks, Gradient Boosting Machines (GBM), and ensemble methods (Citterio, 2024; Kristóf & Virág, 2022; Li et al., 2022; Pham & Ho, 2021; Xia et al., 2025).

Iturriaga and Sanz (2015) adopted an artificial neural network to develop a model that focused on the specific features of the financial crisis to predict the bankruptcy of US commercial banks between 2000 and 2012. They argued that their model had 96% predictive accuracy and outperformed traditional economic models for predicting bank failures, and thus could be used to predict bank failures in the short, medium, and long term. Boyacioglu et al. (2009) carried out an expanded study that tested the application of various machine learning algorithms in predicting bank failure in the U.S. where they adopted four different categories of neural networks such as multi-layer perceptron, competitive learning, self-organising map and learning vector quantisation. They found that multi-layer perceptron and learning vector quantisation were more accurate in predicting future bank failures. Gogas et al. (2018) utilised an ensemble learning method that incorporated several machine learning algorithms to capture non-linear interactions in the dataset to predict bank failures. They found that the ensemble method effectively captured the nonlinearity in the data and achieved higher predictive accuracy than a single machine learning algorithm could achieve. Its result was also more accurate than that of traditional economic models. Similarly, Kou et al. (2019) adopted gradient boosting machines (GBM) to predict financial distress in banks using both financial ratios and macroeconomic variables as sample data. The study concluded that the GBM handled complex non-linear

relationships within the data and showed the highest predictive accuracy score when compared with individual machine learning models and traditional economic models.

While the attention of researchers has been increasingly drawn to the application of artificial intelligence (AI) and machine learning (ML) in predicting bank failures in recent times, the nature, direction, and scope of such applications are broad and heterogeneous. Also, the types of AI/ML models adopted are varied and complex, thus making it hard for people interested in the subject who are not subject matter experts in artificial intelligence to appreciate the usefulness of AI technology in the prediction of bankruptcy in financial firms like banks and improving banking regulation. The generally broad nature of the available literature in this domain spans from differences in the banking systems across countries in which the studies were situated to differences in research questions used by researchers, thus making it difficult to identify aspects of the research that require further investigations. Or even making it more difficult to conclude on whether or not artificial intelligence as a technology is effective in boosting banking supervision and strengthening the banking system as adduced (Bahoo et al., 2024).

Although some researchers have attempted to undertake systematic literature reviews that aggregated and analysed different studies that have investigated the applications of artificial intelligence (AI) models in bankruptcy predictions, most of their reviews were focused on studies predicting corporate bankruptcies of companies other than banks with little or no coverage of those applying AI models to predict bankruptcy of financial firms like banks. Common systematic review studies following this trajectory are the works of Clement et al. (2022), El Madou et al. (2024), Jones (2023), Kuizinienė et al. (2022), Quinteros-Navarro and Rodriguez (2024), Shakeel et al. (2023).

There are only a few strands of prominent systematic literature reviews that have reviewed studies that investigated the application of artificial intelligence (AI) to bank bankruptcy predictions, i.e. bank failures, and the prominent ones include the works of Citterio (2024) and Liu et al. (2021). Other available systematic reviews relating to the application of AI in banking such as Bahoo et al. (2024), Dewasiri et al. (2024), and Satheesh and Nagaraj (2021) covered studies that applied AI in other banking operations such as risk management, customer service, and general banking operations without precisely synthesising studies that predicted bank failures.

The salient dimensions of the two prominent review studies, Citterio (2024) and Liu et al. (2021) which focused on aggregating and synthesising studies modelling bank failure and distress risk are comparatively presented in Table 1 below, alongside those in our review to demonstrate the incremental contribution of the current review.

Table 1 positions the present review relative to the two previous reviews most closely aligned with its scope. Liu et al. (2021) provide a broad synthesis of statistical, artificial intelligence and machine-learning approaches to bank-failure prediction, while Citterio (2024) provides a more extensive review of bank-default prediction research published between 2000 and 2022, with particular attention to definitions of financial distress, prediction techniques and predictor selection.

Accordingly, the incremental contribution of the present review does not rest on providing broader or exhaustive coverage of the literature. Rather, it adopts a deliberately focused post-Global Financial Crisis scope and extends previous synthesis through a structured critical assessment of methodological characteristics that affect the credibility and comparability of reported predictive performance. These include failure definition, sample and failure-event adequacy, class imbalance, validation design, and out-of-sample or temporal testing. The review further distinguishes predictive-performance measures from model-fit statistics and validation procedures and places greater interpretive weight on comparative evidence generated under broadly comparable empirical conditions. This provides a methodological basis for assessing when reported performance differences between econometric and AI/ML approaches constitute credible evidence of incremental predictive value, rather than treating headline accuracy differences as directly comparable across heterogeneous studies.

This systematic review aimed to systematically identify, evaluate and synthesise a focused body of post-Global Financial Crisis literature on bank-failure prediction models, with the objective of identifying methodological and empirical gaps and assessing their implications for future research, banking supervision and financial stability. Against this background, the present study undertakes a focused systematic review of studies applying conventional econometric and AI/ML approaches to bank-failure prediction. Rather than claiming an exhaustive mapping of the broader banking-distress literature, the review adopts a deliberately bounded scope centred on bank-failure prediction in the post-Global Financial Crisis period. The review comparatively examines the modelling approaches adopted, data and predictor variables, operational definitions of bank failure, prediction horizons, validation designs, performance metrics and reported predictive performance. Econometric models are retained as an important benchmark for assessing developments in AI/ML-based predictive modelling. Accordingly, the review does not assume the superiority of either methodological approach but critically evaluates the conditions under which reported performance differences can reasonably be interpreted.

The overarching research question of this systematic review is: What does the post-Global Financial Crisis literature reveal about the methodological characteristics, validation and predictive performance of econometric and AI/ML models for bank-failure prediction? Accordingly, the objective of the review is to synthesise and critically evaluate these modelling approaches, with particular attention to how bank failure is operationalised, the prediction horizons employed, the validation designs and performance metrics used, and the conditions under which AI/ML models demonstrate incremental predictive performance relative to conventional econometric approaches. These dimensions provide the analytical framework for interpreting the evidence while recognising the substantial methodological heterogeneity across the included studies.

We have operationalised our overarching research question through five sub-questions: (1) How is bank failure defined and operationalised across the included studies? (2) How are prediction horizons specified and reported across bank-failure prediction studies, and what implications does variation in prediction horizon have for comparison of model performance? (3) What validation designs are used to assess model performance and generalisability? (4) Which performance metrics are used, particularly in the context of rare failure events and class imbalance? (5) Under what methodological and empirical conditions do AI/ML models demonstrate incremental predictive performance relative to conventional econometric approaches?

This study made two major contributions to bank failure risk modelling research and banking regulation. Firstly, the findings of this study contributes to the literature by identifying gaps in the existing knowledge which would guide researchers within the bank failure prediction domain and spur future research interest in addressing these gaps, therefore, shaping the future direction of research to be undertaken by other researchers who are interested in investigating bank failures and their causes. Secondly, the findings of this study have practical implications for bank regulators and risk managers in financial institutions who wish to develop early warning systems (EWS) to proactively identify potentially unsound banks and implement intervention measures to prevent their failures, by creating empirical evidence that bank regulators and managers can adopt as a basis for selecting appropriate blend of strong models, relevant predictor variables, and integrating non-financial variables and macroeconomic variables to boost the overall performance and reliability of their EWS.

The rest of the paper is organised as follows. Section 2 focused on the research method and materials adopted to select the eligible studies included in the review as well as explained both the inclusion and exclusion criteria used to select the target studies. Section 3 covered the presentation and discussion of the results of the analysis of the relevant studies that were included in the systematic review. Section 4 presented the limitations of this review, and Section 5 presents the summary and conclusion of the paper.

Table 1
Positioning of the present review relative to previous reviews of bank-failure prediction research (Source: Produced by the Authors).

Dimension	Liu et al. (2021)	Citterio (2024)	Present review
Primary purpose	Synthesis of statistical, AI and ML approaches to bank-failure prediction and directions for future research.	Updated review of bank-default prediction models, failure/distress definitions and predictor variables.	Focused comparative and critical synthesis of econometric and AI/ML approaches to bank-failure prediction.
Temporal scope	Broad historical coverage; reviews more than 60 key papers.	2000–2022.	2010–2024, with an explicit post-Global Financial Crisis focus.
Database and search strategy	A reproducible database search strategy is not clearly specified in the methodological description.	Scopus; broad keyword strategy covering failure, distress, default, insolvency, early-warning systems, prediction and banking terms.	Scopus; explicit sequential TITLE-ABS-KEY search strategy and predefined eligibility restrictions.
Review sample	More than 60 key papers.	76 papers.	23 publications, comprising 21 empirical studies and 2 review studies.
Failure/distress definition	Considered within the reviewed literature but not structured as a methodological-quality domain.	A major analytical dimension of the review.	Explicitly extracted and incorporated into the methodological-quality assessment.
Model families	Statistical, artificial-intelligence and machine-learning approaches.	Statistical and intelligent prediction techniques.	Econometric and AI/ML approaches retained within a common comparative framework.
Predictor variables	Predictive indicators considered alongside modelling methodologies.	CAMELS, macroeconomic, non-financial and ESG-related variables.	Financial, macroeconomic and non-financial predictors examined alongside modelling approaches.
Performance evaluation	Predictive performance and findings discussed across the reviewed studies.	Overall predictive accuracy, Type I and Type II errors, and AUC considered in model comparisons.	Multiple predictive-performance measures assessed while distinguishing performance metrics from model-fit statistics and validation procedures.
Validation design	Not treated as a dedicated methodological-quality appraisal domain.	Validation methods are recorded and considered in model comparisons.	Explicit methodological-quality domain, distinguishing internal validation, holdout testing, temporal testing and out-of-sample evidence.
Class imbalance	Not identified as a formal methodological-quality appraisal domain.	Not identified as a formal methodological-quality appraisal domain.	Explicitly assessed as a methodological-quality domain because bank failure is a rare-event classification problem.
Sample/event adequacy	Study sample characteristics are reported.	Sample sizes and their variation across studies are analysed.	Explicit methodological-quality domain considering sample and failure-event adequacy in relation to model complexity.
Methodological quality appraisal	No structured methodological-quality appraisal framework identified.	No structured methodological-quality appraisal framework identified.	Structured appraisal across sample/event adequacy, failure definition, validation, class imbalance and out-of-sample/temporal testing.
Interpretation of comparative performance	Broad synthesis of developments in statistical and machine-learning prediction approaches.	Broad comparison of models, definitions and predictors; no single prediction technique is established as universally superior.	Emphasises conditional interpretation of model performance, methodological credibility and within-study comparisons rather than cross-study ranking by headline accuracy.

2. Methods and materials

We have adopted an aggregative synthesis approach based on the preferred reporting items for systematic reviews and meta-analyses (PRISMA) protocol as the primary methodology of this study following the works of Bougault et al. (2025) and Alkhuzaimi et al. (2025). The adoption of this method was informed by the need to produce a rigorous evidence-based review that was reproducible and verifiable by other researchers as well as minimising the risks of bias in the selection of studies, data extraction, and result interpretation (O'Dea et al., 2021; Sarkis-Onofre et al., 2021).

Scopus index journal database was searched following the PRISMA protocol framework guidelines to define appropriate inclusion and exclusion criteria that ensured that relevant studies were objectively selected for the systematic review. Scopus provides broad multidisciplinary coverage relevant to this intersection and supports structured and reproducible searching of titles, abstracts and keywords using Boolean operators and other search functions appropriate for systematic evidence retrieval (Gusenbauer, 2022). In addition, sources indexed in Scopus are subject to formal selection and continuing evaluation procedures, providing a consistently curated evidence base for the review. The use of a single database also enabled the search strategy, eligibility restrictions and sequential query refinements to be applied consistently across the identified records. However, the decision to restrict the search to Scopus may have excluded relevant studies indexed exclusively in other bibliographic databases. The resulting evidence base should therefore be interpreted as a focused, reproducibly identified body of Scopus-indexed literature rather than an exhaustive representation of all research on bank-failure prediction.Gusenbauer (2022, 2024).

The literature search terms were constructed from the main research question, which was to conduct a systematic literature review of post Global Financial Crisis (GFC) 2008/2009 studies that adopted conventional econometrics models and AI/ML models for the prediction of bank failures.

The search strategy on the Scopus database was initiated using our curated phrasing technique that combined key terms from the research

question and Boolean operators such as 'AND' and 'OR' to broaden the scope of the search and make it specific to capture only relevant results and documents (Booth et al., 2021; MacFarlane et al., 2022). Consequently, the phrase (Bank failure predictions) was initially used as the first search term. This search was set to search within the title, abstract, and keywords (TITLE-ABS-KEY) of the database and identified 593 research documents in the Scopus Index database that were relevant to the search terms. The following are the search constructs used to extract the relevant articles.

TITLE-ABS-KEY (bank AND failure AND predictions)

The search terms mentioned above were refined in a second filter by putting the first two terms in a double quotation mark while leaving predictions as a stand-alone term, and further queries were performed on the Scopus Index database. So, the key search term changed to this.

TITLE-ABS-KEY ("Bank failure" predictions)

The effect of this was significant as it reduced the number of relevant documents to 151 studies.

A third search was conducted on the database that wrapped all three key terms in the search inside a double quotation as shown in the following:

TITLE-ABS-KEY ("Bank failure predictions")

This search prompted the database to consider the three key terms as one and return relevant studies that closely reflected them only. The effect of this search was a reduction in the number of documents from 151 studies to 58 relevant studies.

The fourth search was conducted, which introduced a new key term—Models into the search construct as shown in the following.

TITLE-ABS-KEY ("Bank failure predictions" models)

This search was carried out to zero in our focus on studies that have adopted both traditional econometric models and artificial intelligence models to predict bank failures and ensure that they were included in the systematic review of the literature. After conducting this search, the number of studies identified was reduced to 48 studies.

The search strategy is further refined based on the year of publication of the relevant studies we have identified using the search construct below.

TITLE-ABS-KEY ("Bank failure predictions" models AND PUBYEAR > 2009 AND PUBYEAR < 2025)

The above search was conducted to ensure that only current relevant studies that have been published between the years 2010 and 2024 were selected for our literature review. Thus, older studies predicting bank failures published before 2010 were excluded from this literature survey. This search further reduced the number of studies that were identified to 28.

Finally, we performed the search that defined the type of documents we wanted to include in our study using the search construct below.

TITLE-ABS-KEY ("Bank failure predictions" models AND PUBYEAR > 2009 AND PUBYEAR < 2025 AND LIMIT-TO (DOCTYPE, "ar") AND LIMIT-TO (SRCTYPE, "j") AND LIMIT-TO (LANGUAGE, "English"))

This search removed documents such as conference papers from the identified documents and ensured that only peer-reviewed papers published in quality journals were included in the systematic literature review. It also eliminated research items that were written in other languages and ensured that only studies written in English language were selected in this systematic literature review. This search resulted in the identification of 23 studies which was the final number of studies selected for the systematic literature review.

We adopted the above rigorous literature search procedures and applied the inclusion and exclusion criteria set out as part of the overall search strategy for this review to ensure that this systematic review of the literature included only quality studies that had been subjected to methodological rigour, as well as facilitate the achievement of the main objective of this study, which was to synthesise available empirical evidence that could be used to generalise the findings of studies that adopted both traditional econometric and AI models to predict failures among banks and inform evidence-based decision making by banks and bank regulators regarding the adoption of these models to predict the probability of future bank failures (Aromataris & Pearson, 2014).

2.1. Inclusion and exclusion criteria

Variations in the operational definition of bank failure—including regulatory closure, licence withdrawal, liquidation and clearly specified severe-distress proxies—were extracted as study characteristics and considered in the methodological quality assessment and interpretation of model performance. We have considered these differences in operational definition of bank failure to establish the inclusion and exclusion criteria for this review, which are presented in Table 2 below. These criteria were adopted to select the final 23 studies that were included in the systematic review.

2.2. The PRISMA flowchart diagram

PRISMA flowchart was utilised to visually demonstrate the document selection process that the researchers implemented. This flowchart helped the researcher to minimise biases that are common in study selection in systematic reviews and improve the transparency, reproducibility, and methodological rigour of the study (Haddaway et al., 2022). See Fig. 1 for the PRISMA flowchart for this study.

2.3. Study identification and selection procedure

The Scopus search strategy was developed through sequential query refinement. The initial broad search identified 593 records, after which progressively more specific search expressions returned 151, 58 and 48 records. Restriction to the predefined 2010–2024 publication period returned 28 records, while application of the final document-type, source-type and language eligibility restrictions produced 23 publications. These numerical reductions represent changes in the number of records retrieved following successive modifications of the Scopus search specification and should therefore not be interpreted as manual title/abstract exclusions, duplicate removal or unsuccessful retrieval. The complete search-refinement sequence is retained in Fig. 1 to provide transparency regarding the development and application of the final search strategy. The 23 publications returned by the final specification constituted the evidence set subjected to the eligibility and inclusion procedures reported in the PRISMA flow.

2.4. Methods of literature evidence analysis

We have adopted a qualitative method to analyse the evidence obtained from the included 23 studies. The primary evidence analysis method adopted in this review is narrative synthesis.

2.4.1. Narrative synthesis

According to Lisy and Porritt (2016), narrative synthesis is a method of evidence synthesis that involves the summarising and presenting of findings of other related studies. It has been increasingly used for the synthesis of evidence recently because it does not only describe and summarise the findings of related studies but also facilitate the examination of the similarities and differences among the studies as well as the investigation of the nature and cause of the relationship within the data and the credibility of the literature evidence being considered.

Based on the above strengths of narrative synthesis, we have adopted it to summarise and evaluate the key findings of the 23 studies that were finally included in this systematic literature review. We employed narrative synthesis to investigate the relationships across the findings of the various studies and critically explored the strength and credibility of the various models adopted by this group of selected studies in predicting bank failures.

Table 2
Inclusion and exclusion criteria applied to the Scopus search.

Inclusion Criteria		
SN	**Criteria**	**Remark**
1	Financial institutions – Banks	Studies focusing on banks and the prediction of bank failure were eligible for inclusion.
2	Econometric and Artificial Intelligence/Machine Learning models	Studies developing, applying, evaluating, or comparing conventional econometric and/or AI/ML models for bank-failure prediction were eligible for inclusion.
3	Publication year	Only studies published between 2010 and 2024 were included, consistent with the review's focus on the post-Global Financial Crisis period.
4	Peer-reviewed journal articles	Only peer-reviewed research articles published in academic journals were included.
5	Publication language – English	Only research articles published in the English language were included.
Exclusion Criteria		
SN	**Criteria**	**Remark**
1	Non-bank financial institutions and other organisations	Studies predicting the failure or financial distress of organisations other than banks were excluded.
2	Studies outside the specified modelling scope	Studies that did not develop, apply, evaluate, or compare econometric and/or AI/ML models for bank-failure prediction were excluded.
3	Publication year	Studies published before 2010 or after 2024 were excluded to maintain the review's focus on the post-Global Financial Crisis period.
4	Publication type	Conference papers, conference proceedings, book chapters, editorials, and other non-journal publications were excluded.
5	Publication language – Non-English	Studies published in languages other than English were excluded.

2.4.2. Quality assessment criteria

We established a structured methodological quality assessment framework to evaluate the 21 primary empirical studies included in this systematic review. The two review studies, Liu et al. (2021) and Citterio (2024), were retained for contextual and methodological synthesis but were not subjected to the prediction-model quality assessment because they did not constitute primary model-development or validation studies.

The quality assessment framework was informed by the Prediction Model Risk of Bias Assessment Tool (PROBAST), which evaluates prediction-model studies across the domains of participants, predictors, outcome and analysis (Wolff et al. 2019). PROBAST provides a structured basis for assessing potential methodological limitations and risk of bias in studies developing or validating prediction models. Because the present review concerns bank-failure prediction and includes heterogeneous econometric and AI/ML modelling approaches, the PROBAST principles were adapted to the specific methodological characteristics of this literature rather than applied as a formal verbatim PROBAST assessment.

Accordingly, the 21 empirical studies were assessed across five domains considered particularly relevant to bank-failure prediction: (1) sample and failure-event adequacy, (2) clarity and appropriateness of the operational definition of bank failure, (3) validation method, (4) treatment of class imbalance, and (5) out-of-sample or temporal testing. The criteria used to operationalise these domains are presented in Table 3.

2.4.3. Methodological quality assessment of the included empirical studies

Based on the criteria described above, the 21 primary empirical studies were subjected to methodological quality assessment across five domains: sample and failure-event adequacy, failure definition, validation method, treatment of class imbalance, and out-of-sample/temporal testing. The two review studies—Liu et al. (2021) and Citterio (2024)—were retained within the overall systematic review to inform the contextual synthesis of methodological developments and research trends but were not subjected to the empirical prediction-model quality assessment. They are therefore reported as not applicable (N/A) in Table 4.

It is important to note that of the 23 included publications, 21 were empirical studies and two—Liu et al. (2021) and Citterio (2024)—were review studies. The two review articles were retained to inform the broader synthesis of methodological developments and research trends but were not treated as primary model-development studies in analyses of sample characteristics, predictive-performance metrics, or methodological quality. Where quantitative frequencies relate specifically to empirical model-development studies, the denominator is stated explicitly. The result of this assessment is presented in Table 3 below.

3. Result and discussion

The results are synthesised through the existing thematic structure of the review. Collectively, the sections address the operational research questions concerning model characteristics and failure definitions, data and predictor selection, prediction and evaluation design, performance metrics, validation practices, and comparative model performance. Given the substantial methodological heterogeneity of the included studies, these dimensions are integrated within the thematic synthesis rather than treated as independent quantitative comparisons

3.1. Classification of included studies By primary modelling approach and study type

We have classified the included studies based on the dominant modelling approach adopted and whether or not they are empirical studies or review studies. This classification represents the primary methodological orientation assigned to each empirical study for descriptive synthesis and does not imply that a study evaluated only one model family. Some studies compared models across methodological

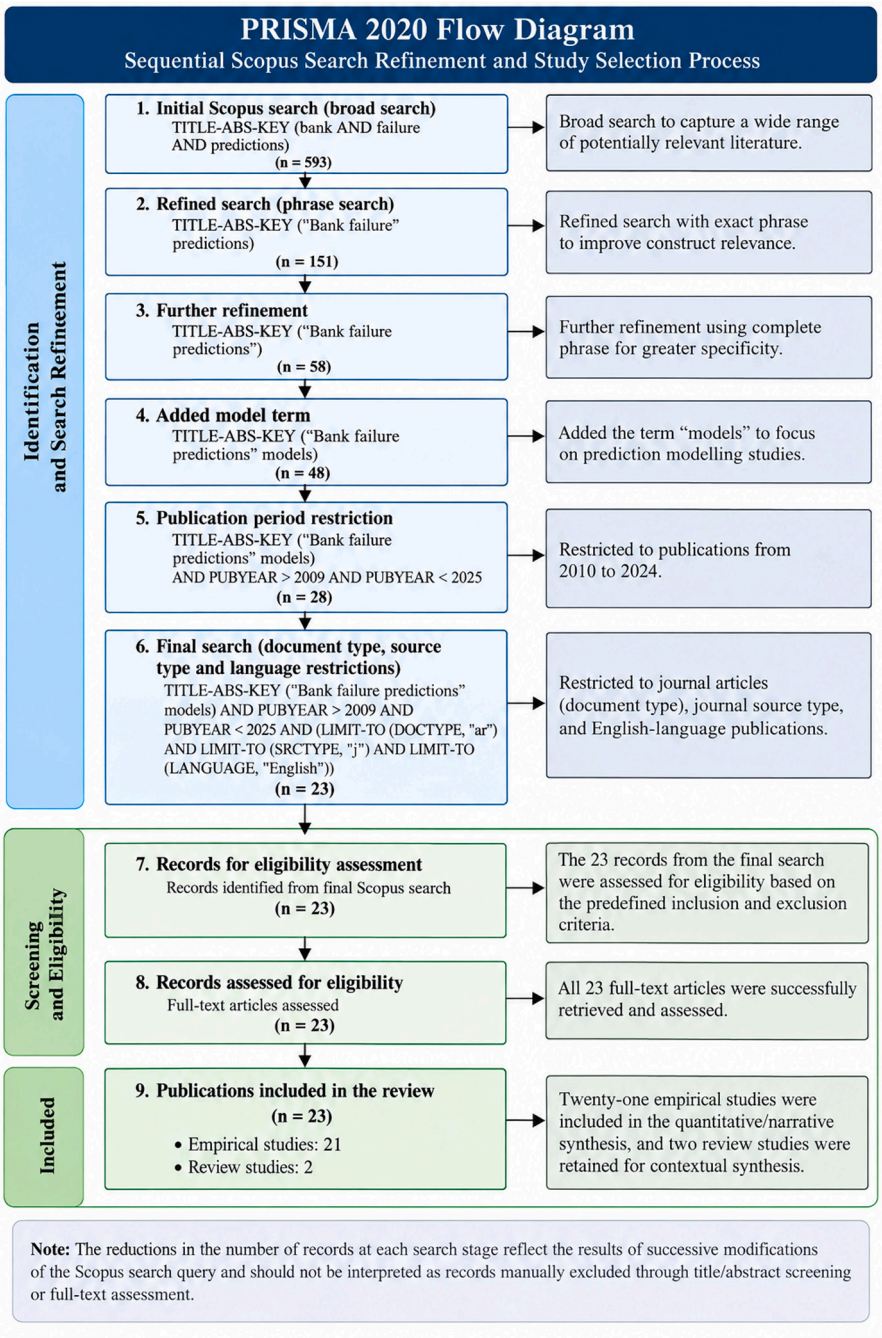


Fig. 1. PRISMA 2020 Flow Diagram of the Sequential Scopus Search Refinement and Study Selection Process.

categories. The Table 5 below presents the distribution of included studies based on these methodologies.

The methodological classification presented in Table 5 is based on the primary modelling orientation of the 21 empirical studies included in the review. The two review studies, Liu et al. (2021) and Citterio (2024), were excluded from this empirical classification because they did not develop or validate primary bank-failure prediction models (Citterio, 2024; Liu et al., 2021). Of the 21 empirical studies, 9 (42.9%) were classified primarily within the traditional econometric modelling category, while 12 (57.1%) were classified primarily within the artificial intelligence/machine learning (AI/ML) category. This classification reflects the primary methodological orientation assigned to each study for descriptive synthesis and does not imply that individual studies evaluated only one model family, since some studies compared econometric and AI/ML approaches within the same empirical design.

The distribution indicates substantial representation of both econometric and AI/ML approaches in the post-Global Financial Crisis bank-failure prediction literature reviewed. Econometric models remain important because of their established application and interpretability, while AI/ML methods have increasingly been applied to capture complex and potentially nonlinear relationships in financial data (Citterio, 2024). Several included empirical studies compared conventional and AI/ML approaches within the same experimental setting (Ecer, 2013; Jing & Fang, 2018; Li et al., 2014; Manthoulis et al., 2020). These within-study comparisons provide a more appropriate basis for assessing relative predictive performance than comparisons of headline performance measures across heterogeneous studies.

Accordingly, the larger proportion of studies classified primarily as AI/ML (57.1%) should be interpreted as a descriptive characteristic of

Table 3
The Quality Assessment Criteria for bank failure prediction studies.

Quality domain	Low concern	Moderate/unclear concern	High concern
Sample and event adequacy	Adequate number of banks and failure events relative to the number of predictors and model complexity; sufficiently large longitudinal or panel sample.	Sample or number of failure events is only partly adequate, or reporting is insufficient to establish adequacy confidently.	Few failure events relative to candidate predictors/model complexity, creating substantial risk of overfitting, unstable estimates or poor generalisability.
Failure definition	Failure is clearly and objectively defined using observable institutional events such as closure, receivership, licence withdrawal, regulatory intervention or equivalent formal failure status.	Failure definition is broadly relevant but combines failure with related events or is not reported with sufficient operational detail.	Outcome is primarily a proxy for distress, fragility, weakness or an accounting threshold rather than observed bank failure.
Validation method	Appropriate internal and/or independent validation is reported using suitable discrimination or classification measures, with clear separation between model development and evaluation.	Validation is present but limited, incompletely described, or based mainly on cross-validation/holdout procedures without strong protection against optimism.	Model performance is assessed predominantly on the estimation sample, or no meaningful validation procedure is reported.
Class imbalance	Rare-event imbalance is appropriately recognised and handled using methods such as resampling, SMOTE, class weighting, cost-sensitive learning, stratification or appropriate rare-event metrics.	Class imbalance is limited, partly addressed, or insufficiently reported to determine its likely effect on model performance.	Severe imbalance is ignored, or evaluation relies on overall accuracy without adequate attention to minority-class performance.
Out-of-sample/temporal testing	Independent, external or clearly chronological out-of-sample testing is conducted, preferably using future observations or a separate population.	Holdout testing exists but is random rather than temporal, or the degree of independence between training and testing data is uncertain.	No independent out-of-sample, external or temporal validation is reported.

Notes: The assessment framework was adapted for the banking context from principles used in prediction-model risk-of-bias assessment, including PROBAST and related methodological guidance (Navarro et al., 2021; Wolff et al., 2019). Ratings are domain-based rather than combined into a numerical quality score. "Unclear" is used where reporting is insufficient to support a definitive judgement.

Table 4
Methodological quality assessment of included studies.

#	Study	Sample adequacy	Failure definition	Validation method	Class imbalance	OOS/temporal testing	Overall concern
1	Affes & Hentati-Kaffel (2019)	Low	Low–Moderate	Low	Moderate	Low	Low–Moderate
2	Citterio (2024)	N/A	N/A	N/A	N/A	N/A	Review study
3	Climent et al. (2019)	Low	Low	Low	Moderate	Low	Low–Moderate
4	Ecer (2013)	High	Low	Moderate	Moderate	Moderate–High	High
5	Ekinci & Erdal (2017)	High	Low	Low–Moderate	Moderate	High	High
6	Özel (2013)	High	Low	High	Moderate	High	High
7	Huang, Chang & Liu (2012)	Low	High	High	Unclear–Moderate	High	High
8	Ilk, Pekkurnaz & Cinko (2014)	High	Low	Low–Moderate	Moderate	Moderate–High	High
9	Jing & Fang (2018)	Low	Low	Low	Moderate	Low	Low
10	Kimmel, Thornton & Bennett (2016)	Low	Low	Low	Moderate	Low	Low–Moderate
11	Kristóf & Virág (2022)	Low	Low	Moderate	Moderate	Moderate	Low–Moderate
12	Li, Shi, Huang & Yang (2014)	Low	Low–Moderate	Low–Moderate	Moderate	High	Moderate
13	Li, Feng & Tang (2022)	Low	Low	Low	Moderate	Low	Low
14	Liu, Liu & Sathye (2021)	N/A	N/A	N/A	N/A	N/A	Review study
15	Manthoulis et al. (2020)	Low	Low	Low	Moderate	Low	Low–Moderate
16	Pham & Ho (2021)	Low	Low	Low–Moderate	Low	Low–Moderate	Low–Moderate
17	Wang, Quek & Ng (2016)	Low	Low–Moderate	Low–Moderate	Moderate	Moderate	Low–Moderate
18	Zaghdoudi (2013)	High	High	High	Unclear–Moderate	High	High
19	Zhen-Jia-Liu (2015)	Low	High	Low–Moderate	Moderate	Low	Moderate
20	Barajas, Krakovich & López-Iturriaga (2023)	Low	Low	Moderate–High	Low	High	Moderate
21	Behbood, Lu & Zhang (2013)	Unclear	Unclear	Low–Moderate	Unclear	Low	Unclear–Moderate
22	Behbood, Lu & Zhang (2014)	Low	Low	Low	Low	Low	Low
23	Behbood, Lu, Zhang & Pedrycz (2015)	Low	Low	Low	Low	Low	Low

Notes: Methodological quality assessment was conducted for the 21 primary empirical studies. Liu et al. (2021) and Citterio (2024) are review studies and were therefore not subjected to the empirical prediction-model quality assessment; their domains are reported as N/A. For Behbood, Lu and Zhang (2013), domains that could not be reliably established from the accessible study information were classified as unclear rather than inferred. OOS = out-of-sample. Ratings are domain-based and should not be interpreted as an additive numerical quality score.

the 21 empirical studies included in this review rather than as evidence of methodological superiority. Reported predictive performance varies according to datasets, operational definitions of bank failure, prediction horizons, class distributions, performance measures and validation designs (Citterio, 2024; Liu et al., 2021). The relative performance of econometric and AI/ML models is therefore evaluated within the methodological context in which the reported results were generated.

Furthermore, the results show increasing adoption of AI/ML methods, with 52% of the selected studies using AI/ML as their primary research methodology. This growing application has been associated with the ability of AI/ML methods to model complex relationships and patterns in financial datasets that may be difficult to capture under some conventional model specifications (Medeiros, 2022; Citterio, 2024; Zheng et al. 2023). The increasing volume and complexity of banking data have also encouraged the application of computational and machine-learning approaches to financial prediction problems (Malali and Gopalakrishnan, 2020).

Nevertheless, the greater adoption of AI/ML methods should not be interpreted as evidence that econometric approaches are inherently

Table 5
Classification of included studies by primary modelling approach and study type.

SN	Research Item	Source Journal	Econometric Model	AI/ML Model	Review
1	Huang et al. (2012)	Quality and Quantity	✓		
2	Zaghdoudi (2013)	International Journal of Economics and Financial Issues	✓		
3	Özel (2013)	Pakistan Journal of Statistics and Operation Research	✓		
4	Ilk et al. (2014)	Optimisation	✓		
5	Zhen-Jia-Liu (2015)	RAE Revista de Administração de Empresas	✓		
6	Kimmel et al. (2016)	North American Journal of Economics and Finance	✓		
7	Jing and Fang (2018)	Journal of Forecasting	✓		
8	Manthoulis et al. (2020)	European Journal of Operational Research	✓		
9	Barajas et al. (2023)	European Journal of Management and Business Economics	✓		
10	Behbood et al. (2013)	International Journal of Computational Intelligence and Applications		✓	
11	Behbood et al. (2014)	IEEE Transactions on Industrial Informatics		✓	
12	Li et al. (2014)	Annals of Data Science		✓	
13	Behbood et al. (2015)	IEEE Transactions on Fuzzy Systems		✓	
14	Wang et al. (2016)	AI Communications		✓	
15	Ekinci and Erdal (2017)	Computational Economics		✓	
16	Affes and Hentati-Kaffel (2019)	Annals of Operations Research		✓	
17	Climent et al. (2019)	Journal of Business Research		✓	
18	Pham and Ho (2021)	International Review of Economics and Finance		✓	
19	Kristóf and Virág (2022)	Research in International Business and Finance		✓	
20	Ecer (2013)	Ekonomska Istraživanja		✓	
21	Li, Feng and Tang (2022)	Annals of Operations Research		✓	
22	Liu et al. (2021)	Journal of Risk and Financial Management			✓
23	Citterio (2024)	Socio-Economic Planning Sciences			✓

inferior. Both approaches have methodological strengths and limitations, and their reported performance is influenced by differences in datasets, failure definitions, predictor variables, class distributions and validation procedures. Therefore, while several studies report higher predictive performance for AI/ML models, the heterogeneity of the reviewed studies limits direct cross-study comparison and supports a more cautious conclusion that AI/ML methods may provide incremental predictive value in particular bank-failure prediction settings (Shobana and Umamaheswari, 2021; Citterio, 2024).

3.2. Data and variable selection

The 21 empirical studies included in this review employed heterogeneous datasets and predictor specifications for bank-failure modelling. Bank-level financial variables constituted an important source of predictive information across the literature, with several studies using accounting ratios reflecting dimensions such as capital adequacy, asset quality, profitability, liquidity, leverage and other aspects of bank financial condition (Affes & Hentati-Kaffel, 2019; Climent et al., 2019; Ecer, 2013; Jing & Fang, 2018; Manthoulis et al., 2020; Özel, 2013; Wang et al., 2016). The specific variables, however, differed across studies, reflecting differences in geographical setting, sample period, failure definition, data availability and modelling approach.

The empirical studies also varied substantially in sample size and composition. For example, some studies were based on relatively small samples of failed and non-failed banks, whereas others analysed several thousand banks or bank-year observations. Manthoulis et al. (2020), for example, analysed approximately 6500 US banks over the 2006–2015 period, while other studies employed much smaller country-specific samples (Ecer, 2013; Ekinci & Erdal, 2017; Manthoulis et al., 2020). Such variation is methodologically important because sample size, the number of observed failure events and the proportion of failed to non-failed banks can influence model estimation, validation and the interpretation of reported predictive performance.

Differences were also observed in the number and type of predictor variables used. Some studies relied primarily on bank-specific financial ratios, whereas others incorporated broader sets of financial or efficiency indicators (Li et al., 2022, 2014; Manthoulis et al., 2020). Consequently, the predictor sets used across the empirical literature should not be treated as directly equivalent. These differences form part of the methodological heterogeneity considered in this review when interpreting the comparative performance of bank-failure prediction models.

Among the 23 research items that we reviewed, 21 items were research articles and two items were systematic reviews. The 21 research articles used different sample sizes with some adopting large sample size and others small sample size. Based on Table 6, Manthoulis et al. (2020) focused his study within the United States and adopted the highest sample size of 6500 banks to develop early warning systems (EWS) to predict bank failures. This was followed by Kristóf and Virág (2022) who adopted a sample of 5614 banks in predicting failures of banks in the United States. On the other hand, the study by Zaghdoudi (2013) adopted the smallest sample size of 14 banks to predict bankruptcy among Tunisian banks. Other studies that adopted smaller sample sizes as shown in Table 6 above include Ecer (2013), Ekinci and Erdal (2017), and Ilk et al. (2014) who utilised 34, 37, and 40 banks respectively.

The variations in sample sizes observed among the selected studies could be attributed to several factors, such as differences in the countries where the studies were conducted, the availability of data, and the research questions being addressed by individual studies. Another factor could be the availability of resources such as time and funding at

Table 6
Sample size and variable used by the included studies.

SN	Research Item	CAMELS based Data Source	No. of Banks	No. of Variables
1	Huang D.-T.; Chang B.; Liu Z.-C. (2012)	√	858	5
2	Zaghdoudi T. (2013)	√	14	18
3	Özel G. (2013)	√	70	37
4	Ilk O.; Pekkurnaz D.; Cinko M. (2014)	√	40	21
5	Zhen-Jia-Liu (2015)	√	772	11
6	Kimmel R.K.; Thornton J.H., Jr.; Bennett S.E. (2016)	√	86	45
7	Jing Z.; Fang Y. (2018)	√	293	16
8	Manthoulis G.; Doumpos M.; Zopounidis C.; Galariotis E. (2020)	√	6500	14
9	Liu L.X.; Liu S.; Sathye M. (2021)	Review study	Review study	Review study
10	Barajas A.; Krakovich V.; López-Iturriaga F.J. (2023)	√	954	7
11	Behbood V.; Lu J.; Zhang G. (2013)	√	3103	9
12	Behbood V.; Lu J.; Zhang G.(2014)	√	3103	9
13	Li Y.; Shi Y.; Huang A.; Yang H. (2014)	√	316	12
14	Behbood V.; Lu J.; Zhang G.; Pedrycz W. (2015)	√	3103	9
15	Wang D.; Quek C.; Ng G.S. (2016)	√	3635	9
16	Ekinci A.; Erdal H. (2017)	√	37	35
17	Affes Z.; Hentati-Kaffel R. (2019)	√	1025	10
18	Climent F.; Momparler A.; Carmona P. (2019)	√	168	28
19	Pham X.T.T.; Ho T.H. (2021)	√	180	16
20	Kristóf T.; Virág M. (2022)	√	5614	15
21	Ecer F. (2013)	√	34	14
22	Citterio A. (2024)	Review study	Review study	Review study
23	Li Z.; Feng C.; Tang Y. (2022)	√	4426	21

the disposal of the researchers when they were conducting these studies (Andrade, 2020).

With regard to the types of variables that were selected and incorporated by these studies into their failure prediction models, the results were strikingly similar as most of the 23 selected studies we considered in this review used financial ratios of sampled banks that were principally derived from the CAMELS framework as shown in Table 6 above. The wide acceptance and use of accounting ratios as independent variables in bankruptcy predictions of both financial and non-financial firms by researchers in this scientific domain could be said to have been inspired by the work of Altman (1968) who combined various accounting ratios to develop his popular Zscore model that predicted bankruptcy of non-financial firms, as well as other benefits of accounting ratios. Billios et al. (2024) and Ghazalat and AlHallaq (2024) have argued that accounting ratios such as profitability, liquidity, and leverage ratios have strong power to predict firm failures. They also contended that these ratios are standardised and comparable as they are obtained from financial statements that are prepared following stringent accounting rules such as the International Financial Reporting Standards (IFRS) and Generally Acceptable Accounting Principles (GAAPs), thus making comparison among the firms being studied easier.

The number of accounting ratios selected and incorporated by individual studies differed considerably. While some studies such as Climent et al. (2019), Ekinci and Erdal (2017), Kimmel et al. (2016), Özel (2013) have selected and used a large number of variables in building their models, others selected and incorporated a small number of variables (Barajas et al., 2023; Behbood et al., 2013; Huang et al., 2012; Wang et al., 2016). Despite the differences observed in the amounts of variables used, the selected studies were similar in many respects regarding the type of accounting ratios used as most studies have used profitability, liquidity, and leverage ratios derived from the CAMELS framework (Affes & Hentati-Kaffel, 2019; Behbood et al., 2013; Özel, 2013; Wang et al., 2016).

3.3. Model performance evaluation and validation procedures reported by included empirical studies

Evaluation of the performances of models developed for predicting any future outcomes is critical for result generalisation as it helps researchers determine the accuracy and reliability of their developed models. It also facilitates the development of early warning systems (EWS) that could provide signals to prompt immediate interventions to forestall undesirable future outcomes (Altman et al., 2017). In this section, we have considered the various evaluation metrics that were adopted by the authors of the 23 selected studies included in our review to evaluate the predictive performances of the models they developed to predict bank failures. Table 7 below presents the frequency of various model evaluation metrics and validation procedures recorded by the 23 included papers.

Table 7 demonstrates considerable heterogeneity in the approaches adopted to evaluate bank-failure prediction models. Rather than representing a set of directly comparable performance measures, the evaluation approaches identified in the reviewed literature comprise classification-performance measures, discrimination measures, model-fit and model-selection statistics, and validation procedures. Based on the 20 empirical studies for which sufficient full-text information was available to verify the reported evaluation approaches, overall accuracy or correct classification rate was the most frequently reported measure, appearing in 16 studies (80%). This was followed by sensitivity/recall in nine studies (45%), ROC-AUC in eight studies (40%), specificity in eight studies (40%), and cross-validation in seven studies (35%). Type I and/or Type II errors were reported in four studies (20%), R-square in three studies (15%), AIC in three studies (15%), precision in two studies (10%), and Kolmogorov–Smirnov distance (KSD) in one study (5%). Notably, none of the reviewed empirical studies explicitly reported the F1-score.

Most of the reviewed studies employed more than one approach to evaluate model performance, although the combinations of measures varied substantially across studies (Affes and Hentati-Kaffel, 2019; Zhen-Jia-Liu, 2015; Ekinci and Erdal, 2017; Barajas et al. 2023; Ecer, 2013; Ilk et al. 2014; Jing and Fang, 2018; Pham and Ho, 2021; Wang et al. 2016). The use of multiple complementary measures can provide a more comprehensive assessment of predictive behaviour because different measures capture different aspects of model performance. However, the use of multiple metrics should not itself be interpreted as a mechanism for preventing overfitting. Although Hicks et al. (2022) highlight the importance of rigorous evaluation when assessing machine-learning models, overfitting is primarily controlled through appropriate model-development and validation procedures, including separation of training and evaluation data, cross-validation, regularisation where appropriate, and testing on unseen observations. Thus, reporting several performance metrics can improve the breadth of model assessment but does not substitute for rigorous validation.

Table 7
Predictive performance measures and validation procedures reported by the included empirical studies.

SN	Measure/procedure reported	Category	No. of studies	% of studies	Interpretation
1	Accuracy/correct classification rate	Classification performance	16	80%	Most frequently reported measure, but potentially misleading where the failed-bank class is substantially smaller than the non-failed class.
2	Sensitivity/Recall/True Positive Rate	Class-specific performance	9	45%	Measures the proportion of actual failed banks correctly identified and is particularly relevant for bank-failure early-warning systems.
3	ROC-AUC/AUROC	Discrimination	8	40%	Evaluates the model's ability to discriminate between failed and non-failed banks across classification thresholds.
4	Specificity/True Negative Rate	Class-specific performance	8	40%	Measures the proportion of non-failed banks correctly identified and should be interpreted jointly with sensitivity.
5	Cross-validation	Validation procedure	7	35%	Provides internal assessment of model performance and reduces dependence on a single train–test split; however, it is a validation procedure rather than a predictive-performance metric.
6	Type I and/or Type II errors	Misclassification performance	4	20%	Distinguishes different types of classification errors, which is important because failure to identify a failing bank may have different consequences from incorrectly flagging a healthy bank.
7	R^2/pseudo-R^2	Model fit	3	15%	Provides information on model fit or explanatory power in some econometric models, but is not directly comparable with classification-performance measures.
8	Akaike Information Criterion (AIC)	Model-selection criterion	3	15%	Used for selecting among competing statistical models; lower AIC indicates preferable relative fit, but AIC is not a measure of predictive classification accuracy.
9	Precision	Class-specific performance	2	10%	Measures the proportion of banks predicted to fail that actually failed and is useful when false-positive failure warnings are important.
10	Kolmogorov–Smirnov Distance (KSD)	Discrimination	1	5%	Measures separation between the score distributions of failed and non-failed banks and was used alongside AUROC in one included study.
11	F1-score/F-measure	Combined class-specific performance	0	0%	Not explicitly reported by the included empirical studies. Its absence is notable because F1 combines precision and recall and is useful when evaluating imbalanced classification problems.

Notes: Frequencies are based on the 20 empirical studies for which the reported performance measures could be reliably established from the accessible study evidence. Liu et al. (2021) and Citterio (2024) were excluded because they are review studies. Behbood et al. (2013) was retained in the systematic review and methodological synthesis but excluded from this frequency calculation because its complete set of reported performance measures could not be reliably verified. Exclusion from this frequency calculation should not be interpreted as evidence that particular performance measures were absent from the original study.

The measures reported in Table 7 should also not be treated as interchangeable indicators of predictive performance. Accuracy measures the overall proportion of correctly classified observations; sensitivity or recall measures the proportion of actual failed banks correctly identified; specificity measures the proportion of non-failed banks correctly classified; and precision measures the proportion of banks predicted to fail that actually failed. ROC-AUC assesses discrimination between failed and non-failed banks across classification thresholds. In contrast, R-square principally provides information on model fit or explanatory power in relevant statistical models, while AIC is primarily a model-selection criterion. Cross-validation is a validation procedure rather than a performance metric. Consequently, these measures were not treated as directly equivalent indicators of predictive performance in the present review.

Differences in evaluation practice were observed between studies employing econometric and AI/ML models, although this distinction was not absolute. Studies employing conventional econometric approaches more commonly reported measures of model fit and model selection such as R-square, and AIC (Barajas et al. 2023; Huang et al. 2012; Ilk et al. 2014; Zaghdoudi, 2013; Ozel, 2013; Zhen-Jia-Liu, 2015; Kimmel et al. 2016). AI/ML studies more frequently reported classification and discrimination measures such as accuracy, sensitivity, specificity and ROC-AUC (Ekinci and Erdal, 2017; Kristóf and Virág, 2022; Pham and Ho, 2021; Affes and Hentati-Kaffel, 2019; Wang et al. 2016; Jing and Fang, 2018; Li et al. 2022b; Climent et al. 2019; Behbood et al. 2014, 2015; Ecer, 2013). Nevertheless, several studies evaluated both conventional and AI/ML approaches using common performance measures, meaning that the choice of evaluation criterion cannot be attributed exclusively to model family (Jing and Fang, 2018; Manthoulis et al. 2020).

The distinction between econometric and AI/ML evaluation should therefore be expressed cautiously. The original assumption that econometric models are necessarily linear while AI models are necessarily nonlinear is overly general. Some conventional statistical models impose particular functional-form assumptions, whereas AI/ML methods can accommodate complex nonlinear relationships and interactions. Consequently, conventional goodness-of-fit measures such as R-square cannot automatically be extended to all machine-learning classification

models, and model evaluation should instead be aligned with the statistical properties and predictive objective of the model concerned (Mac Nally, 2018). The appropriateness of an evaluation measure, therefore, depends more fundamentally on the outcome structure, modelling objective, and validation design than on a simple econometric-versus-AI distinction.

A particularly important finding from Table 7 is the predominance of overall accuracy or correct classification rate, which was reported in 80% of the fully appraised empirical studies. Accuracy is widely used because it is intuitive and provides a readily interpretable measure of the proportion of observations correctly classified. However, its interpretation in bank-failure prediction requires considerable caution because failure is generally a minority-class or rare-event outcome. Under class imbalance, a model can obtain apparently high accuracy by correctly classifying the majority of non-failed banks while performing considerably less effectively in detecting actual failures. Thus, reliance on accuracy alone may provide an overly optimistic assessment of model performance (Johnson and Smith, 2023).

This limitation is particularly relevant to the present review because the objective of a bank-failure early-warning model is not merely to maximise the total proportion of correctly classified banks but, critically, to identify institutions at elevated risk of failure. Sensitivity or recall is therefore especially informative because it measures the proportion of actual failures correctly detected. Precision provides complementary information by measuring the proportion of predicted failures that actually fail, while specificity measures the model's ability to correctly identify non-failed banks. Type I and Type II errors are similarly important because the consequences of failing to identify a genuinely failing bank may differ substantially from those of incorrectly flagging a healthy institution as likely to fail (Barajas et al. 2023; Ecer, 2013; Jing and Fang, 2018). The re-audit of the reviewed studies further shows that precision was explicitly reported in only 10% of the fully appraised empirical studies and that none explicitly reported the F1-score. This is an important finding given the class-imbalance characteristics of bank-failure prediction.

The F1-score combines precision and recall and can provide complementary information about minority-class performance where both missed failures and false failure warnings are important. Its absence, together with the comparatively limited reporting of precision, suggests that a substantial part of the literature has emphasised aggregate classification performance more strongly than the joint assessment of minority-class detection and the reliability of positive failure predictions. ROC-AUC was reported by 40% of the fully appraised empirical studies. Unlike accuracy at a particular classification threshold, ROC-AUC assesses a model's ability to discriminate between failed and non-failed banks across a range of thresholds and has consequently become an important evaluation measure in classification research (Isangediok and Gajamannage, 2022). Nevertheless, ROC-AUC should not be characterised as inherently more "advanced" than other measures or as eliminating the methodological challenges created by class imbalance. A model evaluated on a highly imbalanced dataset may still exhibit apparently strong ROC-AUC performance while generating relatively weak precision for the minority failure class. ROC-AUC should therefore be interpreted alongside sensitivity/recall, specificity, precision, class prevalence and the validation design rather than being treated as an intrinsically superior performance measure.

More fundamentally, the substantial heterogeneity among the reviewed studies means that reported performance values cannot be directly compared across studies. The studies differ considerably in sample size, banking system and geographical setting, number and prevalence of failures, definition of bank failure, predictor variables, prediction horizon, treatment of class imbalance, sampling strategy and validation methodology (Affes and Hentati-Kaffel, 2019; Ekinci and Erdal, 2017; Jing and Fang, 2018; Manthoulis et al. 2020; Pham and Ho, 2021; Wang et al. 2016). For example, Jing and Fang (2018) explicitly estimated models using observations from 2002–2009 and evaluated them out-of-sample using 2010 observations, whereas other studies relied on internal partitions or cross-validation. Consequently, performance obtained through an in-sample assessment or random train–test partition is not methodologically equivalent to performance obtained from a genuinely future-period temporal test.

Accordingly, an accuracy of 95% reported by one study cannot automatically be interpreted as evidence of a superior model to an accuracy of 90% reported by another study. Differences in the underlying datasets, failure prevalence, prediction horizons and validation designs may explain part of the observed variation in performance. The present review therefore treats the performance values reported by individual studies as study-specific descriptive evidence rather than as directly comparable estimates from which algorithms can be ranked across studies.

This consideration is particularly important when interpreting exceptionally high predictive-performance values reported for some models, including ANN and hybrid/boosting configurations for which reported accuracy approached or reached 100%. Although several reviewed studies found strong performance for AI/ML techniques (Ecer, 2013; Ekinci and Erdal, 2017; Jing and Fang, 2018; Kristóf and Virág, 2022; Pham and Ho, 2021; Wang et al. 2016), extremely high classification performance should be interpreted cautiously. Perfect or near-perfect performance within a particular dataset does not necessarily imply equivalent performance on unseen banks or future periods. Such estimates may be influenced by sample size and composition, class balancing, predictor-selection procedures, model complexity, overfitting or optimistic validation designs.

Data leakage represents an additional potential source of inflated performance where information from the evaluation sample inadvertently enters model development through preprocessing, feature selection, resampling or other stages. However, the present review does not infer that data leakage occurred merely because a study reported exceptionally high performance. Such an inference would require evidence from the methodology of the individual study. Consequently, the 100% or near-100% accuracy values reported in the literature are retained as findings of the original studies but are not interpreted as evidence of perfect generalisability or universal superiority of the corresponding algorithms.

The methodological quality assessment undertaken in this review provides an important additional context for interpreting these performance results. A high accuracy or ROC-AUC should not automatically be considered strong predictive evidence where the underlying study has substantial methodological concerns relating to sample/event adequacy, definition of bank failure, validation methodology, treatment of class imbalance or absence of out-of-sample testing. Conversely, a somewhat lower performance estimate obtained under rigorous temporal or external validation may provide stronger evidence of generalisable predictive capability than near-perfect performance obtained under less demanding evaluation conditions. This reinforces the need to interpret model performance jointly with methodological quality rather than treating headline performance measures in isolation. For this reason, greater interpretive weight is placed on within-study comparisons, where competing models are evaluated using the same dataset, failure definition, predictor information, prediction horizon and validation procedure.

Several included studies directly compared conventional econometric models with AI/ML approaches under broadly common experimental conditions (Jing and Fang, 2018; Li et al. 2014; Manthoulis et al. 2020), while others compared alternative AI/ML architectures using common data and evaluation designs (Ecer, 2013; Ekinci and Erdal, 2017; Pham and Ho, 2021; Wang et al. 2016). Such comparisons provide a more credible basis for assessing relative predictive performance because major sources of between-study heterogeneity are held approximately constant.

The evidence from these within-study comparisons indicates that AI/ML models frequently achieved competitive or higher reported predictive performance than conventional approaches in particular applications (Ecer, 2013; Jing and Fang, 2018; Li et al. 2014; Manthoulis et al. 2020). However, this advantage was not uniform across models, prediction horizons, datasets or performance measures. For example, Jing and Fang (2018) found that although data-mining models performed strongly in-sample, the logit model generated fewer missed failures and false alarms in the out-of-sample period. Such evidence demonstrates why model superiority cannot be established solely from headline accuracy and why validation context and the particular regulatory cost of classification errors must also be considered.

Consequently, it cannot be argued that econometric based models performed poorly in predicting bank failures as this cannot be supported considering the differences in study design and methodological quality. Although some conventional econometric models reported lower performance than AI/ML models, the observed differences cannot be attributed solely to model class because the studies differed substantially in datasets, failure prevalence, predictor specification, prediction horizons, class-imbalance treatment and validation designs. The evidence, therefore, supports a more qualified conclusion: AI/ML models frequently demonstrate potential incremental predictive value over conventional econometric approaches in within-study comparisons, but the magnitude and consistency of this advantage vary across studies and evaluation criteria and do not establish universal superiority (Jing and Fang, 2018; Li et al. 2014; Manthoulis et al. 2020).

Overall, Table 7 reveals both substantial heterogeneity in model-evaluation practices and an important methodological limitation within the bank-failure prediction literature. The predominance of accuracy, comparatively lower reporting of class-specific measures such as precision, and the absence of explicitly reported F1-scores indicate that aggregate classification performance has frequently received greater attention than minority-class predictive performance. Future bank-failure prediction research would therefore benefit from more systematic reporting of complementary measures, including sensitivity/recall, specificity, precision and F1-score, alongside discrimination measures such as ROC-AUC and, where appropriate, calibration measures. These should be accompanied by rigorous temporal or external validation and transparent reporting of class-imbalance treatment. Accordingly, the present review interprets model performance jointly with methodological quality and treats reported performance estimates as descriptive, study-specific evidence rather than as a basis for directly ranking econometric and AI/ML algorithms across heterogeneous studies.

3.4. Models and algorithms selection

The 21 included primary studies reviewed in this systematic review have used various models and algorithms to predict future bank failures. The selection of models and algorithms is a very important aspect of both econometric and AI modelling projects, as choosing the right models and algorithms ensures an improvement in the performance and efficiency of the resulting early warning systems and promotes the efficiency, scalability, and interpretability of the applications that are being developed (Zöller & Huber, 2021).

We observed that both studies that were econometric-based and AI-based selected and adopted diverse techniques and algorithms to achieve the common goal of predicting bank failures. Table 8 below shows the techniques and algorithms adopted by both strands of studies we have included in this review.

Table 8 demonstrates the methodological diversity of the 21 primary empirical studies included in the review. Because several studies evaluated more than one modelling technique, the frequencies reported in the table represent the number of studies in which a particular technique or algorithm was evaluated rather than mutually exclusive classifications of the studies.

Table 8
Techniques and algorithms reported in the included empirical studies.

SN	Technique/Algorithm	No. of Studies	% of Studies
1	Logistic regression/logit model	7	33.3%
2	Fuzzy bridge refinement	3	14.3%
3	Neural fuzzy inference	2	9.5%
4	Deep-learning neural network (DL-NN)	1	4.8%
5	Hybrid modelling approaches	4	19.0%
6	Data envelopment analysis (DEA)	1	4.8%
7	Multiple discriminant analysis	1	4.8%
8	Cox proportional-hazards model	1	4.8%
9	Artificial neural networks (ANN)	2	9.5%
10	Support vector machines (SVM)	3	14.3%
11	Classification and regression trees (CART)	1	4.8%
12	Multivariate adaptive regression splines (MARS)	1	4.8%
13	Other survival modelling approaches	1	4.8%

Note: Percentages are calculated using the 21 primary empirical studies as the denominator. Liu et al. (2021) and Citterio (2024) are review studies and are excluded from the calculation. Individual empirical studies may evaluate more than one technique or algorithm; consequently, the categories are not mutually exclusive and the percentages are not expected to sum to 100%.

Logistic regression was the most frequently identified individual modelling technique, appearing in seven of the 21 empirical studies (33.3%) (Barajas et al., 2023; Huang et al., 2012; Ilk et al., 2014; Kimmel et al., 2016; Özel, 2013; Zaghdoudi, 2013; Zhen-Jia-Liu, 2015). Its continued use in bank-failure research reflects, among other considerations, its suitability for binary outcomes and the relative interpretability of estimated relationships between predictor variables and the probability of failure. Nevertheless, the appropriateness of logistic regression depends on model specification and the characteristics of the data, and its predictive performance should therefore be evaluated empirically rather than assumed from model class alone (Citterio, 2024).

The reviewed literature also demonstrates substantial use of AI/ML and hybrid approaches. Fuzzy bridge refinement was reported in three studies (14.3%), support vector machines in three studies (14.3%), artificial neural networks in two studies (9.5%), and neural fuzzy inference approaches in two studies (9.5%). Hybrid modelling approaches were identified in four studies (19.0%). These frequencies indicate that the post-Global Financial Crisis literature has explored a diverse range of computational approaches alongside conventional econometric models.

Other techniques were less frequently represented in the included evidence base. These included data envelopment analysis (DEA), multiple discriminant analysis, Cox proportional-hazards modelling, classification and regression trees (CART), multivariate adaptive regression splines (MARS), deep-learning neural-network approaches and other survival modelling approaches. Their lower frequency within this particular evidence base should not be interpreted as evidence of inferior predictive performance; it indicates only that these methods were less frequently represented among the 21 empirical studies identified by the review.

Importantly, the frequencies in Table 8 describe the methodological composition of the reviewed literature and do not provide a ranking of model performance. Several included studies compared multiple econometric, AI/ML or hybrid approaches within the same empirical design (Affes & Hentati-Kaffel, 2019; Behbood et al., 2014; Jing & Fang, 2018; Manthoulis et al., 2020). Accordingly, comparative predictive performance is considered separately in the subsequent section, with greater weight placed on within-study comparisons, validation design, prediction horizon, class distribution and the performance measures reported by the original studies.

3.5. Predictive performance and validation of bank-failure models

The predictive performance of bank-failure models is a central concern in the development of Early Warning Systems (EWS), as banks

and regulatory authorities require models that can identify institutions at elevated risk of failure with sufficient reliability and timeliness to support supervisory intervention (Citterio, 2024). However, predictive performance cannot be evaluated meaningfully using overall classification accuracy alone. This is particularly important in bank-failure prediction, where failures may represent a relatively small proportion of observations and where model performance is influenced by the operational definition of failure, class distribution, prediction horizon, sampling strategy, decision threshold and validation design.

Accordingly, the present review does not rank the included studies according to their reported headline accuracy. Instead, the predictive evidence from the 21 primary empirical studies is evaluated within the methodological context in which it was generated. The two review studies, Liu et al. (2021) and Citterio (2024), are retained for contextual synthesis but are excluded from the empirical performance comparison. Where an empirical study evaluated multiple models or algorithms, the analysis retains the reported comparative evidence rather than selecting only the model with the highest accuracy. This avoids systematically privileging studies that evaluated larger numbers of candidate algorithms and preserves the experimental conditions under which competing models were compared.

Table 9 therefore presents a structured evidence matrix covering the operational definition of bank failure, geographical setting and observation period, sample and failure-event composition, treatment of class imbalance, models evaluated, prediction horizon, validation and temporal splitting procedures, reported performance measures, and relevant temporal-validity considerations. Performance measures such as accuracy, sensitivity or recall, specificity, precision, Type I and Type II errors, and ROC-AUC are interpreted jointly where reported rather than treated as interchangeable indicators of predictive quality. Particular attention is also given to whether performance was evaluated in-sample, through cross-validation or random holdout procedures, or through temporally separated out-of-sample testing.

These methodological distinctions are necessary because the reported performance estimates are not generated under equivalent experimental conditions. A higher accuracy or AUC reported in one study does not necessarily indicate a superior modelling approach to a lower value reported in another study when the underlying studies differ in failure prevalence, outcome definition, prediction horizon, predictor information and validation design. Consequently, greater interpretive weight is placed on within-study comparisons in which competing models are evaluated using broadly common data, outcome definitions, prediction horizons and validation procedures. The evidence matrix is therefore intended to support a structured assessment of the conditions under which different modelling approaches perform effectively rather than a cross-study hierarchy based on headline predictive accuracy.

The empirical studies included in this review demonstrate substantial diversity in the modelling approaches used to predict bank failure. Many studies evaluated more than one econometric, AI/ML, or hybrid modelling approach rather than relying on a single specification (Affes & Hentati-Kaffel, 2019; Behbood et al., 2014; Jing & Fang, 2018; Manthoulis et al., 2020). However, where several models were evaluated within the same study, the present synthesis does not select only the model reporting the highest headline accuracy. Such an approach could introduce optimistic selection bias and disproportionately favour studies that evaluated a larger number of candidate algorithms. Instead, the comparative results are retained within the methodological setting in which they were generated, with greater interpretive weight placed on within-study comparisons where competing models were evaluated using broadly common data, outcome definitions, prediction horizons and validation procedures.

As shown in Table 9, substantial heterogeneity exists across the empirical evidence base. The studies differ in geographical setting, observation period, sample size, failure prevalence, operational definition of bank failure, predictor variables, treatment of class imbalance, prediction horizon and validation design. Reported predictive-performance measures also vary considerably and include overall accuracy or correct classification, sensitivity or recall, specificity, precision, Type I and Type II errors, ROC-AUC and other model-specific measures. Consequently, a predictive accuracy of, for example, 95% in one study cannot automatically be interpreted as evidence of superior predictive performance relative to a lower accuracy reported in another study. The estimates were generated under different experimental conditions and are therefore not directly comparable.

This issue is particularly important because bank failure is commonly a rare-event classification problem. Overall accuracy can provide an incomplete or potentially misleading representation of predictive performance when failed banks constitute a relatively small proportion of the sample. A model may correctly classify a large proportion of surviving banks and consequently achieve high overall accuracy while still failing to identify a substantial proportion of the failed institutions that are of primary interest to regulators. For this reason, the present review interprets accuracy jointly with class-specific measures such as sensitivity or recall, specificity, precision and Type I and Type II errors, together with threshold-independent measures such as ROC-AUC where these are reported. The treatment of class imbalance is also considered when interpreting the reported results.

The validation design provides a further important basis for distinguishing the strength of the predictive evidence. Some studies employed cross-validation or training/test partitions, whereas others incorporated explicit temporal separation between model development and subsequent evaluation. For example, Jing and Fang (2018) estimated their models using observations from 2002–2009 and evaluated predictive performance using 2010 observations (Jing & Fang, 2018). Behbood et al. (2014) similarly trained their models using earlier observations and subsequently evaluated performance in later test periods, including 1995, 1998 and 2000 (Behbood et al., 2014). Their use of SMOTE on the training data rather than the subsequent test observations is also methodologically relevant when interpreting the reported predictive results. Wang et al. (2016), meanwhile, explicitly evaluated early-warning performance at one- and two-year prediction horizons. These designs provide different evidence from studies reporting primarily in-sample classification or internal validation, and their reported performance estimates should not be interpreted as methodologically equivalent.

Prediction horizon is similarly important for interpreting model performance. The supervisory usefulness of a bank-failure model depends not only on whether it correctly identifies a failing institution, but also on whether the warning is generated sufficiently early to permit meaningful intervention. The included studies vary in this respect. Some explicitly evaluate predictions at defined pre-failure horizons, while others report contemporaneous classification or do not specify the forecasting horizon sufficiently clearly. For example, Climent et al. (2019) evaluate predictive performance across multiple pre-failure horizons, while Wang et al. (2016) consider one- and two-year early-warning scenarios. Consequently, models evaluated at different prediction horizons should not be ranked solely according to their reported classification performance.

The treatment of class imbalance also varies materially across the studies. Behbood et al. (2014), for example, apply SMOTE to the training data in a sample containing substantially fewer failed than surviving banks (Behbood et al., 2014), while Manthoulis et al. (2020) incorporate random undersampling through RUSBOOST in a large US banking dataset containing 430 failures among approximately 59,347 bank-year observations (Manthoulis et al., 2020). Other studies employ approximately balanced samples by design. Ecer (2013), for example, evaluates ANN and SVM using 17 failed and 17 non-failed banks. Although a balanced design can facilitate comparison between algorithms, predictive performance obtained from such a sample should not be interpreted as directly equivalent to performance in a naturally imbalanced banking population. These differences further demonstrate why headline accuracy alone provides an insufficient basis for cross-study comparison.

Table 9
Structured evidence matrix of methodological characteristics, validation designs and predictive performance of the included empirical studies.

Study	Failure definition	Geography/period	Sample/failures	Class balance/treatment	Models evaluated	Prediction horizon	Validation/temporal split	Metrics reported	Temporal validity considerations
Huang, Chang & Liu (2012)	Financial distress proxied by current-year cash flow relative to liabilities rather than observed regulatory closure	Developing and developed-country banks; study period as reported in original study	858 banks; 2114 observations	No explicit resampling procedure identified	Bank-failure/ distress prediction model based on financial ratios	NR	No clearly separated temporal or external predictive validation identified	Classification/ predictive performance measures	High interpretive caution: outcome is a distress proxy rather than observed bank closure; temporal OOS protection not established
Zaghdoudi (2013)	Banking weakness/ failure index constructed from changes in deposits, credit and borrowing	Tunisia; 2002–2010	14 banks	No explicit imbalance treatment reported	Logistic regression	NR	Model estimated and evaluated using the study sample; no independent temporal OOS test identified	Accuracy and statistical /model-fit measures	Temporal generalisability unclear; outcome is a banking-weakness proxy rather than observed closure
Özel, G. (2013)	Banks classified as failed/ transferred to the Savings Deposit Insurance Fund (SDIF)	Turkey; 2000–2008	70 banks; failed and non-failed banks	No explicit resampling procedure reported	Cox proportional hazards, stratified Cox, ordinary logistic regression and conditional logistic regression	NR	Reported logistic classification is in-sample; no independent temporal holdout identified	Correct classification; likelihood/ model-fit statistics	No independent temporal OOS validation identified; look-ahead and survivorship protection not explicitly established
Ilk, Pekkurnaz & Cinko (2014)	Bank failure status as defined in the original Turkish-bank study	Turkey; longitudinal study period reported by authors	Approximately 40 banks	No explicit synthetic balancing procedure identified	Longitudinal/ econometric failure models including logistic-regression-based analysis	NR	Internal model evaluation; independent temporal validation not clearly established	Classification and model-fit measures	Temporal-validity protection unclear from reported design; survivorship considerations NR
Liu, Z.-J. (2015)	Financial distress proxied by current cash flow being insufficient relative to liabilities	Cross-country sample; 2002–2012	772 banks; 6773 bank-year observations	No explicit synthetic resampling procedure reported	Bank financial-distress prediction models	NR	In-sample and out-of-sample assessments reported	Accuracy, sensitivity, specificity and ROC-based measures	OOS assessment improves validation; interpretation remains constrained by use of a distress proxy rather than observed bank closure
Kimmel, Thornton & Bennett (2016)	Problem/failing-bank status based on supervisory/ bank-failure classification used in the study	United States; period reported in original study	Large US banking sample	No explicit synthetic balancing procedure reported	Logit, Cox proportional-hazards and multiple-discriminant approaches	Early-warning prediction prior to bank problems/ failure	Predictive evaluation beyond model estimation reported	Classification and model-performance measures	Temporal structure strengthens early-warning interpretation; exact look-ahead and survivorship protection should be interpreted from the reported design
Jing & Fang (2018)	Observed US bank failures	United States; 2002–2010	293 banks; 16 predictors	No synthetic balancing procedure identified in the principal comparison	Logit and multiple data-mining models, including ANN	Prediction using information available prior to failure	Models estimated on 2002–2009 data and evaluated on 2010 observations	Accuracy, missed failures, false alarms and other classification measures	Explicit temporal OOS design reduces look-ahead concern; within-study comparison is more informative than cross-study accuracy ranking
Manthoulis et al. (2020)	Failed versus surviving US banks, extended to ordinal pre-failure states	United States; 2006–2015	Approx. 6500 banks; 59,347 bank-year observations; 430 failures	Class imbalance explicitly addressed; random undersampling used within RUSBOOST	Ordinal classification, logistic regression, SVM and RUSBOOST	Multiple pre-failure states represented through an ordinal framework	Comparative validation incorporating resampling/ testing procedures	AUROC, KSD and classification measures	Strong attention to rare-event prediction and validation; interpretation should retain distinction between resampling-based and temporal validation
Barajas, Krakovich & López-Iturriaga (2023)	Failure includes cessation/ loss of banking licence, including liquidation, licence withdrawal and relevant exit events	Russia; 2012–2019	954 banks; 13,578 bank-quarter observations	Failure prevalence explicitly represented; no synthetic balancing required/ reported	Logistic-regression/ survival-oriented bank-failure analysis	NR	Model estimation and predictive assessment; no independent temporal/ external holdout clearly identified	Correct classification and regression/ model-performance measures	Independent temporal generalisability remains uncertain; failure definition is broader than simple regulatory closure
Behbood, Lu & Zhang (2013)	Failed and surviving US banks	United States; historical banking sample	3103 banks reported in the study dataset	Imbalanced failure data addressed within modelling framework	Fuzzy-logic/ fuzzy-bridge refinement approach	Pre-failure prediction	Training/ testing evaluation reported	Accuracy/ classification measures	OOS testing reported; some design details remain insufficiently reported for stronger temporal-bias judgement

(*continued on next page*)

Table 9 (*continued*).

Study	Failure definition	Geography/period	Sample/failures	Class balance/treatment	Models evaluated	Prediction horizon	Validation/temporal split	Metrics reported	Temporal validity considerations
Behbood, Lu & Zhang (2014)	Observed failed and surviving US banks	United States; historical sample extending to 2000	3103 banks: 548 failed and 2555 surviving	SMOTE applied to training data only	Fuzzy-bridge refinement and comparator models	Pre-failure prediction	Training observations through 1990; later test periods include 1995, 1998 and 2000	Accuracy, sensitivity, specificity and geometric mean	Temporal separation and training-only SMOTE reduce leakage concern and provide comparatively strong OOS evidence
Li, Shi, Huang & Yang (2014)	Failed and non-failed commercial banks based on FDIC-related banking data	United States; period reported in original study	316 banks	No explicit synthetic resampling procedure identified	Trait Recognition (TR), SVM and hybrid TR–SVM	NR	Training/testing predictive comparison reported; independent temporal validation not clearly established	Overall accuracy, precision, TPR and TNR	Multiple class-specific metrics improve interpretation; temporal generalisability remains less certain
Behbood, Lu, Zhang & Pedrycz (2015)	Observed failed and surviving US banks	United States; 1979–2000	3103 banks: 548 failed and 2555 surviving	SMOTE applied to training data only	Fuzzy-rule/ fuzzy-bridge refinement prediction system	Pre-failure prediction	Training through 1990; subsequent temporal tests include 1995, 1998 and 2000	Sensitivity, specificity, geometric mean and classification performance	Explicit temporal testing and training only oversampling provide comparatively strong protection against optimistic evaluation
Wang, Quek & Ng (2016)	Failed versus non-failed US banks identified from longitudinal financial information	United States; approximately 21 years	3635 banks; nine financial covariates	Training samples constructed to control failed/ non-failed composition	GARSINFIS neural-fuzzy system and benchmark models	One-year and two-year prior prediction	Repeated training/ testing experiments across prediction horizons	Accuracy, sensitivity and specificity	Explicit one- and two-year horizons strengthen temporal interpretation; sensitivity and specificity permit minority-class assessment
Ekinci & Erdal (2017)	Failed versus non-failed banks	Bank sample reported by authors	37 banks: 17 failed and 20 non-failed; 35 financial ratios	Approximately balanced by study design	Hybrid prediction model and comparator models	NR	Stratified 10-fold cross-validation	Classification rate, sensitivity, specificity and AUC	Cross-validation reduces dependence on a single split but does not constitute independent temporal/ external validation
Affes & Hentati-Kaffel (2019)	Bankruptcy/ failure status of US banks	United States; complete market-cycle sample with annual evaluation	1025 banks reported in review extraction	Natural class imbalance retained; failed and non-failed classification reported separately	CART, MARS, K-means and hybrid K-means–MARS	Annual failure forecasting	K-fold cross-validation with training/ testing assessment	Correct classification, sensitivity, specificity, Type I error, Type II error and AUC	Class-specific errors provide stronger rare-event information than accuracy alone; validation is internal rather than independent external validation
Climent, Momparler & Carmona (2019)	Failed/ distressed Eurozone banks identified using the study's distress/ failure criterion	Eurozone; 2006–2016	168 banks; 28 variables reported in review extraction	Class imbalance considered in model development/ evaluation	XGBoost and comparative prediction framework	One-, two- and three-year pre-failure horizons	Machine-learning training/ testing across pre-failure horizons	Accuracy, AUC, sensitivity and specificity	Explicit horizons improve temporal interpretation; outcome coding and classification threshold should be considered when comparing results with other studies
Pham & Ho (2021)	Failed and non-failed banks according to the study's bank-failure classification	Bank sample reported by authors	180 banks; 16 predictors	Class imbalance explicitly addressed in modelling/ evaluation	Boosting algorithms, including XGBoost and hybrid/ ensemble approaches	Pre-failure prediction	Training/ testing and comparative boosting-model evaluation	Accuracy, precision, recall and AUC	Use of precision/ recall is valuable under imbalance; reported headline accuracy should be interpreted within the study's sampling and validation design
Kristóf & Virág (2022)	Bank failure status across EU-27 banking sample	EU-27; study period reported by authors	5614 banks; 15 variables reported in review extraction	No explicit synthetic balancing procedure identified	C5.0 decision trees and deep-learning neural networks	Pre-failure prediction	Separate training and testing performance reported	Classification accuracy and AUROC	Training AUROC and testing AUROC are reported separately, permitting assessment of overtraining; independent external validation not identified

(*continued on next page*)

Table 9 (*continued*).

Study	Failure definition	Geography/period	Sample/failures	Class balance/treatment	Models evaluated	Prediction horizon	Validation/temporal split	Metrics reported	Temporal validity considerations
Ecer F. (2013)	Failed versus non-failed banks	Turkey; study period reported by authors	34 banks: 17 failed and 17 non-failed; 14 selected financial ratios	Balanced failed/ non-failed sample by design	Artificial Neural Network (ANN) and Support Vector Machine (SVM)	NR	Training/validation comparison	Classification accuracy and comparative predictive performance	Balanced design facilitates model comparison but differs substantially from natural rare-event prevalence; temporal/ external validation is limited
Li, Feng & Tang (2022)	Failed versus non-failed US commercial and savings banks	United States; 2002–2016	4426 banks; 21 variables reported in review extraction	No explicit synthetic oversampling procedure reported	Malmquist DEA with Worst Practice Frontier and alternative static/ dynamic DEA specifications	Dynamic multi-period prediction	Dynamic and static specifications compared longitudinally; robustness and paired statistical testing performed	AUC and other predictive/ classification indicators; paired tests of model performance	Longitudinal design incorporates historical information; dynamic and static specifications should be compared within-study rather than ranked against unrelated algorithms

Notes: NR = not reported or not sufficiently specified in the original study; OOS = out-of-sample; AUC/AUROC = area under the receiver operating characteristic curve; ANN = artificial neural network; SVM = support vector machine; DEA = data envelopment analysis; SMOTE = Synthetic Minority Over-sampling Technique; TPR = true-positive rate; TNR = true-negative rate; KSD = Kolmogorov–Smirnov distance; GARSINFIS = Genetic Algorithm and Rough Set Incorporated Neural Fuzzy Inference System. NR should not be interpreted as evidence that a methodological procedure was not undertaken; it indicates that the information could not be established sufficiently reliably from the reported study.

The table contains the 21 primary empirical studies. Liu et al. (2021) and Citterio (2024) are review studies and are therefore excluded from this empirical evidence matrix. Reported performance measures are presented as study-specific descriptive evidence and are not treated as directly comparable estimates across studies because of differences in failure definitions, geographical settings, observation periods, sample composition, class prevalence, predictor sets, prediction horizons, decision thresholds and validation designs.

Calibration assessment was recorded where explicitly reported. Where no clear assessment of agreement between predicted probabilities and observed outcomes could be established, calibration was treated as NR rather than assumed to be absent. Similarly, potential look-ahead or survivorship concerns were classified conservatively and were not inferred where the original study provided insufficient information.

The very high predictive-performance values reported by some studies therefore require interpretation within their respective experimental designs rather than as evidence of an absolute hierarchy of algorithms. Studies including Pham and Ho (2021), Behbood et al. (2014), Li et al. (2022), Jing and Fang (2018) and other AI/ML and econometric applications report strong predictive results under particular modelling conditions (Behbood et al., 2014; Jing & Fang, 2018; Li et al., 2022; Pham & Ho, 2021). However, exceptionally high reported accuracy may be influenced by sample composition, class balancing, feature selection, prediction horizon, decision threshold and the method used to separate model-development and evaluation data. High reported performance is therefore not treated in this review as sufficient evidence, by itself, that one model or modelling family is generally superior to another. Nor is data leakage inferred solely from a high performance estimate; such a conclusion would require evidence from the underlying study design.

The evidence also indicates that sample size alone does not explain differences in reported predictive performance. Some studies use several thousand banks, including Behbood et al. (2014) and Li et al. (2022), whereas others obtain strong reported performance from substantially smaller samples. The informational content of the predictors, number of failure events, class distribution, model specification, feature-selection procedure, prediction horizon and validation strategy are therefore also relevant. Financial statement variables and accounting ratios are prominent predictors throughout the literature and continue to provide important information concerning bank capitalisation, asset quality, profitability and liquidity. Their frequent use across both econometric and AI/ML studies suggests that the predictive information contained in bank-level financial characteristics remains central to bank-failure modelling, although the way in which different modelling approaches exploit that information varies.

The within-study comparative evidence provides a more appropriate basis for assessing the relative predictive performance of AI/ML and conventional econometric approaches than comparisons of headline performance measures across different studies. Several included studies evaluated competing model families using broadly common datasets, outcome definitions and evaluation procedures (Ecer, 2013; Jing & Fang, 2018; Li et al., 2014; Manthoulis et al., 2020).

Within these comparative settings, AI/ML approaches frequently produced competitive or higher predictive performance according to one or more of the measures reported by the original studies. However, this pattern was not uniform across algorithms, datasets, prediction horizons or evaluation criteria. Consequently, the evidence does not support a general conclusion that AI/ML models are universally superior to conventional econometric approaches.

Jing and Fang (2018) provide an important illustration of this qualification. Although the data-mining models evaluated in their study demonstrated strong predictive performance under some evaluation conditions, logistic regression produced fewer missed failures and false alarms in the temporally separated out-of-sample evaluation. This finding demonstrates that the relative performance of competing approaches may depend on the performance criterion being considered and that higher performance on one measure does not necessarily imply superior supervisory usefulness across all relevant criteria.

The evidence therefore suggests that the potential incremental predictive value of AI/ML is conditional on the empirical and methodological setting in which the models are evaluated. Model architecture alone does not determine predictive performance. Sample composition, failure prevalence, predictor information, class-imbalance treatment, prediction horizon, decision threshold and validation design also influence the reported results. Accordingly, the present review interprets AI/ML as providing potential incremental predictive value in particular bank-failure prediction settings rather than as constituting a universally superior alternative to econometric modelling.

From an early-warning-system perspective, this distinction is important. A model intended to support supervisory intervention should be evaluated not only according to aggregate predictive performance but also according to its ability to identify failed or distressed institutions sufficiently early, its false-negative and false-positive behaviour, and its performance on observations that were not used during model development. The practical value of a bank-failure prediction model therefore depends jointly on predictive performance, validation quality, prediction horizon and the decision context in which the model is intended to operate.

3.6. Integration of external macroeconomic and non-financial data sources

The evidence in the literature has shown that bank failures are caused by both internal and external factors, therefore, to gain a complete understanding of these causes and predict future failures of banks, both internal and external factors must be considered in forecasting bank failures (Heffernan, 2003; Zaghdoudi, 2013). In this section, we reviewed the selected studies based on their level of integration of external macroeconomic variables and non-financial data like Environmental Social and Governance (ESG) performance of banks into the models they have developed to predict bank failures. The table below presents our analysis of the level of data points that were used by the selected studies to develop models that predicted bank failures (see Table 10).

The evidence reviewed indicates that bank-level financial variables dominate the predictor sets used in the primary empirical studies, while broader macroeconomic and non-financial risk factors are less consistently represented. Some empirical studies incorporated variables extending beyond conventional bank-level accounting indicators; however, the extent and specification of these variables varied across studies. Consequently, the evidence does not support the conclusion that macroeconomic information is entirely absent from the empirical bank-failure prediction literature.

The comparatively limited and heterogeneous incorporation of broader risk factors nevertheless identifies an important area for further research. Macroeconomic conditions such as economic growth, inflation, interest rates and labour-market conditions may affect the environment in which bank-level vulnerabilities develop. Their potential contribution to bank-failure prediction should therefore be evaluated alongside bank-specific financial indicators rather than assumed to provide either universally superior or negligible predictive information.

Environmental, social and governance (ESG) factors represent a further potential extension of the predictor space. Citterio (2024), as a review study rather than a primary model-development study, discusses the relevance of broader determinants of bank distress and failure. The evidence identified in the present review, however, does not provide a sufficient basis for concluding that ESG variables have been systematically incorporated into the primary empirical bank-failure prediction models. This should therefore be interpreted as an area requiring further empirical investigation rather than as an established determinant of predictive performance.

This is a startling observation as there is strong evidence in the literature that documented the influence that macroeconomic variables like the Gross Domestic Product (GDP), inflation rate, interest rate and unemployment have on the financial stability of banks. Mare (2015) documented evidence from a study he conducted in Italy in which he argued that distressed conditions in the macroeconomic environment can escalate the risk of bank failing to fulfil their obligations to depositors and creditors. Citterio (2024) conducted a systematic review of 76 research papers that developed models to predict bank failures and concluded that bank failures are multifaceted in nature as they are caused by several factors including micro-level variables within the bank as well as macroeconomic variables external to the banks. He argued that these factors reinforce each other and build up over time prior to the actual occurrence of the main failures.

However, evidence from the broader literature provides additional context for this research gap. Wu and Cole (2024) had argued that internal bank level data such as accounting ratios are more sensitive and

Table 10
Integration of macroeconomic variables in reviewed studies.

SN	Research Item	Accounting ratios	Macroeconomic variables	ESG Variables
1	Huang D.-T.; Chang B.; Liu Z.-C. (2012)	YES	NO	NO
2	Zaghdoudi T. (2013)	YES	NO	NO
3	Özel G. (2013)	YES	NO	NO
4	Ilk O.; Pekkurnaz D.; Cinko M. (2014)	YES	NO	NO
5	Zhen-Jia-Liu (2015)	YES	NO	NO
6	Kimmel R.K.; Thornton J.H., Jr.; Bennett S.E. (2016)	YES	NO	NO
7	Jing Z.; Fang Y. (2018)	YES	NO	NO
8	Manthoulis G.; Doumpos M.; Zopounidis C.; Galariotis E. (2020)	YES	NO	NO
9	Liu L.X.; Liu S.; Sathye M. (2021)	YES	NO	NO
10	Barajas A.; Krakovich V.; López-Iturriaga F.J. (2023)	YES	NO	NO
11	Behbood V.; Lu J.; Zhang G. (2013)	YES	NO	NO
12	Behbood V.; Lu J.; Zhang G.(2014)	YES	NO	NO
13	Li Y.; Shi Y.; Huang A.; Yang H. (2014)	YES	NO	NO
14	Behbood V.; Lu J.; Zhang G.; Pedrycz W. (2015)	YES	NO	NO
15	Wang D.; Quek C.; Ng G.S. (2016)	YES	NO	NO
16	Ekinci A.; Erdal H. (2017)	YES	NO	NO
17	Affes Z.; Hentati-Kaffel R. (2019)	YES	NO	NO
18	Climent F.; Momparler A.; Carmona P. (2019)	YES	NO	NO
19	Pham X.T.T.; Ho T.H. (2021)	YES	NO	NO
20	Kristóf T.; Virág M. (2022)	YES	NO	NO
21	Ecer F. (2013)	YES	NO	NO
22	Citterio A. (2024)	YES	YES	YES
23	Li Z.; Feng C.; Tang Y. (2022)	YES	NO	NO

essential determinants of bank failures than external macroeconomic variables. According to them, external macroeconomic variables influence only the risk of failure of weak and vulnerable banks. Therefore, they discounted the significance of external macroeconomic variables in predicting bank failures.

Another risk factor in banking system that has increasingly drawn the attention of investors, regulators, and market participants is the Environmental Social and Governance (ESG) risks that banks are exposed to. Citterio (2024) argued that integration of ESG risk factors into models that are designed to predict bank failures has become very important because the ESG performance of a bank can impact its financial performance and contributes the probability of failure of the bank. However, in another related study by Menicucci and Paolucci (2023) where the author investigated the impact of ESG performance on bank performance among Italian banks, they argued that there was a negative relationship between ESG performance of banks and their market and operational performance. The authors further submitted that the results on all dimensions of ESG and their effects on bank performance were mixed. According to them, improving the environmental dimension of ESG through waste reduction and emission controls resulted in improved bank operational and market performance. However, improved performance in the social dimension of ESG through responsible product design and marketing resulted in a decrease in accounting profits of banks.

3.7. Directions for future research

The following discussion presents forward-looking research directions arising from gaps identified in the reviewed evidence and should not be interpreted as an empirical finding of the 21 primary studies included in this systematic review. In particular, generative artificial intelligence (GenAI) and synthetic-data methods were not sufficiently represented in the included empirical evidence to permit conclusions regarding their effectiveness for bank-failure prediction. They are therefore considered here as potential directions for future research rather than as established findings of the review.

As artificial intelligence (AI) research and development progress and more sophisticated AI models are being developed, there is a greater chance that researchers, bank regulators, and financial institutions will find opportunities to innovate and adopt more advanced and secure AI algorithms for predicting bank failures. One of these innovations is the emergence of Generative AI (GenAI), which has given birth to a set of novel models that are capable of creating new contents such as texts, videos, images, audios, codes, etc. Zhao et al. (2024).

GenAI has several branches, such as the popular Large Language Models (LLM), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, etc. These subsets have distinct capabilities that enable each of them to excel in the creation of specific outputs. For example, while LLM excel in creating textual outputs; diffusion models excel in creating images as outputs (Bandi et al., 2023).

The application of GenAI in finance is gaining rapid momentum considering the continuous generation of large volumes of data in the industry and the willingness of financial firms such as banks to invest in innovations that could promote efficiencies in their operations and improve firms' performance (Chen et al., 2023; Husain et al., 2022).

GenAI has been increasingly applied to various functions and aspects of operations in the financial industry, such as stock trading, fraud detection, stress testing and risk management, robo-advisors, financial planning, synthetic financial data generation, etc. For example, pretrained Large Language Models and transformer models are being adapted and finetuned into domain-specific models within investment banking to conduct fundamental stock analysis that integrates investor sentiments from social networks to improve portfolio performance (Lee et al., 2024).

One important characteristic of GenAI models that are being adapted in the finance industry, like banking is that they are very data-hungry models and require a large volume of data for their training and testing. This volume of data is not always available due to data privacy and security concerns, high cost of data collection, and inadequate access to the right quality data (Kannan, 2024). To overcome this obstacle, AI researchers and financial firms are exploring the possibility of creating artificial data samples, called synthetic data generation. Synthetic data is a look-alike of the real data created to mimic the statistical properties of the real data while maintaining the privacy and security of the original data (Kannan, 2024).

Synthetic data is becoming an emerging innovation in artificial intelligence(AI) and is being tested across diverse industry verticals such as health care, finance, drug discovery, manufacturing, etc. It has the potential to address some of the critical problems that AI researchers and practitioners have faced over time, such as ensuring the privacy of data being used for training AI models; dealing with unbalanced dataset in model training; and reducing overfitting and biases in the trained model.

Synthetic data generation mechanisms are capable of taking the actual data observations and learning from their statistical property to create a new dataset of high similarity without revealing the true identities of individuals to whom the transactions or original data belong, i.e data anonymisation (Shahul Hameed et al., 2024)

Some of the problems that AI researchers and scientists attempt to solve do not occur frequently and evenly. And this has created the challenge of imbalanced data, which is a condition in the dataset where one category dominates another by occurring more frequently than the other (Cheah et al., 2023). For example, a dataset of a sample of 1000 global banks collected to predict bank failures may have more surviving banks dominating the entire dataset than failed banks. This is because more banks succeed than they fail as evidenced by the work of Correia et al. (2024) who established a panel of US commercial banks from 1863 to 2023 and found that among all the banks in the panel, failed banks were in the minority and surviving banks were in the majority. In this situation, any AI model trained on this panel dataset will be predicting results that favour the dominant class, which is the surviving banks. Synthetic data could help to address this problem by generating more data in the minority class to increase the numbers of failed banks in the panel dataset and make it more balanced. Class imbalance may be addressed using several approaches, including resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic minority-class observations through interpolation between neighbouring observations and should therefore be distinguished from generative-AI approaches such as generative adversarial networks (GANs), variational autoencoders (VAEs), and emerging LLM-based synthetic-data generation methods. Future research could comparatively evaluate these approaches in the specific context of rare bank-failure events.

Bank failure prediction problem is a risk management issue in banking that suffers from the above challenges that the innovation in deep generative models for generating synthetic data can help to address. The banking system is highly regulated, so data protection laws are very strict. As a result, evolving future research in this domain to leverage synthetic data for training AI models that will be developed to serve as the foundation on which AI-powered Early Warning Systems (EWSs) will be built to forecast future bank failures will not only help banking institutions to avoid the risk of data privacy breaches, but also significantly reduce the problem of model's biases due to imbalanced data, which is common with naturally rare events such as bank defaults.

4. Limitations

This review has several limitations. First, the literature search was restricted to the Scopus database. Although Scopus provides broad multidisciplinary coverage relevant to finance, economics, decision sciences and computer science, reliance on a single database may have resulted in the omission of relevant studies indexed exclusively in other bibliographic databases. The evidence base should therefore be interpreted as a focused body of Scopus-indexed literature rather than an exhaustive representation of all research on bank-failure prediction.

Second, the search strategy deliberately focused on the bank-failure prediction construct rather than the broader literature on banking distress, fragility, insolvency, default and Early Warning Systems. Although this approach improved conceptual specificity, the relatively restrictive search terminology may have omitted relevant studies examining bank failure using alternative terminology. Accordingly, the findings should be interpreted within the defined conceptual and search boundaries of the review.

Third, study identification and eligibility assessment were conducted by a single reviewer. Independent duplicate screening was therefore not undertaken. Although the use of explicit eligibility criteria and a structured selection procedure improved consistency, single-reviewer assessment may increase the possibility of selection error or omission relative to independent screening by multiple reviewers.

Fourth, the review was restricted to English-language peer-reviewed journal articles, introducing the possibility of language and publication bias and potentially excluding relevant evidence published in other languages or publication formats.

Fifth, substantial methodological heterogeneity was observed across the included empirical studies, including differences in sample size, geographical setting, definitions of bank failure, class distributions, predictor variables, prediction horizons, evaluation metrics and validation procedures. This heterogeneity limits direct quantitative comparison of reported predictive performance and precludes strong conclusions regarding the universal superiority of one modelling approach over another.

Finally, a structured methodological quality appraisal was introduced to account for differences in study quality, considering sample/event adequacy, failure definition, validation methodology, treatment of class imbalance and out-of-sample or temporal testing. However, because the appraisal framework was adapted to the bank-failure prediction context and the included studies employed heterogeneous econometric and AI/ML methodologies, the resulting assessments should be interpreted as indicators of methodological confidence rather than definitive risk-of-bias scores. Two included publications were review studies rather than primary empirical model-development studies and were therefore excluded from the empirical methodological-quality and predictive-performance assessments. Collectively, these limitations require cautious interpretation of cross-study performance differences.

5. Conclusion

This systematic literature review synthesised a focused body of post-Global Financial Crisis research on bank-failure prediction, with particular attention to the methodological characteristics, validation practices and reported predictive performance of econometric and artificial intelligence/machine learning (AI/ML) approaches. The review identified 23 publications, comprising 21 primary empirical studies and two review studies.

Several conclusions emerge from the evidence. First, bank-level financial and accounting variables remain prominent in the empirical bank-failure prediction literature. Frequently examined dimensions include capital adequacy, asset quality, profitability, liquidity, leverage and other indicators of bank financial condition. However, predictor specifications varied substantially across studies, and broader macroeconomic and non-financial risk factors were less consistently represented. Consequently, the reviewed evidence does not support treating the predictor sets employed across studies as directly equivalent.

Second, the empirical literature demonstrates considerable methodological diversity. Conventional econometric models coexist with machine-learning, fuzzy, neural-network, support-vector-machine, tree-based, data-envelopment, survival and hybrid modelling approaches. Several studies directly compared different model families within the same empirical setting. These within-study comparisons provide stronger evidence for assessing relative predictive performance than comparisons of headline performance measures across studies using different datasets and research designs.

Third, AI/ML approaches demonstrated competitive or higher predictive performance than conventional approaches in some within-study comparisons. However, this pattern was not uniform across datasets, algorithms, failure definitions, prediction horizons, class distributions, evaluation metrics or validation designs. The evidence therefore supports the potential for incremental predictive value from AI/ML in particular settings, but does not establish the universal superiority of AI/ML, hybrid or ensemble approaches over conventional econometric models.

Fourth, substantial heterogeneity was observed in model-evaluation practices. Studies differed in their use of accuracy, sensitivity or recall, specificity, precision, ROC-AUC, model-fit measures and validation procedures. Given the rarity of bank failures, reliance on aggregate

accuracy alone may provide an incomplete assessment of predictive usefulness. Future empirical studies would benefit from reporting complementary class-specific and discrimination measures and from applying rigorous out-of-sample, temporal or external validation where feasible.

The review also highlights the importance of interpretability in the application of predictive models to banking supervision. Although complex AI/ML models may capture nonlinear relationships and interactions, predictive performance alone does not determine their suitability for supervisory use. The transparency of model outputs, validation quality, stability across time and the consequences of false-negative and false-positive classifications are also relevant when evaluating models intended to support early-warning decisions.

Several areas warrant further investigation. These include the incorporation and evaluation of broader macroeconomic and non-financial predictors, improved treatment of rare-event and class-imbalance problems, greater use of temporal and external validation, and further examination of explainable AI methods in bank-failure prediction. Generative AI and synthetic-data approaches may also represent potential avenues for future research, particularly where access to failure observations is constrained. However, these approaches were not sufficiently represented in the primary empirical studies included in this review to permit conclusions regarding their effectiveness. They should therefore be regarded as forward-looking research directions rather than as empirically established findings of the present review.

Overall, the evidence indicates that the usefulness of a bank-failure prediction model depends not solely on the modelling technique employed but on the interaction between data quality, predictor specification, failure definition, prediction horizon, class distribution, validation design and the performance criteria relevant to the intended decision context. Future research should therefore prioritise transparent and methodologically comparable evaluation frameworks that permit more reliable assessment of the incremental predictive value and supervisory usefulness of emerging modelling approaches.

CRediT authorship contribution statement

Japheth Jev: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Lukman A. Akanbi:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Muhammad Bilal:** Writing – review & editing, Supervision, Formal analysis. **Arun Sukumar:** Writing – review & editing, Supervision. **Simeon Babatunde:** Writing – review & editing, Formal analysis. **Hakeem Owolabi:** Writing – review & editing, Formal analysis. **Vasileios Kolovos:** Writing – review & editing, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Acharya, V. V., & Mora, N. (2015). A crisis of banks as liquidity providers. *The Journal of Finance*, *70*(1), 1–43.

Affes, Z., & Hentati-Kaffel, R. (2019). Forecast bankruptcy using a blend of clustering and MARS model: Case of US banks. *Annals of Operations Research*, *281*(1), 27–64.

Aikman, D., Bridges, J., Burgess, S., Galletly, R., Levina, I., O'Neill, C., & Varadi, A. (2018). *Measuring risks to UK financial stability*: *Technical report*, Bank of England.

Alkhuzaimi, F., Wilson, C. B., & Wong, W. Y. A. (2025). Key factors influencing undergraduate nursing students' perceptions of the use of learning management systems: A systematic literature review. *BMC Nursing*, *24*(1), 323.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, *28*(2), 131–171.

Andaur Navarro, C. L., Damen, J. A. A. G., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ*, *375*, n2281. http://dx.doi.org/10.1136/bmj.n2281.

Andrade, C. (2020). Sample size and its importance in research. *Indian Journal of Psychological Medicine*, *42*(1), 102–103.

Aromataris, E., & Pearson, A. (2014). The systematic review: An overview. *AJN the American Journal of Nursing*, *114*(3), 53–58.

Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in Finance: A comprehensive review through bibliometric and content analysis. *SN Business & Economics*, *4*(2), 23.

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260.

Barajas, A., Krakovich, V., & López-Iturriaga, F. J. (2023). Survival of Russian banks: How efficient are the control measures? *European Journal of Management and Business Economics*, *32*(3), 320–341.

Behbood, V., Lu, J., & Zhang, G. (2013). Fuzzy bridged refinement domain adaptation: Long-term bank failure prediction. *International Journal of Computational Intelligence and Applications*, *12*, http://dx.doi.org/10.1142/S146902681350003X.

Behbood, V., Lu, J., & Zhang, G. (2014). Fuzzy refinement domain adaptation for long term prediction in banking ecosystem. *Industrial Informatics, IEEE Transactions on*, *10*, 1637–1646. http://dx.doi.org/10.1109/TII.2012.2232935.

Behbood, V., Lu, J., Zhang, G., & Pedrycz, W. (2015). Multistep fuzzy bridged refinement domain adaptation algorithm and its application to bank failure prediction. *IEEE Transactions on Fuzzy Systems*, *23*(6), 1917–1935.

Betz, J. (2014). Emerging powers and global financial governance. *Strategic Analysis*, *38*(3), 293–306.

Billios, D., Seretidou, D., & Stavropoulos, A. (2024). The power of numerical indicators in predicting bankruptcy: A systematic review. *Journal of Risk and Financial Management*, *17*(10), 433.

Board, F. S. (2010). The basel committee on banking supervision. *Macroeconomic Assessment Group*.

Booth, A., Martyn-St James, M., Clowes, M., & Sutton, A. (2021). *Systematic approaches to a successful literature review*. SAGE Publications Ltd.

Bougault, P., El Haddad, R. A., et al. (2025). Industrial applications of AI in aircraft manufacturing: A PRISMA systematic literature review.

Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, *36*(2), 3355–3366.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Cheah, P. C. Y., Yang, Y., & Lee, B. G. (2023). Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. *International Journal of Financial Studies*, *11*(3), 110. http://dx.doi.org/10.3390/ijfs11030110.

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: The growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*, *21*(4), 471–496.

Citterio, A. (2024). Bank failure prediction models: Review and outlook. *Socio-Economic Planning Sciences*, Article 101818.

Clement, C., David, M., & Jemna, D. V. (2022). Bankruptcy prediction using machine learning–A meta-analysis. *Journal of Public Administration, Finance and Law*, *26*, 63–77.

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the eurozone: An extreme gradient boosting approach. *Journal of Business Research*, *101*, 885–896.

Correia, S. A., Luck, S., & Verner, E. (2024). *Failing Banks*: *Working Paper 32907*, National Bureau of Economic Research, http://dx.doi.org/10.3386/w32907, URL: https://www.nber.org/papers/w32907.

Cox, R. A., & Wang, G. W. Y. (2014). Predicting the US bank failure: A discriminant analysis. *Economic Analysis and Policy*, *44*(2), 202–211.

Demirgüç-Kunt, A., & Detragiache, E. (1998). The determinants of banking crises in developing and developed countries. *Staff Papers*, *45*(1), 81–109.

Dewasiri, N. J., Dharmarathna, D. G., & Choudhary, M. (2024). Leveraging artificial intelligence for enhanced risk management in banking: A systematic literature review. *Artificial Intelligence Enabled Management: An Emerging Economy Perspective*, 197.

Ecer, F. (2013). Comparing the bank failure prediction performance of neural networks and support vector machines: The Turkish case. *Economic Research-Ekonomska Istraživanja*, *26*(3), 81–98.

Ekinci, A., & Erdal, H. İ. (2017). Forecasting bank failure: Base learners, ensembles and hybrid ensembles. *Computational Economics*, *49*(4), 677–686.

El Madou, K., Marso, S., El Kharrim, M., & El Merouani, M. (2024). Evolutions in machine learning technology for financial distress prediction: A comprehensive review and comparative analysis. *Expert Systems*, *41*(2), Article e13485.

Federal Deposit Insurance Corporation (2025). *FDIC glossary*. Federal Deposit Insurance Corporation, Definition of Bank Failure: A bank is closed by its chartering authority and the FDIC is named as receiver.

Ghazalat, A., & AlHallaq, S. (2024). Predicting and assessing bankruptcy risk: The role of accounting conservatism and business strategies. *Journal of Financial Reporting and Accounting*.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing: A machine learning approach. *International Journal of Forecasting*, *34*(3), 440–455.

Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, *127*(5), 2683–2745. http://dx.doi.org/10.1007/s11192-022-04289-7.

Gusenbauer, M. (2024). Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Research Synthesis Methods*, *15*(5), 802–817. http://dx.doi.org/10.1002/jrsm.1729.

Haddaway, N., Page, M., Pritchard, C., & McGuinness, L. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and. *Open Synthesis Campbell Systematic Reviews*, *18*, 1230. http://dx.doi.org/10.1002/cl2.1230.

Heffernan, S. (2003). 14. The causes of bank failures. *Handbook of International Banking*, 366.

Huang, D. T., Chang, B., & Liu, Z. C. (2012). Bank failure prediction models: For the developing and developed countries. *Quality & Quantity*, *46*, 553–558.

Husain, A. R. A. M., Hamdan, A., & Fadhul, S. M. (2022). The impact of artificial intelligence on the banking industry performance. In *Future of organizations and work after the 4th industrial revolution: The role of artificial intelligence, big data, automation, and robotics* (pp. 145–156). Springer.

Ilk, O., Pekkurnaz, D., & Cinko, M. (2014). Modeling company failure: A longitudinal study of Turkish banks. *Optimization*, *63*(12), 1837–1849.

IMF Research Dept. (2010). *World economic outlook, april 2010: Rebalancing growth*. International Monetary Fund, http://dx.doi.org/10.5089/9781589069152.081.

Iturriaga, F. J. L., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of US commercial banks. *Expert Systems with Applications*, *42*(6), 2857–2869.

Jing, Z., & Fang, Y. (2018). Predicting US bank failures: A comparison of logit and data mining models. *Journal of Forecasting*, *37*(2), 235–256.

Jones, S. (2023). A literature survey of corporate failure prediction models. *Journal of Accounting Literature*, *45*(2), 364–405.

Kannan, N. (2024). A review of deep generative models for synthetic financial data generation. *International Journal of Financial Data Science (IJFDS)*, *1233*, 1259.

Kimmel, R. K., Thornton, J. H., Jr., & Bennett, S. E. (2016). Can statistics-based early warning systems detect problem banks before markets? *The North American Journal of Economics and Finance*, *37*, 190–216.

Kolari, J., Glennon, D., Shin, H., & Caputo, M. (2002). Predicting large US commercial bank failures. *Journal of Economics and Business*, *54*(4), 361–387.

Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., Herrera Viedma, E., et al. (2019). *Machine learning methods for systemic risk analysis in financial sectors*. Vilnius Gediminas Technical University Press.

Kristóf, T., & Virág, M. (2022). EU-27 bank failure prediction with C5. 0 decision trees and deep learning neural networks. *Research in International Business and Finance*, *61*, Article 101644.

Kuizinienė, D., Krilavičius, T., Damaševičius, R., & Maskeliūnas, R. (2022). Systematic review of financial distress identification using artificial intelligence methods. *Applied Artificial Intelligence*, *36*(1), Article 2138124.

Kumar, P., & Sharma, V. (2022). Probabilistic assessment to measure insolvency risk of Indian public sector banks for sustainable financial health. *Journal of Risk Finance*, *23*(1), 1–19.

Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the Cox proportional hazards model to bank failure. *Journal of Banking & Finance*, *10*(4), 511–531.

Lee, D. K. C., Guan, C., Yu, Y., & Ding, Q. (2024). A comprehensive review of generative AI in finance. *FinTech*, *3*(3), 460–478.

Li, Z., Feng, C., & Tang, Y. (2022). Bank efficiency and failure prediction: A nonparametric and dynamic model based on data envelopment analysis. *Annals of Operations Research*, *315*(1), 279–315.

Li, Y., Shi, Y., Huang, A., & Yang, H. (2014). Failure prediction in commercial banks with a hybrid prediction model. *Annals of Data Science*, *1*, 209–220.

Lisy, K., & Porritt, K. (2016). Narrative synthesis: Considerations and challenges. *JBI Evidence Implementation*, *14*(4), 201.

Liu, L. X., Liu, S., & Sathye, M. (2021). Predicting bank failures: A synthesis of literature and directions for future research. *Journal of Risk and Financial Management*, *14*(10), 474.

MacFarlane, A., Russell-Rose, T., & Shokraneh, F. (2022). Search strategy formulation for systematic reviews: Issues, challenges and opportunities. *Intelligent Systems with Applications*, *15*, Article 200091.

Manthoulis, G., Doumpos, M., Zopounidis, C., & Galariotis, E. (2020). An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for US banks. *European Journal of Operational Research*, *282*(2), 786–801.

Mare, D. S. (2015). Contribution of macroeconomic factors to the prediction of small bank failures. *Journal of International Financial Markets, Institutions and Money*, *39*, 25–39.

Menicucci, E., & Paolucci, G. (2023). ESG dimensions and bank performance: An empirical investigation in Italy. *Corporate Governance: The International Journal of Business in Society*, *23*(3), 563–586.

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., et al. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A PRISMA extension. *Biological Reviews*, *96*(5), 1695–1722.

Özel, G. (2013). Probabilistic prediction of bank failures with financial ratios: An empirical study on Turkish banks. *Pakistan Journal of Statistics and Operation Research*, [ISSN: 1816-2711] *9*(4), 407–426. http://dx.doi.org/10.18187/pjsor.v9i4.492.

Pham, X. T., & Ho, T. H. (2021). Using boosting algorithms to predict bank failure: An untold story. *International Review of Economics & Finance*, *76*, 40–54.

Quinteros-Navarro, D., & Rodriguez, C. (2024). Prediction model applying machine learning to forecast the bankruptcy of companies. A systematic review of the literature. *Journal of Theoretical and Applied Information Technology*, *102*(15).

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA statement. *Systematic Reviews*, *10*(1), 117.

Satheesh, M., & Nagaraj, S. (2021). Applications of artificial intelligence on customer experience and service quality of the banking sector. *International Management Review*, *17*(1), 9–86.

Shahul Hameed, M. A., Qureshi, A. M., & Kaushik, A. (2024). Bias mitigation via synthetic data generation: A review. *Electronics*, *13*(19), 3909. http://dx.doi.org/10.3390/electronics13193909.

Shakeel, M. R., Siddiqui, T. A., & Alam, S. (2023). Feature selection in corporate bankruptcy prediction using ml techniques: A systematic literature review. *Advances in Signal Processing, Embedded Systems and IoT*, 345–363.

Tarawneh, A., Abdul-Rahman, A., Mohd Amin, S. I., & Ghazali, M. F. (2024). A systematic review of fintech and banking profitability. *International Journal of Financial Studies*, *12*(1), 3.

Wang, D., Quek, C., & Ng, G. S. (2016). Bank failure prediction using an accurate and interpretable neural fuzzy inference system. *AI Communications*, *29*(4), 477–495.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. http://dx.doi.org/10.7326/M18-1376.

Wu, Q., & Cole, R. A. (2024). Macroeconomic conditions and bank failure. *Journal of Forecasting*, *43*(5), 1212–1234.

Xia, Y., Xiong, A., Zhang, Z., Zou, Q., & Cui, F. (2025). A comprehensive review of deep learning-based approaches for drug–drug interaction prediction. *Briefings in Functional Genomics*, *24*, elae052.

Zaghdoudi, T. (2013). Bank failure prediction with logistic regression. *International Journal of Economics and Financial Issues*, *3*(2), 537–543.

Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., Xu, S., Dai, H., Zhao, L., Mai, G., et al. (2024). Revolutionizing finance with llms: An overview of applications and insights. arXiv preprint arXiv:2401.11641.

Zhen-Jia-Liu (2015). Cross-country study on the determinants of bank financial distress. *Revista de Administração de Empresas*, *55*(5), 593–603. http://dx.doi.org/10.1590/S0034-759020150510.

Zöller, M. A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, *70*, 409–472.