North American Journal of Economics and Finance 69 (2024) 102018

Contents lists available at ScienceDirect

North American Journal of Economics and Finance

journal homepage: www.elsevier.com/locate/najef

Constructing early warning indicators for banks using machine learning models

Coskun Tarkocin [a,b,1,*], Murat Donduran [a,2]

[a] Yildiz Technical University, Istanbul, Turkey
[b] King's College London, London, United Kingdom

ARTICLE INFO

JEL Codes:
C51
C88
G21

Keywords:
Early warning indicators
Financial stress
Machine learning
Ensemble model
Liquidity risk
Crisis management
COVID-19 crisis

ABSTRACT

This research contributes to bank liquidity risk management by employing supervised machine learning models to provide banks with early warnings of liquidity stress using market-based indicators. Identifying increasing levels of stress as early as possible provides management with a crucial window of time in which to assess and develop a potential response. This study uses publicly available data from 2007 to 2021, covering two severe stress periods: the 2007–2008 global financial crisis and the COVID-19 crisis. The current version of the developed model then applies backtesting using the data from the COVID-19 crisis. The findings of this study show that the ensemble model with the RUSBoost algorithm predicts "red" and "amber" days with a success rate 21% greater than the average of other machine learning models; thus, it can greatly contribute to bank risk management.

1. Introduction

Bank risk management has become more complicated with the evolving regulatory framework following the 2007–2008 global financial crisis (GFC). With increased regulation and advancements in risk management practices, it is expected that banks and the wider financial system will become more resilient to shocks. However, any new crisis may unfold uniquely, which will require management to be on alert if the market stress level changes unexpectedly.

Banks have adapted to the current relatively complex environment through substantial costly investments. Extensive transformation projects have been implemented for systems, reporting, modelling and governance. Thus, the granularity of the risk data produced and reported has increased significantly. However, producing and reporting a significant amount of information does not necessarily provide all the answers. Several questions remain on an ongoing basis: what are the stress levels in the market? what are the perceptions of clients and counterparties regarding a specific bank's risk level? in the very short term, is there any emerging stress? These questions must be answered to define an alert level for any development in a stressed market which may jeopardise banks' survival. Therefore, banks regularly monitor the market and internal bank indicators.

It is standard practice to monitor several indicators with defined thresholds to inform management of the stress levels of the market.

* Corresponding author.
E-mail addresses: coskun.tarkocin@kcl.ac.uk, ctarkocin@hotmail.com (C. Tarkocin), donduran@yildiz.edu.tr (M. Donduran).
[1] Head of Liquidity Optimisation, HSBC Group, London, United Kingdom and Visiting Research Fellow at Qatar Centre for Global Banking & Finance, King's Business School, King's College London.
[2] Director of Graduate School of Social Sciences, Yildiz Technical University, Turkey.

https://doi.org/10.1016/j.najef.2023.102018
Received 29 June 2022; Received in revised form 6 July 2023; Accepted 30 September 2023
Available online 2 October 2023
1062-9408/© 2023 Elsevier Inc. All rights reserved.

However, this study proposes a novel method of analysing these indicators by employing supervised machine learning models and transforming the indicators into a classification problem without creating another stress index. This will provide practitioners with a method to transform multiple indicators into a simple system that can facilitate swift decision-making in times of stress.

There are three main motivations behind this study. First, early warning indicator (EWI) research in the literature often focuses on the policymaker's perspective, not the perspective of a bank as an individual agent in the financial system (see Yoshitomi and Shirai, 2000; BCBS, 2008; BCBS, 2013; Surjaningsih, Yumanita and Deriantino, 2014; Drudi and Nobili, 2021; Guerra, Castelli and Côrte-Real, 2022). Policymakers prioritise wider financial system health and cost to taxpayers while individual banks prioritise their own survival and behave accordingly. Therefore, presenting the construction of EWIs from an individual bank's perspective, which can include high-frequency data such as daily liquidity positions and very short-term forecasts, will be a novel approach compared to those taken by studies in the extant literature. Second, research using machine learning models to detect financial stress levels is still in its infancy; therefore, the addition of the ensemble classifier with a random under-sampling algorithm, as applied by this study, will provide a unique contribution to the field. Last, this study focuses on the immediate nature of EWIs based on publicly available daily data, easily linking real-time market data with an institution's internal EWIs. This is particularly important in the selection of variables since policymakers and institutions have access to data with different frequencies and granularities. In addition, the flexibility of the proposed supervised machine learning model makes its use possible in different markets and by different stakeholders such as investors, regulators, or central banks.

The remainder of this study is organised as follows: Section 2 reviews the relevant extant literature on EWIs; Section 3 provides definitions of, and a framework for, EWIs. Section 4 outlines the data selected for this study and summarises the data transformation process. Section 5 describes the methodology of this study, and Section 6 discusses the results. Section 7 summarises the findings of the empirical analysis and discusses their policy implications.

2. Literature review

The literature around EWIs is extensive, since understanding emerging stress as early as possible – though not so early that it may increase the cost of policies implemented to prevent or reduce the impact of stress – is vital. The main motivation behind EWIs has been to detect stress signals early and then provide governments, central banks and regulators sufficient time to prevent or reduce the impact of the emerging crisis (BCBS, 2013).

Literature on EWIs in banking can be grouped into two categories. Literature in the first category aims to provide a warning for financial or banking crises. Literature in the second category aims to provide early warning of individual bank failure, which may negatively impact the wider financial system (Gaytán and Johnson, 2002). Models predicting distress in banks from a policymaker perspective help to better understand underlying vulnerabilities and identify patterns preceding financial stress (Betz et al., 2013). This study will bridge these two categories of literature by focusing on a specific financial crisis from the perspective of an individual bank, and detecting market signals, which can be incorporated into the bank's internal indicators for decision-making rather than policy-making. Therefore, selected literature in this area will be discussed.

Individual banks use EWIs to predict banking stress and any potential adverse event impacting their risk positions which may increase their probability of failure. Combining market and internal bank-specific indicators will provide an ideal set of indicators to estimate these EWIs. The model presented in this study is based on daily data, making it useful to incorporate into day-to-day management. Beutel et al. (2018) researched whether machine learning methods are better for prediction compared to traditional models and found that traditional models perform better. However, their study aimed to predict financial stress for 5 to 12 quarters ahead. In this study, we aim to predict very short-term stress where individual banks can take action.

Demirguc-Kunt and Detragiache (1998) studied determinants of banking crisis using multivariate logit models for a large panel of countries and found that banking crises are more likely in a weak macroeconomic environment. Navajas and Aaron (2013) used a simple multivariate logit model on financial soundness indicators and macroeconomic control variables to investigate a correlation between the occurrence of banking crises and other variables. While capital adequacy ratio and return on equity have a negative correlation, analysis found that lagged return on equity may be a leading indicator for a banking crisis.

Imbierowicz and Rauch (2014) investigated the relationship between liquidity and credit risks for a commercial bank in the United States and found that the interaction of both risk types has a significant impact on bank default probability. Galletta, Mazzù and Scannella (2021) investigated the relationship of bank governance characteristics to liquidity risk for the period after the global financial crisis (2011–2017) and showed that liquidity risk was reduced when a certain threshold of risk supervisors was defined. Arif and Anees (2012) examined liquidity risk in Pakistani banks to assess its impact on bank' profitability using multiple regressions for datafrom 2004 to 2009 and showed that liquidity risk affects bank profitability significantly.

Goodhart (2008) referred to liquidity and solvency as heavenly twins of banking and stated that an illiquid bank can rapidly become insolvent and an insolvent bank illiquid. Goodhart, also tried to determine best distribution of responsibility for liquidity management between commercial banks and a central bank. Chen et al. (2018) investigated bank liquidity risk and performance for 12 advanced countries from1994 to 2006, and their results showed that liquidity risk is an endogenous factor of bank performance. They further reported that components of liquid assets, dependence on external funding, supervisory and regulatory factors and macroeconomic factors are determinants of liquidity risk.

History has shown there has been an increased interest in understanding the drivers of banking crises and, consequently, proposing policy recommendations following each major crisis period. Yoshitomi and Shirai (2000) conducted a critical and comprehensive review of research following the 1997 Asian crisis, making policy recommendations to prevent another one. Gaytán and Johnson (2002) reviewed the literature on early warning systems for banking crises and classified studies by methodological approaches.

Regulatory guidance before the global financial crisis does not include liquidity EWIs. The 1992 and 2000 framework guidance for the managing liquidity by the Basel Committee do not refer any EWIs to be monitored by individual banks (BCBS, 1992; BCBS, 2000). However, BCBS (2008) – "Principles for Sound Liquidity Risk Management and Supervision" – included a recommendation for individual banks to have EWIs. BCBS (2008) highlights the importance of designing a set of indicators to identify increased risk or vulnerabilities. Although BCBS (2008) did not suggest composite early warning indicators, Aikman et al. (2018) suggested composite measures can be simple communication tools for macroprudential policymakers and the public.

Literature around early warning indicators peaked after the 2007–2008 Global Financial Crisis. There was a similar increase in research interest in this area following the financial crisis in Asia (Bräuning et al., 2019). For instance, Surjaningsih, Yumanita and Deriantino (2014) developed EWIs for banking liquidity risk and could detect liquidity imbalances one year before October 2008, reporting 67 % predictive power.

Iachini and Nobili (2014) introduced a coincident indicator of systemic liquidity risk in the Italian financial markets using standard portfolio theory. Individual raw indicators from the equity and corporate markets, Italian government bond market and money market were used. The calculated indicator was then compared to the results of a survey of the most liquidity stressful events for the Italian financial markets, which showed the systemic liquidity risk indicator accurately identified high systemic risk events.

Eross et al. (2015) employed an autoregressive Markov regime-switching model and showed that the US LIBOR–OIS spread can be used to predict when a liquidity crisis may occur within the interbank market. Acharya et al. (2017) developed a model-based measure of systemic risk and, with a variety of tests, showed that market data from equity and credit default swaps (CDS) were able to predict financial firms' largest contributors to systemic risks. Aldasoro, Borio and Drehmann (2018) assessed household and international debt as EWIs for banking distress, and found that these indicators provide useful information when modelling rare events such as crises. In addition, they found that combining information from credit and asset markets into composite indicators can improve performance, EWIs are designed to capture vulnerabilities linked to the financial cycle, and other vulnerabilities are not considered. In contrast, we consider the short-term nature of the liquidity crisis and, more specifically market variables in this study.

Lang and Sarlin (2018) proposed a general-purpose framework to build early warning models and applied it to predict bank distress in Europe. This study also provided a flexible modelling solution, combining the loss function approach to evaluate models with regularised logistic regression and cross validation, which can be used to analyse emerging vulnerabilities at the micro and macro level. Aikman et al. (2018) assessed build-ups of risk in the wider financial system by using composite measures and showed how they influence downside risks to economic growth.

Padhan and KPP (2019) documented the history of early warning models predicting the financial crisis and also proposed a new agenda to augment existing models. Models were listed under six different types and, using machine learning, family artificial neural networks and genetic algorithms were reported. In a more recent study employing a machine learning model to predict bank distress, Bräuning et al. (2019) employed a decision tree model with the Quinlan C5.0 algorithm to identify individual bank distress for small European banks. The study reported low Type I and Type II errors, with the decision tree model outperforming the logit model. For the review of theoretical algorithms for multivariate portfolio optimisation algorithms under illiquid market conditions and how to integrate asset liquidity risk into an LVAR model, see Al Janabi (2021).

Drudi and Nobili (2021) developed an early warning system to identify Italian banks facing liquidity difficulties. For modelling purposes, they used the Bank of Italy's confidential data from 2014 to 2020. Using central bank data superiority, they modelled to forecast a 3-month horizon. Guerra, Castelli and Côrte-Real (2022) used machine learning models to classify Portuguese banks using supervisory data from 2014 to 2021. They showed that machine learning models outperform traditional statistical methods in the risk classification of a bank from a supervisory lens.

3. EWI definitions and framework

In this study, EWIs are defined as any data or information used to predict a potential stress event. These can be numerical or categorical data. Market-based indicators reflect the health of the economy, bank, and firm, and can predict changes in financial conditions (Kliesen & McCracken, 2020).

Examples of EWIs include, but are not limited to, the following:

- Market EWIs: These are mainly external data, but not specific to the individual bank, such as equity prices, interest rates, spreads, commodity prices, macroeconomic variables, credit ratings (other banks), futures, foreign exchange rates, policy rates, stock indices, volatility indices, and secured funding spreads.

Internal EWIs: These are bank-specific measures including capital position, deposit outflows, maturity mismatch, stress test results, bank own credit rating and CDS spread, credit loss measures, liquidity and funding metrics, market sentiment, share price, funding spread compared to peers, concentration metrics, negative publicity, and currency mismatches. Most internal bank indicators are not available publicly on a daily basis. For example, measurements such as the liquidity coverage ratio, available high-quality liquid assets, or unexpected deposit outflows are important indicators of a bank's liquidity. However, these data are not available to model in this study. Individual banks can use the approach presented in this study by incorporating their internal data.

A comprehensive list of market and internal EWIs can be seen in Venkat and Baird (2016, Chapter 6, authored by Bruce Choy and Girish Adake) and the BCBS (2008) suggested EWIs for liquidity risk.

Table 1 summarises data asymmetry in the modelling literature on EWIs. Data asymmetry is important since it defines which EWIs can be used in the modelling depending on the stakeholders' perspective. For instance, individual banks cannot build models which

include other banks' daily risk data, which is only available to the other banks with production frequency and to the regulator with reporting frequency.

The proposed specific modelling and data experiment both focus on market EWIs; however, individual banks can employ the proposed ensemble classifier random under-sampling algorithm to generate the RAG status based on internally available daily data only. This provides a comprehensive view of different stress types, detecting vulnerabilities that may not be visible based on public information.

RAG status is a common framework for risk measurement and monitoring. A green indicator suggests a normal market environment, an amber indicator denotes elevated stress levels which may require further investigation of underlying risks, and a red indicator highlights major stress levels in the market, demonstrating the need for immediate management attention and a potential response.

Back-testing of the thresholds for green, amber and red statuses is of critical importance to ensure management receives warning light indicators when most relevant. Back-testing helps to determine if recalibration is needed. For example, when a measure frequently exceeds the threshold, it will report false alarms (Venkat & Baird, 2016, Chapter 6, authored by Bruce Choy and Girish Adake). A recent example showing the importance of back-testing is the St. Louis Fed Financial Stress Index (STLFSI2), which was revised from STLFSI1 as that earlier version was not able to detect market movements around August 2011 or, most importantly, the turmoil triggered by the COVID-19 pandemic around February/March 2020 (Kliesen & McCracken, 2020). Models should aim to have a limited number of false alarms, and a recalibration and review of the EWIs should be performed following a major stress period.

The MERIT framework (Table 2) provides criteria for the successful implementation of EWIs for effective management response. EWI models should be completed with a proper escalation and reporting process (Venkat and Baird, 2016, Chapter 6, authored by Bruce Choy and Girish Adake).

4. Data

Publicly available data from 2007 to 2021 (3537 days) were used to train the machine learning models in this study. Historical data for potential EWIs for the same period were sourced from Federal Reserve Economic Data (FRED) and Yahoo Finance for different markets such as Equity, FX, interest rates, and spreads.

The list of market data used informs of wider market stress which is relevant to both commercial and investment banks. Data such as the Nasdaq100, TED Spread, and CPFF are more specific to the US market than they are to other developed or developing countries. Therefore, our model is more relevant to US banks. One limitation of this study is that it is US-centric. Therefore, to develop it for other markets such as developing countries, future research needs to be conducted.

4.1. Data transformation and cleaning

The St. Louis Fed Financial Stress Index, which began in late 1993, was used to define the level of stress in the market and train the machine learning model. In this index, zero represents normal market conditions. When the value drops below zero it means below-average financial market stress, whilst above zero is interpreted as above-average market stress (Federal Reserve Bank of St. Louis, 2021).

In this study, historical distribution of the STLFSI2 was reviewed to make assumptions regarding the Green, Amber, and Red days. Interrogation of the historical data shows the largest group is around 85 % of the days; therefore, it was assumed this represented normal days (Green). The second largest cluster, around 10 % of the days, was assumed to be moderate stress days (Amber), and significantly outlying index days, around 5 %, were taken to represent severe stress days (Red). This assumption intuitively made sense since the number of severe stress days would be expected to be significantly less than the number of normal days, which were coded with a Green status.

One limitation of the St. Louis Fed Financial Stress Index is that it provides weekly rather than daily data; therefore, the same RAG status was assumed for the whole week. Although it was recognised that some days may have been a different level, this assumption likely had a limited impact on the main aim of this study.

With a few exceptions, no issues identified within the historical data. There were two minor issues addressed during preparation of the data:

- On 20 and 21 April 2021, crude oil futures prices fell below zero dollars, which resulted in a substantial outlier for the range for these two days compared to the previous day range used.

Table 1
Information Asymmetry Between Stakeholders.

	Investors	Bank Management	Policymakers (Regulator, Central Bank)
Market Data	All publicly available data (including data retrieved from vendors such as Bloomberg and Reuters)	All publicly available data (including data retrieved from vendors such as Bloomberg and Reuters)	All publicly available data (including data retrieved from vendors such as Bloomberg and Reuters)
Internal Data (Bank-specific)	Public data only with disclosed frequency	Internal bank data with full granularity and maximum frequency	Public data and all banks' reported data

Table 2
The MERIT Framework.

	Definition
Measures	Measures are the essential building block of a robust EWI framework.
Escalation	An appropriate escalation framework is essential to ensure that the EWI framework is embedded in a management information system. Overall status changes to Amber or Red should be linked to internal escalation frameworks for crisis management.
Reporting	Timely reporting is required to ensure that escalation can happen as required. Best practice is to have EWIs daily monitored.
Integrated Systems	Appropriate systems and data are required to ensure that the reporting can be conducted on a timely basis.
Thresholds	Measures must have relevant and appropriately calibrated thresholds applied for the escalation process to be effective (for this, our model will automatically calculate status but effective back-testing is critical).

- CPFF (3-Month Commercial Paper Minus Federal Funds Rate) was not available for 78 days during the 2008 financial crisis and the COVID-19 crisis. For those days, the elevated spread level from the last available date was used.

The raw data referred to in Table 3 was transformed as shown in Table 4 for use in the modelling. A seven-day rolling standard deviation was calculated to capture the increasing volatility in the market. The high-low range was calculated to capture the intra-day variation for the variables referred to in Table 3. The TED spread, defined as the difference between the 3-month US Treasury Bill and the 3-month USD Libor, was used as a movement in the model. In order to capture daily movement of the variables in Table 3, the logarithmic return was calculated.

Visual inspection shows the STLFSI2 captures two recession periods and major financial crises. Fig. 1 shows that STLFSI2 increases significantly during the 2008–2009 global financial crisis and the COVID-19 crisis around February–March 2020. This supports the hypothesis that the model can be used as a benchmark to train a machine learning model for RAG status estimation. One limitation is that STLFSI2 is published weekly; therefore, day t features were used to predict the $t + 1$ index level.

5. Methodology

A subset of artificial intelligence, machine learning models were employed to conduct data experiments in this study. The popularity of machine learning models in the banking and financial sector has increased in recent years mainly due to the increasing availability of data, computing power, and improved software (BOE, 2019).

Machine learning models can be grouped into three main categories: supervised, unsupervised, and reinforcement learning. Supervised learning models train data based on a given input and output. In contrast, unsupervised learning models analyse data without a given output, and find potential relationships through clustering data. In reinforcement learning models, the aim is to maximise the defined reward for a specific action (McKinsey&Company, 2021). In addition, FSB (2017) identifies deep learning models under the machine learning umbrella and defines these as algorithms that work with layers similar to how the human brain functions.

To the best of our knowledge, this study will be the first in the literature to use ensemble classifiers with a random under-sampling algorithm to construct early warning indicators. EWIs produced by the model may trigger a management response, which may have cost implications. For this reason, the explainability of the model and, where necessary, overlaying the model with expert judgment, is of utmost importance. In the following section, the selected classifier and algorithm are briefly discussed. Machine learning models tend to have an overfitting problem if cross validation is not performed. To prevent overfitting in this study, K-fold cross validation was used, with $K = 5$.

5.1. Ensemble classifiers

Ensemble classifiers will be used in this study because the underlying data is materially imbalanced. Ensemble learning aims to combine predictions from several base estimators to build a more robust single estimator. Two families of ensemble methods are widely used. Averaging methods, such as bagging methods and forests of randomised trees, use models that independently make predictions and then average those predictions. The second group – boosting methods – builds base estimators sequentially to achieve a combined estimator with reduced bias. Some examples of boosting methods are RUSBoost, AdaBoost, and Gradient Tree Boosting (scikit-learn, 2020).

The RUSBoost algorithm was first introduced by Seiffert et al. (2008) to reduce class imbalance problems in the data set. RUSBoost

Table 3
List of Raw Data and Transformation.

Group	Transformed for Model	Source Data
Volatility	Logarithmic return, 7 days standard deviation, high-low range to low	CBOE Volatility Index (^VIX)
Commodity	Logarithmic return, 7 days standard deviation, high-low range to low	Crude Oil Mar 21 (CL = F)
Futures	Logarithmic return, high-low range to low	Nasdaq 100 Mar 21 (NQ = F)
Equity	Logarithmic return, 7 days standard deviation, high-low range to low	EURONEXT 100 (^N100)
Spread	Movement between two dates	TED Spread
Spread	Actual spread as %, 7 days standard deviation	CPFF: 3-Month Commercial Paper Minus Federal Funds Rate

Table 4
Measures used in tranforming raw data for modelling.

Measure	Formula
High-low range	$\text{HLR} = (\frac{X_{High} - X_{Low}}{X_{Low}})$
TED Spread change	$\text{TED\_Change} = (TED_X - TED_{t-1})$
Logarithmic return	$\text{R\_ln} = \log(\frac{X_t}{X_{t-1}})$
Seven days rolling standard deviation	$\text{7d\_Stdev} = \text{Stdev}(X_{t-7} to X_t)$

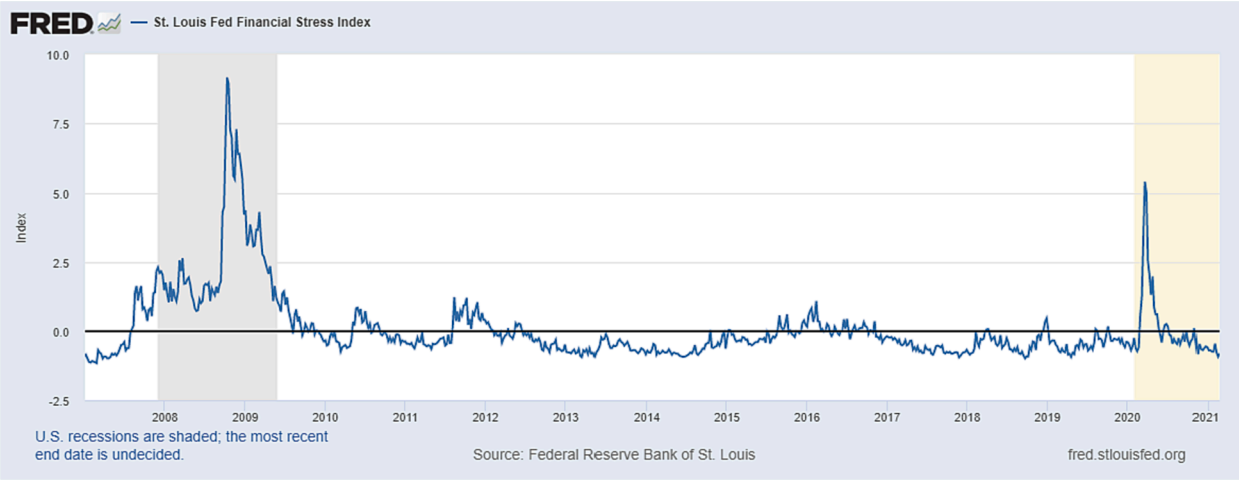

Fig. 1. St. Louis Fed Financial Stress Index (STLFSI2). Source: Federal Reserve Bank of St. Louis.

uses random data sampling with boosting, which improves the classification performance of the training data. Financial stress or distress bank classification problems have imbalanced data, wherein one class has far fewer members than others. For this reason, the RUSBoost algorithm is used for the modelling in this study. For a comprehensive overview of the RUSBoost algorithm, see Seiffert et al. (2010).

The 2019a version of MATLAB, used for implementation of the models, has eleven ensemble learning algorithms. Of the eleven methods, Random Undersampling Boosting (RUSBoost method) is a better fit for imbalanced data and can be used for binary and multiclass classification. MATLAB's random under-sampling algorithm takes the same number of observations from each class of data, which is the same as the number of the minority class. Following sampling, adaptive boosting is applied to construct the ensemble. For full details see MATLAB documentation under Ensemble Algorithms (MATLAB, 2019).

The boosting procedure for RUSBoost applies adaptive boosting for multiclass classification in calibrating weights and constructing ensembles. Adaptive boosting for multiclass classification in MATLAB uses weighted pseudo-loss for N observations and K classes. Calculated pseudo-loss is used as a measure of classification accuracy (MATLAB, 2019).

$$\varepsilon_t = \frac{1}{2}\sum_{n=1}^{N}\sum\nolimits_{k\neq y_n}^{K} d_{n,k}^{t}(1-h_t(x_n,y_n)+(h_t(x_n,k)),$$

- Each step represented by *t; k* represents class; *N* represents number of observations.
- x_n is a vector of predictor values for observation *n*.
- y_n represents the true class value taking one of the *K* values.
- h_t represents the prediction of the learner for each step *t*.
- $h_t(x_n, k)$ is the confidence of learner prediction at step *t,* class *k* ranges from zero to one.
- $d_{n,k}^t$ represents observation weights of class *k* in step *t*.

5.2. Performance evaluation metrics and definitions

Confusion Matrix

The confusion matrix is used to evaluate the performances of the models in the data experiments. Using data presented in the confusion matrix, several performance measures are calculated.

Table 5 summarises the information presented in the confusion matrix. As shown in the table, $G + A + R = N$, the total number of

observations.

Formulas for the measures above are outlined below:

- **Accuracy (Acc)** $= \frac{TG+TA+TR}{N}$. This metric measures the percentage of true class predictions in all observations.
- **Error (Err)** $= 1 - Accuracy$. This metric provides the misclassification percentage.

Since Accuracy and Error evaluations will result in the same conclusion, only accuracy measures will be presented.
In this study, we will extend two measures that can be used specifically for predicting the RAG status of EWIs:

- **Warning Score (WS)** $= \frac{TA+FR2+FA2+TR}{A+R}$. The rationale behind this score is that an institution would like to be warned of elevated stress levels. Once a Warning Score is received in combination with another internal measure, the final management status can be decided. This measure will ignore misclassification between Red and Amber days.
- **Weighted Warning Score (WWS)** $= \frac{TA+w_1\times FR2+w_2\times FA2+TR}{A+R}$. This is a slightly revised version of the Warning Score. It applies weight less than 1 to FR2 and FA2 for misclassification. In this study 0.5 was used as a weight.
- **True Red Accuracy (TRA)** $= \frac{TR}{R}$. This measure is to control the accuracy of red status days since red days are most impacted by the imbalanced data. For instance, if overall AUC is very high but TRA is below 50 %, the model may not be fit for EWI usage.

ROC Curves

Another method used for analysing a model's results performance is the receiver operating characteristics (ROC) curve. ROC curves represent true positive rates versus false positive rates to show the trade-off between a correctly identified and a false prediction. The area under the curve (AUC) is used to compare the models' classification performances (Seiffert et al., 2010).

Precision and Recall

Two additional measures to evaluate the performance of the models' results are the precision and the recall. They are calculated as follows:

- **Precision (PR)** $= \frac{TG}{G^*}$. This measures the proportion of true green in all predicted greens, Using the same logic, it calculates for amber and red as follows: $\frac{TA}{A^*}$ and $\frac{TR}{R^*}$.
- **Recall (RC)** $= \frac{TG}{G}$. This measure calculates correctly classified true greens in all actual greens. Using the same logic, it calculates for amber and red as follow: $\frac{TA}{A}$ and $\frac{TR}{R}$.

6. Empirical results

Twenty-four classification models under six model types were run to predict the RAG status of each day. Six performance evaluation metrics, presented in Table 6, compare the performances of the models in their prediction of RAG status of the EWIs. For each model type, the results for the model with highest accuracy are presented in bold for comparison purposes.

When using Accuracy (or, inversely, Error) as a measure, 'Ensemble – Boosted Trees' showed the highest predictive power. However, when the focus was on predicting only red days or both red/amber days, 'Ensemble – RUSBoost' clearly performed better. The Warning Score for the Ensemble- RUSBoost returned 95 % accuracy, compared to an average of just 74 % for the other best performing models. This trend was supported by the TRA measure.

K-fold cross validation was performed to prevent overfitting. Without cross validation, in-sample accuracy would be very high, but at the cost of sacrificing the ability to make predictions beyond the sample.

Ensemble – Boosted Trees perfectly predicts normal days based on Acc and AUC measures. However, with imbalanced data, predicting normal days is not the aim. For distress prediction, WS, WWS, and TRA are the proposed measures for determining model selection. 'Ensemble – RUSBoost' showed the best performance for this task. Therefore, the results for this model will be examined more closely.

Fig. 2 shows the confusion matrix for the best performing model (Ensemble – RUSBoost) showing a true positive rate of 83 % for red

Table 5
Confusion Matrix.

		RAG Status Predicted Class			
		Green	**Amber**	**Red**	**Total**
RAG Status Actual Class	**Green**	True Green (TG)	False Amber_1 (FA1)	False Red_1 (FR1)	*G*
	Amber	False Green_1 (FG1)	True Amber (TA)	False Red_2 (FR2)	*A*
	Red	False Green_2 (FG2)	False Amber_2 (FA2)	True Red (TR)	*R*
	Total	*G**	*A**	*R**	***N***

Table 6
Comparison of Models.

MATLAB Model Name	Acc	WS	WWS	TRA	AUC
Ensemble – RUSBoost	88 %	**95 %**	**89 %**	**83 %**	0.98
Average of models below	91 %	74 %	69 %	70 %	0.97
Tree – Medium Tree	92 %	73 %	68 %	77 %	0.97
Quadratic Discriminant	89 %	74 %	65 %	60 %	0.96
Naïve Bayes	90 %	85 %	77 %	73 %	0.94
Support Vector Machine – Quadratic	93 %	79 %	73 %	71 %	0.96
k-Nearest Neighbour (weighted KNN)	92 %	58 %	54 %	62 %	0.98
Ensemble – Boosted Trees	**94 %**	78 %	73 %	78 %	**0.99**

status; 84 % for amber status, and 89 % for the normal days with green status. Improved red and amber day prediction came at the cost of missing some of the normal days; since it is preferable to miss green days rather than the higher-risk red and amber days, this is acceptable. Significantly, the 'Ensemble – RUSBoost' model did not misclassify any red status days as green. This means a warning will be produced only when stress levels are elevated. Achieving 95 % WS, 89 % WWS, and 83 % TRA supports the hypothesis that the trained model can be automated to predict daily RAG statuses.

The ROC curve corroborates the observation that 'Ensemble – RUSBoost' is the best performing model. A high number in the AUC suggests a classifier is performing well. Perfect results, where no misclassification occurs, appear in the top left corner of the plot; random results appear along a diagonal line at 45 degrees across the plot, and poor results appear below this line, towards the bottom right corner. Fig. 3 shows the ROC curve for 'Ensemble Model – RUSBoost', where red days are true positives. It also shows that if a classifier's TPR is increased it will compromise false positive rates.

Predicting a high number of red days from actual red days should be the priority. Following this for the amber days, in this case of rare events, the recall measure becomes crucial. Thus, a high recall for red and amber days is achieved with Ensemble- RUSBoost algorithm. Table 7 shows that the recalls for amber (87 %) and red (83 %) are significantly higher than the other best-performing models; averages.

Based on the above empirical results, it can be concluded that the predictive performance of the Ensemble Model with the RUSBoost algorithm is appropriate for EWI RAG status prediction. Future work may involve employing more features and different model combinations, as well as performing hyperparameter tuning to further optimise a model's results before deployment.

7. Conclusions

This paper employed supervised machine learning models to predict EWI RAG statuses for the market across 3537 days between January 2007 and January 2021. Data used for training the model covered two major stress periods: the 2007–2008 global financial crisis and the COVID-19 market crisis in 2020. The St. Louis Fed Financial Stress Index (STLFSI2) was used to create three types of

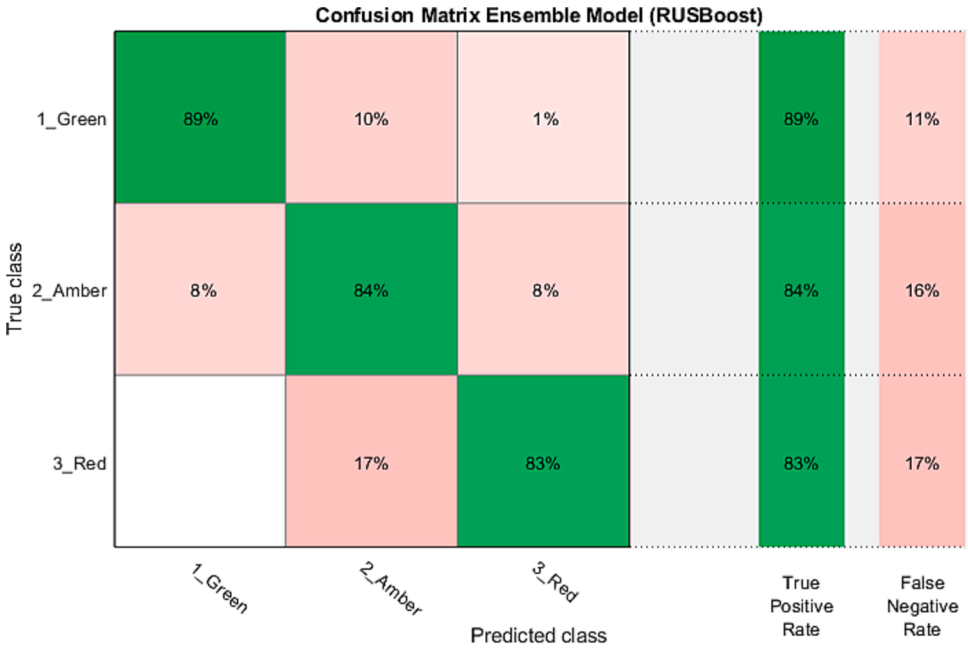

Fig. 2. Confusion Matrix: Ensemble Model – RUSBoost.

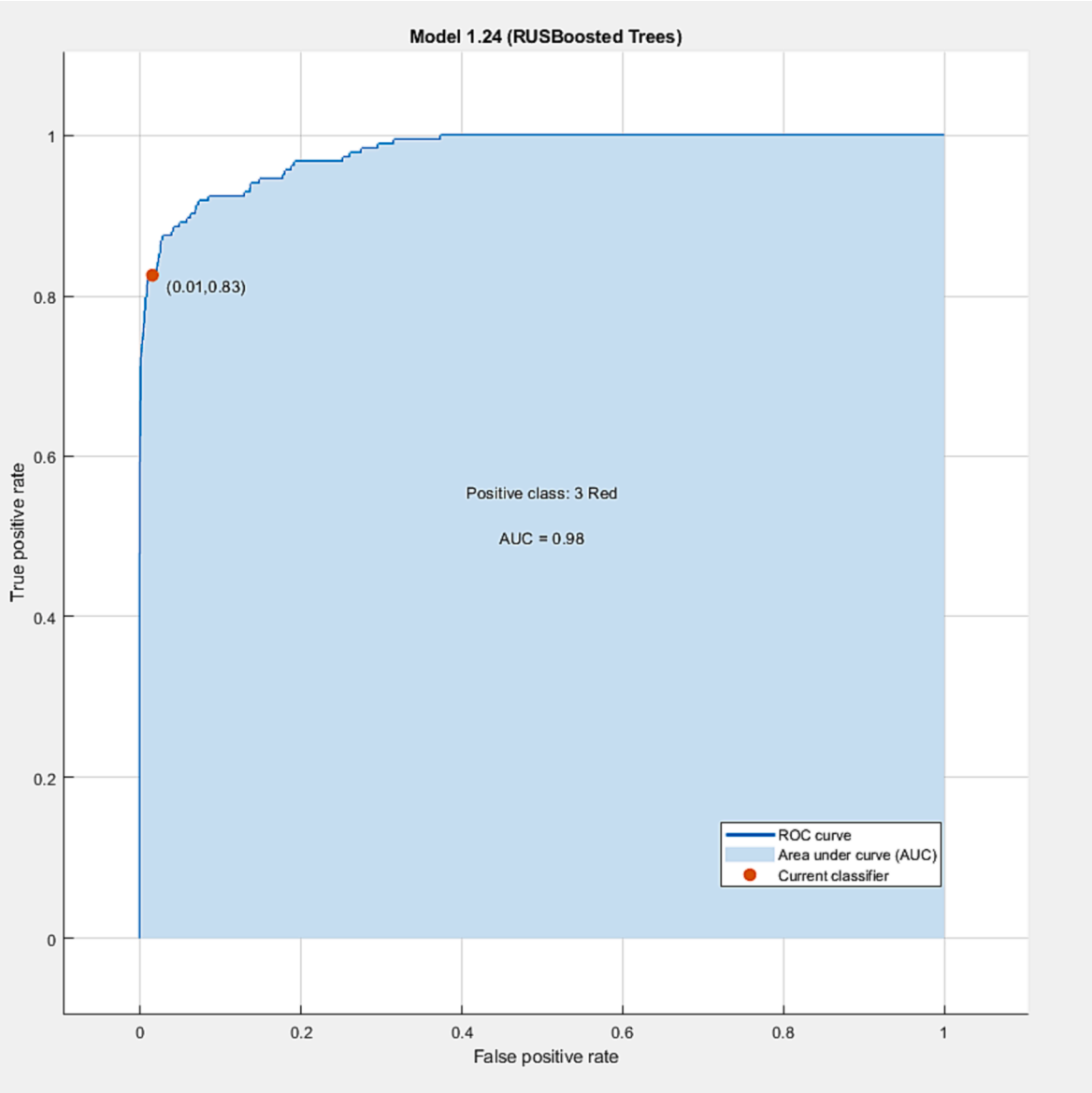

Fig. 3. ROC Curve for 'Ensemble – RUSBoost'.

Table 7
Comparison of Models Using Precision-Recall Measures.

	Green		Amber		Red	
MATLAB Model Name	Precision	Recall	Precision	Recall	Precision	Recall
Ensemble - RUSBoost	99 %	89 %	49 %	84 %	75 %	83 %
Average of the six best performers of other model groups	95 %	97 %	67 %	59 %	77 %	70 %
Tree- Medium Tree	95 %	97 %	68 %	57 %	78 %	77 %
Quadratic Discriminant	95 %	95 %	55 %	54 %	59 %	60 %
Naïve Bayes	97 %	94 %	57 %	68 %	62 %	73 %
Support Vector Machine- Quadratic	96 %	98 %	69 %	65 %	82 %	71 %
k-Nearest Neighbor (weighted KNN)	93 %	100 %	79 %	45 %	93 %	62 %
Ensemble - Boosted Trees	96 %	98 %	75 %	65 %	89 %	78 %

classification problems and to train the machine learning models.

This study showed that the ensemble method with random under-sampling algorithm (Ensemble – RUSBoost) outperformed 23 other models in the MATLAB program based on warning score, weighted warning score and true red accuracy measures. This method and measuring approach for EWIs distinguishes this study from the existing literature.

The model and framework proposed in this study can be applied in a bank setting. This will enable financial institutions to combine their internal metrics with market stress measures, thereby producing stronger risk warning signals specific to their business model or balance sheet structure. Utilising a machine learning model for this purpose is dynamic in nature and it can be automated for integration into management information systems. With further calibration, a link to real-time data may be possible.

The main limitation of this study is tour ability to use critical individual bank data which is not publicly available such as bank liquidity positions, liquidity forecasts, collateral calls and deposit outflows eg. Furthermore, regulatory authorities' capability to compare individual banks' positions and risks comparatively allows them to invest in risk signal detection with advanced methods and where required inform the bank in question to work on reducing the impact of the stress promptly by improving liquidity and capital position or reducing the risk. Another limitation is that the data used are US-centric. For developing countries or other regions different market data need to be used for the training and testing to explore the performance of the proposed framework and models.

Although this study addresses the problem of risk measurement from the perspective of an individual bank and how EWIs can be constructed using machine learning models, future research can be conducted by regulators and central banks using their superior data availability to incorporate dynamic EWI RAG status prediction. Specifically for the US market, the Federal Reserve can use its data superiority for individual bank data and combine it with market indicators to better measure EWIs regarding emerging stress around specific banks. The recent bank failures of Silicon Valley Bank, Signature Bank and First Republic Bank highlight the importance of further development in this area.

Disclaimer: The views and opinions expressed in this paper are those of the authors and they do not necessarily reflect the views of the HSBC Group or Yildiz Technical University. The initial version of this paper presented at the 7th International Conference (ICE-TEA Conference) on April 9–11,2021.

CRediT authorship contribution statement

Coskun Tarkocin: Conceptualization, Methodology, Data curation, Investigation, Formal analysis, Writing – original draft. **Murat Donduran:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing, Software.

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.

References

Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring Systemic Risk. *The Review fo Financial Studies, 30*(1), 2–47.

Aikman, D., Bridges, J., Burgess, S., Galletly, R., Levina, I., O'Neill, C., & Varadi, A. (2018). Measuring risks to UK financial stability. *Staff Working Paper No, 738*.

Al Janabi, M. A. (2021). Multivariate portfolio optimization under illiquid market prospects: A review of theoretical algorithms and practical techniques for liquidity risk management. *Journal of Modelling in Management, 16*(1), 288–309.

Aldasoro, I., Borio, C., & Drehmann, M. (2018). Early warning indicators fo banking crises; expanding the family. *BIS Quarterly Review*, 29–45.

Arif, A., & Anees, A. N. (2012). Liquidity risk and performance of banking system. *Journal of Financial Regulation and Compliance, 20*(2), 182–195.

BCBS. (1992). *A Framework for Measuring And Managing Liquidity*.

BCBS. (2000). *Sound Practices for Managing Liquidity in Banking Organisations*.

BCBS. (2008). *Principles for Sound Liquidity Risk Management and Supervision*.

BCBS. (2013). Evaluating early warning indicators of banking crisis: Satisfying policy requirements. BIS Working Papers No 421.

Betz F., Oprica S., Peltonen T.A., & Sarlin P. (2013, October 11). *Predicting Distress in European Banks.* Retrieved from ECB Working Paper no. 1597: https://ssrn.com/abstract=2338998.

Beutel, J., List, S., & Schweinitz, G. v. (2018). *An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?* Retrieved from Deutsche Bundesbank Discussion Paper no. 48/2018: http://dx.doi.org?10.2139/ssrn.3312608.

BOE. (2019, October). *Machine Learning in UK Financial Services.* Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/report/2019/machine-learning-in-uk-financial-services.pdf.

Bräuning M., Malikkidou D., Scalone S., & Scricco G. (2019). *A new approach to Early Warning Systems for small European banks.* ECB Working Papers Series No 2348.

Chen, Y.-K., Shen, C.-H., Kao, L., & Yeh, C.-Y. (2018). Bank Liquidity Risk and Performance. *Review of Pacific Basin Financial Markets and Policies, 21 (1)*.

Demirguc-Kunt, A., & Detragiache, E. (1998). The Determinants of Banking Crises in Developing and Developed Countries. *IMF Staff papers, 45 (1)*.

Drudi M.L., & Nobili S. (2021, June 22). *A liquidity risk early warning indicator for Italian banks:a machine learning approach.* Retrieved from Bank of Italy Temi di Discussione (Working Paper) No.1337: http://dx.doi.org/10.2139/ssrn.3891566.

Eross A., Urquhart A., & Wolfe S. (2015, September 29). *An Early Warning Indicator for Liquidity Shortages in the Interbank Market.* Retrieved from Eross, Andrea and Urquhart, Andrew and Wolfe, Simon, An Early Warning Indicator for Liquidity Shortages in the Interbank https://ssrn.com/abstract=2658797.

Federal Reserve Bank of St. Louis. (2021, February). Retrieved from https://fred.stlouisfed.org/series/STLFSI2.

FSB. (2017). *Artificial intelligence and machine learning in financial services*. Financial Stability Board.

Galletta, S., Mazzù, S., & Scannella, E. (2021). Risk committee complexity and liquidity risk in the European banking industry. *Journal of Economic Behavior & Organization*, 691–703.

Gaytán, A., & Johnson, C. A. (2002, October). A Review of the Literature on Early Warning Systems for Banking Crises. *Working Papers Central Bank of Chile 183*. Retrieved from https://ideas.repec.org/p/chb/bcchwp/183.html.

Goodhart C. (2008). Liquidity risk management. *Banque de France Financial Stability Review, Special issue on liquidity No.11*. 39-44.

Guerra, P., Castelli, M., & Côrte-Real, N. (2022). Machine learning for liquidity risk modelling: A supervisory perspective. *Economic Analysis and Policy*, 175–187.

Iachini E. & Nobili S. (2014, April 23). *An Indicator of Systemic Liquidity Risk in the Italian Financial Markets.* Retrieved from Bank of Italy Occasional Paper No. 217: https://ssrn.com/abstract=2489885.

Imbierowicz, B., & Rauch, C. (2014). The Relationship between Liquidity Risk and Credit Risk in Banks. *Journal of Banking & Finance, 40*(1), 242–256.

Kliesen K. & McCracken M. (2020, March 26). *The St. Louis Fed's Financial Stress Index, Version 2.0*. Retrieved from https://fredblog.stlouisfed.org/2020/03/the-st-louis-feds-financial-stress-index-version-2-0/.

Lang J.H.A.P.T., & P. Sarlin. (2018). *A framework for early-warning modeling with an application to banks.* ECB Working Paper Series.

MATLAB. (2019). *Ensemble Algorithms*. Retrieved from https://uk.mathworks.com/help/stats/ensemble-algorithms.html.

McKinsey&Company. (2021, 03 09). *An executive's guide to AI*. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai.

Navajas M.C., & Aaron T. (2013). *Financial Soundness Indicators and Banking Crisis.* IMF Working paper.

Padhan, R., & KPP. (2019). Effectiveness of Early Warning Models: A Critical Review and New Agenda for Future Directions. *Bulletin of Monetary Economics and Banking, 22*(4), 457–484.

scikit-learn. (2020). *1.11. Ensemble methods*. Retrieved from Scikit Learn Documentation: https://scikit-learn.org/stable/modules/ensemble.html.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2008). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *19th International Conference on Pattern Recognition*, (pp. 1-4).

Seiffert, C., Khoshgoftaar, T. M., Van, H. J., & Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans, 40*(1), 185–197.

Surjaningsih, N., Yumanita, D., & Deriantino, E. (2014). *Early Warning Indicators: Banking Liquidity Risk*. Bank Indonesia Working Paper.

Venkat, S., & Baird, S. (2016). *Liquidity Risk Management A Practitioner's Perspective*. Wiley Finance Series.