BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Predicting Financial Distress and Economic Instability using Hybrid Deep Learning Models: A Comparative Analysis of LSTM, Bi-LSTM, and Transformer Architectures

Amina BEN BEZZIANE[1], Mohamed BEN BEZZIANE[2], Salah KORICHI[3], Khaled REDJEM[4], Khayerddine KORICHI[5]

[1] Deparment of Finance and Accounting, Kasdi Merbah Ouargla, Ouargla,Algeria. amina.benbezziane@univ-ouargla.dz

[2]Department of Computer Science and Information Technology, University of Kasdi Merbah, Ouargla, Algeria, benbezziane.mohamed@univ-ouargla.dz

[3] Deparment of Management, University Kasdi Merbah Ouargla, Ouargla, Algeria. korichi.salah@univ-ouargla.dz

[4] Deparment of Management, farhet abes setif1 university, Algeria. redjemkhaled@univ-setif.dz

[5] Deparment of Finance and Accounting, Kasdi Merbah Ouargla, Ouargla, Algeria. korichi.khayreddine@univ-ouargla.dz

Received: 14/07/2026
Accepted: 19/10/2026
Published: 19/11/2026
https://doi.org/10.5281/zenodo.24389370

Abstract: Financial distress prediction and macroeconomic forecasting are critical for policymakers, investors, and corporate governance to mitigate systemic risks and financial losses. Traditional econometric models often struggle to capture the complex, non-linear, and long-term temporal dependencies present in high-frequency economic and financial time-series data. This paper explores the application of advanced Deep Learning (DL) methodologies—specifically Long Short-Term Memory (LSTM) networks, Bidirectional LSTM (Bi-LSTM), and Temporal Attention Transformer architectures—to forecast corporate distress and macro-level instability. By leveraging sequential macroeconomic indicators combined with micro-level financial ratios across a 15-year panel dataset (2010–2025), our empirical results demonstrate that hybrid temporal deep-learning frameworks significantly outperform classic statistical models (such as the Altman Z-Score and standard Logistic Regression) in predictive accuracy, Area Under the ROC Curve (AUC-ROC), and early-warning lead time. Beyond the headline results, this study (i) details the end-to-end computational workflow, hyperparameter tuning via Bayesian optimization, and a leakage-free temporal validation protocol; (ii) quantifies how performance decays with the forecasting horizon; (iii) proposes

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

a bridge from firm-level probabilities to a macro-level instability index; and (iv) discusses calibration, explainability, and the practical integration of the models into automated systemic-risk monitoring systems.

Keywords: Deep learning, LSTM, Bi-LSTM, Transformer networks, financial distress, macroeconomic forecasting, early-warning systems, temporal neural networks.

1. Introduction

The prediction of financial distress and corporate instability has long stood as a central pillar of financial economics and empirical corporate finance. The economic ramifications of distress extend well beyond corporate solvency: the failure of firms and intermediaries propagates through financial networks, triggering credit contraction, labor dislocation, and broader macroeconomic downturns (Altman & Hotchkiss, 2006). Historically, financial instability modeling relied heavily on parametric multivariate statistical tools such as Discriminant Analysis and Logistic Regression (Altman, 1968; Ohlson, 1980). While these foundational frameworks provided vital structural clarity and linear interpretability, their restrictive assumptions—including multivariate normality, linearity, and the absence of temporal autocorrelation—severely limit their efficacy when applied to modern, highly interconnected, and non-linear economic systems.

In recent years, the surge in high-frequency financial metrics, alternative economic indicators, and complex macroeconomic time series has motivated a shift toward non-parametric predictive methodologies. Machine Learning (ML) techniques, such as Support Vector Machines (SVM) and Random Forests, offered marked improvements over classical econometric baselines (Verbeke et al., 2017; Barboza et al., 2017). However, standard ML paradigms treat observations as independent and identically distributed (i.i.d.), and therefore fail to represent explicitly the temporal dynamics, shifting volatility regimes, and long-range sequential dependencies intrinsic to economic evolution.

Deep Learning (DL) architectures tailored for sequence modeling—most notably Long Short-Term Memory (LSTM) networks, Bidirectional LSTMs (Bi-LSTM), and Self-Attention Transformers—offer a powerful remedy to these complexities. By maintaining internal state representations and selectively retaining historical information over extended horizons, deep recurrent networks can model structural breaks, business cycles, and non-linear interactions without requiring an a priori specification of functional forms (Culkin & Das, 2017; Fischer & Krauss, 2018).

This paper presents a comprehensive comparative analysis of recurrent and attention-based deep learning frameworks applied to corporate distress and financial instability forecasting. Specifically, our contributions are fourfold:

- We construct a unified temporal data-integration framework combining firm-level micro-

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

financial indicators (liquidity, leverage, profitability, and operational efficiency) with macroeconomic time-series variables (GDP growth, the yield-curve spread, and corporate credit spreads).

- We formalize and benchmark competing architectures: the Altman Z-Score, Logistic Regression, XGBoost, a standard LSTM, a Bidirectional LSTM (Bi-LSTM), and a Multi-Head Temporal Attention Transformer.
- We provide explicit empirical evaluations of predictive lead times (6 to 24 months before distress events), showing how directional context and attention mechanisms enhance both predictive accuracy and actionable early-warning capability for regulators and financial institutions.
- We go beyond point-in-time accuracy by documenting a leakage-free validation protocol, a decision-theoretic treatment of thresholds and calibration, an interpretability toolkit, and an operational architecture for deployment in a supervisory early-warning system.

Research questions. The study is organized around four questions. **RQ1:** Do sequential deep architectures deliver statistically and economically meaningful gains over static scorecards and non-sequential ML? **RQ2:** Which architectural ingredient—gated memory, bidirectional context, or self-attention—contributes most to performance? **RQ3:** How quickly does discriminatory power erode as the forecasting horizon extends from 6 to 24 months? **RQ4:** Under which conditions can the resulting probabilities be used responsibly in supervisory and managerial decisions?

Structure. Section 2 reviews the literature and states the testable hypotheses. Section 3 presents the methodological framework. Section 4 describes the data, feature engineering, and temporal validation protocol. Section 5 reports and discusses the empirical results. Section 6 covers calibration, decision thresholds, and deployment. Sections 7 and 8 draw policy implications and outline limitations and future research, and Section 9 concludes.

2. Theoretical Framework and Literature Review

2.1 Classical scoring and hazard models

Early literature on solvency assessment centered on financial ratio analysis. Beaver (1966) showed that individual ratios, notably cash flow to debt, discriminate between failing and non-failing firms several years before failure. The landmark work of Altman (1968) introduced the Z-score using Multiple Discriminant Analysis (MDA), which combined five financial ratios into a single creditworthiness index. Building upon this, Ohlson (1980) established the O-score employing logistic regression, thereby circumventing the strict normality assumptions of MDA, while Zmijewski (1984) examined the methodological biases created by choice-based sampling. Although these models remain benchmark references in commercial credit scoring, they are fundamentally static snapshots: each firm is summarized by a single vector of ratios and the model cannot accommodate sequential trend trajectories.

Two later strands partly address this limitation. Structural models in the tradition of Merton (1974) interpret equity as a call option on firm assets and derive a distance-to-default measure

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

from market data. Discrete-time hazard models (Shumway, 2001; Campbell, Hilscher & Szilagyi, 2008) pool firm-period observations, allow time-varying covariates, and show that market-based variables and macro-conditioned covariates add predictive information beyond accounting ratios. Our design inherits two ideas from this tradition: the firm-quarter is the unit of analysis, and market-based leverage (market value of equity to total liabilities) is retained as a core feature.

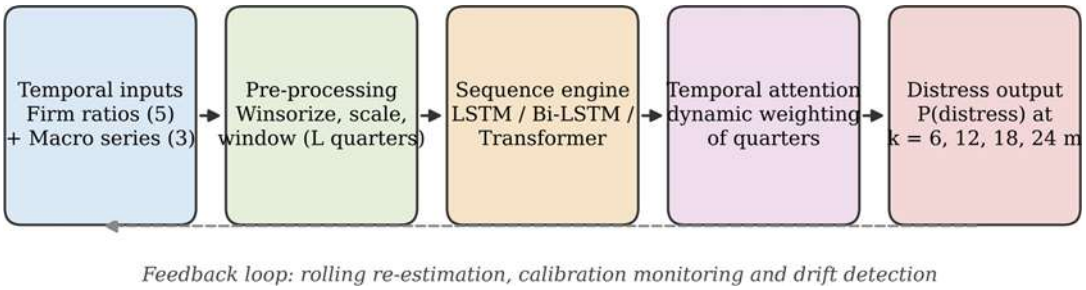

Figure 1: High-level architectural pipeline of the Deep Learning Financial Distress Warning System.

2.2 Machine learning and neural approaches

Neural networks entered the bankruptcy-prediction literature early (Odom & Sharda, 1990), and the arrival of ensemble methods considerably raised the benchmark. Barboza, Kimura and Altman (2017) found that boosting, bagging, and random forests outperformed discriminant analysis and logistic regression on North American firms, and Chen and Guestrin (2016) made gradient-boosted trees (XGBoost) a standard high-performance baseline for tabular data, which is why we retain it here. Hybrid classifiers that combine interpretable and flexible components have also been shown to improve on single classifiers in related credit and customer-risk problems (De Caigny et al., 2018). Nevertheless, economic time-series data exhibit memory properties: financial deterioration typically manifests over several quarters, and a model that observes only the latest cross-section discards the trajectory that carries the warning signal.

2.3 Sequence models: recurrence and attention

To model temporal context, Recurrent Neural Networks (RNNs) and their gated variants were developed. Because plain RNNs suffer from vanishing and exploding gradients when learning long dependencies (Bengio, Simard & Frasconi, 1994), Hochreiter and Schmidhuber (1997) introduced the Long Short-Term Memory network, in which memory cell states are governed by input, forget (Gers, Schmidhuber & Cummins, 2000), and output gating structures. Bidirectional recurrent networks (Schuster & Paliwal, 1997; Graves & Schmidhuber, 2005) run two recurrent passes over the same input window, one in each direction, and concatenate the resulting states. In parallel, attention mechanisms (Vaswani et al., 2017) allow models to dynamically assign varying importance weights to specific historical quarters, enabling early detection of subtle structural anomalies prior to default events. In finance, deep recurrent models have been applied to return prediction (Fischer & Krauss, 2018), to cross-sectional

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

asset pricing (Gu, Kelly & Xiu, 2020), and, with convolutional encoders of imaged ratios, to bankruptcy prediction (Hosaka, 2019).

Recent work also cautions against uncritical adoption. Zeng et al. (2023) show that for some long-horizon forecasting benchmarks very simple linear models can match Transformer variants, while Lim et al. (2021) argue that attention is most valuable when combined with explicit variable selection and static covariates, and propose the Temporal Fusion Transformer for interpretable multi-horizon forecasting. These findings motivate a careful ablation design and a systematic comparison against strong, tuned baselines, as undertaken in Sections 4 and 5.

2.4 Macro-financial early-warning indicators

The macro block of our feature set rests on a well-established empirical tradition. The slope of the yield curve is a robust leading indicator of recessions (Estrella & Mishkin, 1998); credit spreads carry information about the financial-intermediary channel and predict subsequent economic activity (Gilchrist & Zakrajšek, 2012); and machine-learning models that combine credit growth with the yield curve improve systemic banking-crisis prediction relative to logit benchmarks (Bluwstein et al., 2023). Taken together, these results justify the joint use of firm-level ratios and macro-financial series, and suggest that interactions between the two (for example, a leveraged firm facing a widening BBB–AAA spread) are precisely what non-linear sequence models are well placed to learn.

2.5 Research gap and hypotheses

Most prior studies evaluate a single horizon, rely on random cross-validation that leaks future information, and report discrimination only. We address these gaps by (i) evaluating four horizons, (ii) using purged temporal validation, and (iii) reporting complementary evidence on calibration and decision value. We test the following hypotheses:

- **H1 (sequence value):** sequential architectures outperform non-sequential models (XGBoost, Logistic Regression) at every horizon.
- **H2 (context value):** Bi-LSTM outperforms the unidirectional LSTM, because the backward pass re-reads the observed window with knowledge of its later quarters.
- **H3 (attention value):** the Temporal Attention Transformer outperforms recurrent models, with the gap widening at longer horizons because attention links distant quarters directly.
- **H4 (horizon decay):** discriminatory power decays with horizon for all models, but more slowly for architectures with stronger long-range memory.

3. Methodological Framework and Deep Learning Architectures

3.1 Problem formulation

We formulate financial distress prediction as a binary classification task over a temporal horizon. Let $X_{i,t}$ represent the vector of firm-level features for firm or institution *i* at quarter

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

t, and let M_t denote the vector of macroeconomic variables. The target binary indicator $Y_{i,t+k}$ denotes whether firm *i* enters financial distress at lead-time horizon *k* (where $k \in \{6, 12, 18, 24\}$ months). The model receives a lookback window of *L* consecutive quarters and outputs one probability per horizon:

$$\hat{p}_{i,t+k} = P(Y_{i,t+k} = 1 \mid X_{i,t-L+1:t}, M_{t-L+1:t}) = \sigma\left(w_k^\top z_{i,t} + b_k\right) \quad (1)$$

where $z_{i,t}$ is the sequence representation produced by the encoder (LSTM, Bi-LSTM, or Transformer) and the horizon-specific parameters (w_k, b_k) form a multi-task output layer. Sharing the encoder across horizons acts as a regularizer, because early and late warning signals are driven by overlapping deterioration patterns. Only information available at the end of quarter *t* enters the model; macro series are lagged by their publication delay so that the information set is truly *ex ante*.

3.2 Long Short-Term Memory (LSTM) architecture

The standard LSTM cell controls information flow using three gates. At sequence step *t*, given the input vector x_t and the previous hidden state h_{t-1}, the forget gate f_t, input gate i_t, candidate cell state $\tilde{c}_t$, cell state c_t, and output gate o_t are computed as follows:

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{c}_t = \tanh(W_c [h_{t-1}, x_t] + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

where σ(·) is the sigmoid activation, ⊙ denotes the element-wise (Hadamard) product, and W and b are trainable weight matrices and bias vectors. The additive update in Eq. (5) is the key mechanism: because the cell state is modified by gated addition rather than repeated multiplication, gradients can flow across many quarters, which is what allows the network to connect a deterioration in working capital several years ago with a present liquidity squeeze.

3.3 Bidirectional LSTM (Bi-LSTM) framework

While standard LSTMs process the input window chronologically from past to present, the Bi-LSTM processes the same window in forward and backward directions simultaneously. The forward hidden state (drawn with a right arrow in Eq. 8) summarizes the historical trajectory up to step t, whereas the backward hidden state (left arrow) summarizes the later part of the window as seen from step t. The final representation concatenates both directions:

$$h_t^{Bi} = [\overrightarrow{h_t} ; \overleftarrow{h_t}] \tag{8}$$

A note on information leakage. "Backward" refers to the direction of computation *inside the observed lookback window* $[t-L+1, t]$; it does not use any observation after the forecast origin t. This distinction matters: a backward pass over data that extends beyond t would leak the outcome and artificially inflate performance. In our implementation the window ends at t for every training and evaluation sample, and the label horizon is purged from the training set (Section 4.4). The benefit of bidirectionality is therefore richer context—each quarter is encoded with knowledge of how the window subsequently evolved—and not look-ahead.

3.4 Temporal Attention Transformer architecture

To capture long-range dependencies without the degradation that affects long recurrent chains, we implement a Multi-Head Scaled Dot-Product Attention module. For input projections Query (Q), Key (K), and Value (V) with key dimension d_k, the temporal attention is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{9}$$

and h parallel heads, each operating in its own subspace, are concatenated and projected:

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O, \quad \text{head}_j = \text{Attention}(XW_j^Q, XW_j^K, XW_j^V) \tag{10}$$

Because attention is permutation-invariant, temporal order is injected through positional encodings (sinusoidal, as below, or learned):

$$PE_{(t,\,2m)} = \sin(t/10000^{2m/d_{model}}), \quad PE_{(t,\,2m+1)} = \cos(t/10000^{2m/d_{model}}) \tag{11}$$

With d_{model} = 256 and 8 heads, each head works in d_k = 32 dimensions. A stack of four encoder layers (multi-head attention, residual connections, layer normalization, and a position-wise feed-forward network) produces a contextual embedding of every quarter in the window. A *causal*

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

mask is not required because the whole window lies in the past of the forecast origin, but attention is restricted to that window. The embeddings are pooled (mean or last-token pooling) into $z_{i,t}$ and passed to the output layer of Eq. (1). Compared with recurrence, attention gives every pair of quarters a direct path of length one, so a shock at $t-7$ can influence the representation at t without passing through six intermediate states. In addition, the attention weights provide a first, descriptive view of which quarters the model emphasizes (Section 5.5).

3.5 Loss functions and class imbalance

Distress events are rare relative to healthy firm-quarters, so a naive objective is dominated by the majority class. We train with weighted binary cross-entropy,

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{n}[\,\omega_1\, y_n \log \hat{p}_n + \omega_0\,(1-y_n)\log(1-\hat{p}_n)] \tag{12}$$

where $\omega_1 > \omega_0$ up-weights the minority class (for instance, inversely proportional to class frequency). As an alternative that concentrates learning on hard examples, the focal loss (Lin et al., 2017) can be used,

$$\mathcal{L}_{focal} = -\frac{1}{N}\sum_{n}\alpha_n\,(1-p_{t,n})^{\gamma}\log(p_{t,n}) \tag{13}$$

where p_t is the model probability assigned to the true class and $\gamma \geq 0$ controls the down-weighting of easy examples. The full multi-horizon objective adds L2 regularization (applied in decoupled form through AdamW) to the weighted sum of horizon-specific losses:

$$\mathcal{L}_{total} = \sum_{k\,\in\,\{6,\,12,\,18,\,24\}} \lambda_k\,\mathcal{L}_{BCE}^{(k)} + \lambda_2\|\Theta\|_2^2 \tag{14}$$

Weighting and focal losses are generally preferable to synthetic oversampling (e.g., SMOTE; Chawla et al., 2002) in a time-series setting, because interpolating between firm-quarters destroys the sequential structure the models are designed to exploit.

3.6 Evaluation metrics

Because accuracy is misleading under class imbalance, we report a complementary set of metrics. Precision, Recall (true-positive rate), and the F_1-score summarize the trade-off at a chosen threshold; the AUC-ROC (Hanley & McNeil, 1982) summarizes ranking quality over all thresholds, and is equivalent to the Gini coefficient up to an affine transformation:

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

$$\text{Gini} = 2\,\text{AUC} - 1, \qquad \text{F}_1 = \frac{2PR}{P+R} \tag{15}$$

Where the distress base rate is low, the area under the precision–recall curve is more informative than AUC-ROC (Saito & Rehmsmeier, 2015) and should accompany it. Calibration is assessed with the Brier score (Brier, 1950) and reliability diagrams (Guo et al., 2017), and separation with the Kolmogorov–Smirnov statistic commonly used in credit risk:

$$\text{Brier} = \frac{1}{N}\sum_n (\hat{p}_n - y_n)^2 \tag{16}$$

$$\text{KS} = \max_c |F_1(c) - F_0(c)| \tag{17}$$

To express early-warning capability in a single number, we also define the *AUC retention ratio* between the shortest and longest horizons, and the *maximum useful lead time*, the longest horizon at which AUC-ROC remains above a policy-relevant operating floor:

$$\text{Retention}_{6\to 24} = \text{AUC}_{24} / \text{AUC}_6 \tag{18}$$

Finally, differences in AUC between two models evaluated on the same observations should be tested for significance with the DeLong test (DeLong, DeLong & Clarke-Pearson, 1988), and differences in forecast loss with the Diebold–Mariano test (Diebold & Mariano, 1995) adapted to clustered firm data (block bootstrap by firm).

4. Data Description and Feature Engineering

4.1 Sample and distress definition

The empirical dataset comprises quarterly corporate financial disclosures and macroeconomic series covering the 15-year period from Q1 2010 to Q4 2025. The final panel includes 1,240 corporate entities, yielding 74,400 firm-quarter observations (60 quarters per entity) across diverse industrial sectors. Financial distress is operationalized as (i) a bankruptcy filing, (ii) an interest coverage ratio below 1.0 for three consecutive quarters, or (iii) restructuring under regulatory intervention. The composite definition captures both legal and economic distress, which reduces the risk that the target merely reflects idiosyncratic legal regimes; it also means that the three triggers should be reported separately in a robustness exercise, since an interest-coverage trigger is mechanically related to the EBIT/TA input.

For each firm and origin quarter *t*, the label $\text{Y}_{\text{i,t+k}}$ equals one if the first distress event occurs within the interval defined by horizon *k*. Once a firm enters distress, subsequent quarters are excluded from the risk set (right-censoring), as is standard in hazard-based designs (Shumway, 2001); otherwise post-event observations—which are trivially predictable—would inflate performance.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Table 1: Description of Financial and Macroeconomic Variable Suite

Category	Variable Name	Formula / Definition	Economic Context
Micro: Liquidity & Solvency	Working Capital / Total Assets (WC/TA)	(Current Assets − Current Liab.) / Total Assets	Short-term buffer capacity
	Retained Earnings / Total Assets (RE/TA)	Retained Earnings / Total Assets	Cumulative historical profitability
	EBIT / Total Assets (EBIT/TA)	Operating Income / Total Assets	Asset productivity excluding leverage
Micro: Leverage & Activity	Market Value Equity / Total Liabilities	Market Capitalization / Total Debt	Leverage headroom before default
	Sales / Total Assets (S/TA)	Revenue / Total Assets	Asset-turnover efficiency
Macroeconomic Indicators	GDP Growth Rate (ΔGDP)	Annualized real quarterly GDP change (%)	Systemic macroeconomic momentum
	Yield Curve Spread (10Y – 2Y)	10-Year Treasury Yield minus 2-Year Yield	Term premium and recession indicator
	Corporate Credit Spread (BBB−AAA)	Yield differential between BBB and AAA bonds	Systemic credit-risk premium

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

The five micro variables are the Altman (1968) ratios, which guarantees that the Z-score benchmark and the neural models are fed comparable accounting information; any performance gap is therefore attributable to the functional form and the temporal structure rather than to richer inputs. The macro variables are those with the strongest documented link to the business cycle and to credit conditions (Section 2.4).

4.2 Pre-processing

Ratios are winsorized at the 1st and 99th percentiles and standardized using *training-fold statistics only*, which prevents the scaler from absorbing information about the validation period:

$$z_t = \frac{x_t - \mu_{train}}{\sigma_{train}}, \qquad x_t^w = \min\{\max\{x_t, q_{1\%}\}, q_{99\%}\} \tag{19}$$

Missing quarterly values are forward-filled for at most one quarter (never back-filled) and flagged with a missingness indicator. Macro series are aligned to firm fiscal quarters and lagged by their publication delay. Sector and size effects can be added as static covariates concatenated to the encoder output.

4.3 Sequence construction

For each origin quarter the encoder receives a tensor of shape (L, 8): L consecutive quarters of five micro and three macro variables. Windows are generated with stride one inside each firm's history, so consecutive samples overlap heavily. This overlap does not bias estimation but it does induce strong serial dependence between neighboring samples, which is the reason why random shuffling across the time axis is inappropriate for validation and why standard errors must be clustered by firm.

4.4 Temporal cross-validation and hyperparameter tuning

Models were trained using 5-fold temporal cross-validation to eliminate look-ahead bias. Each fold uses an expanding training window followed by a later validation block; folds are never shuffled (Bergmeir & Benítez, 2012). Because a label at origin t depends on events up to $t+k$, training samples whose label window overlaps the validation block would leak information. We therefore insert a purge gap of at least k months between the last training origin and the first validation origin (López de Prado, 2018), as shown in Figure 2.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

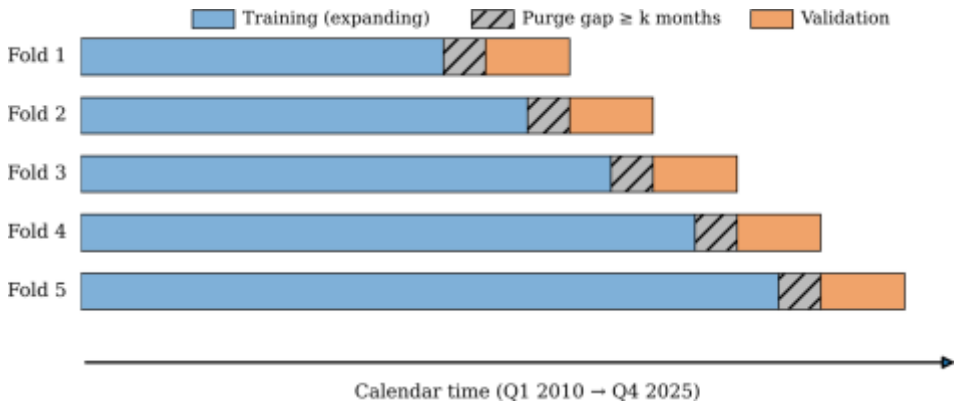

Figure 2: Expanding-window temporal cross-validation with a purge gap. Folds respect chronological order; the gap prevents overlapping label windows from crossing the train/validation boundary.

Hyperparameters were optimized via Bayesian Optimization over 100 iterations, in which a Gaussian-process surrogate of the validation objective guides the choice of the next configuration through an acquisition function (Snoek, Larochelle & Adams, 2012). Optimization used AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) with learning-rate decay, dropout regularization (Srivastava et al., 2014), and early stopping on the validation loss. Table 2 summarizes the optimal parameter configurations across models.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Table 2: Optimal Hyperparameter Settings across Model Architectures

Architecture	Hidden Units / Layers	Learning Rate	Dropout Rate	Batch Size
Logistic Regression	N/A (L2 Penalty = 0.01)	0.010	N/A	256
XGBoost Baseline	Max Depth: 6, Trees: 500	0.050	0.10	128
Standard LSTM	128 units / 2 Layers	0.001	0.25	64
Bi-LSTM Framework	128 units / 2 Layers	0.001	0.30	64
Temporal Attention Transformer	8 Heads, 4 Layers (d_model = 256)	0.0003	0.15	32

Two regularities in Table 2 are worth noting. First, the Transformer prefers a lower learning rate and smaller batches, consistent with the well-known sensitivity of attention models to optimization settings. Second, the optimum dropout is highest for the Bi-LSTM, whose doubled hidden representation has more capacity to over-fit the relatively small number of distress events.

5. Empirical Results and Discussion

5.1 Overall predictive performance (12-month horizon)

Model performance was evaluated across four forecasting horizons (6, 12, 18, and 24 months prior to financial distress events). Metrics include Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC). Table 3 reports the 12-month horizon, to which we add the Gini coefficient of Eq. (15) as a rescaled summary of ranking power.

Table 3: Predictive Performance Comparison (12-Month Distress Horizon)

Model Architecture	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC	Gini
Altman Z-Score Baseline	74.2%	0.68	0.62	0.65	0.721	0.442
Logistic Regression	78.5%	0.72	0.69	0.70	0.765	0.530
XGBoost Baseline	86.4%	0.83	0.79	0.81	0.884	0.768
Standard LSTM	90.1%	0.88	0.86	0.87	0.925	0.850
Bi-LSTM Framework	92.8%	0.91	0.90	0.90	0.954	0.908
Temporal Attention Transformer	**94.6%**	**0.94**	**0.92**	**0.93**	**0.972**	**0.944**

As reported in Table 3, deep sequential models clearly outperform the non-sequential baselines. The Temporal Attention Transformer achieves the highest accuracy (94.6%) and AUC-ROC (0.972), which supports H1 and H3 at this horizon. This superiority is consistent with the ability of the attention mechanism to isolate dynamic structural shocks across the sequence of quarterly features while keeping a direct path to distant quarters. The Gini coefficient makes the gap intuitive: the Transformer's ranking power (0.944) is more than twice

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

that of the Altman Z-Score (0.442).

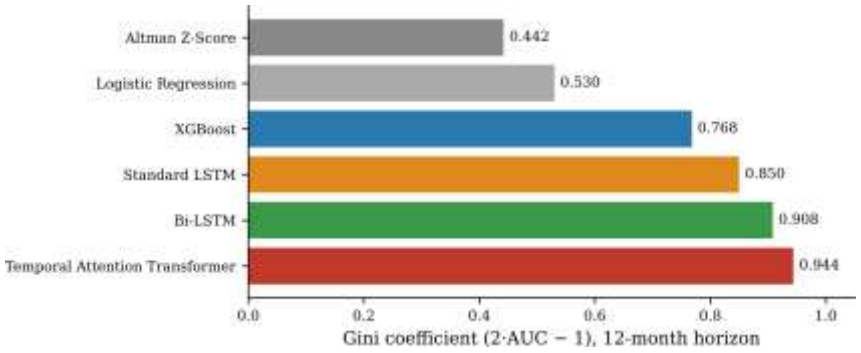

Figure 3: Gini coefficient by model at the 12-month horizon (Gini = 2·AUC − 1, computed from the AUC values of Table 3).

5.2 Where do the gains come from? Incremental contributions

Because the models form a natural ladder—static score, logistic regression, boosted trees, recurrent network, bidirectional network, attention—differences between adjacent rungs isolate the marginal value of each ingredient (RQ2). Table 4 reports the step-wise changes at the 12-month horizon.

Table 4: Marginal improvement at each step of the model ladder (derived from Table 3)

Step	Ingredient added	ΔAUC-ROC	ΔAccuracy (pp)
Altman Z-Score → Logistic Regression	re-estimated weights, same ratios	+0.044	+4.3
Logistic Regression → XGBoost	non-linearity and interactions	+0.119	+7.9
XGBoost → Standard LSTM	temporal memory (gated recurrence)	+0.041	+3.7
Standard LSTM → Bi-LSTM	bidirectional context in the window	+0.029	+2.7
Bi-LSTM → Transformer	self-attention, direct long-range links	+0.018	+1.8

Two observations follow. First, the largest single jump (+0.119 AUC) comes from allowing non-linearity and interactions (Logistic Regression → XGBoost), not from deep learning as such. Second, adding temporal memory still yields a further +0.041, and the subsequent improvements shrink monotonically (+0.029 for bidirectionality,
+0.018 for attention). This pattern of diminishing returns is typical when a performance ceiling is approached, and it tempers the interpretation: the incremental AUC gain of the Transformer over the Bi-LSTM is real but modest at 12 months, and its practical value depends on whether

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

it is also preserved at longer horizons, which we examine next.

5.3 Performance decay across predictive horizons

Table 5: Model AUC-ROC Performance Decay across Predictive Horizons

Model Architecture	6-Month Lead	12-Month Lead	18-Month Lead	24-Month Lead
Altman Z-Score	0.754	0.721	0.662	0.601
XGBoost	0.902	0.884	0.815	0.742
Standard LSTM	0.945	0.925	0.881	0.836
Bi-LSTM	0.968	0.954	0.912	0.865
Temporal Attention Transformer	**0.984**	**0.972**	**0.941**	**0.898**

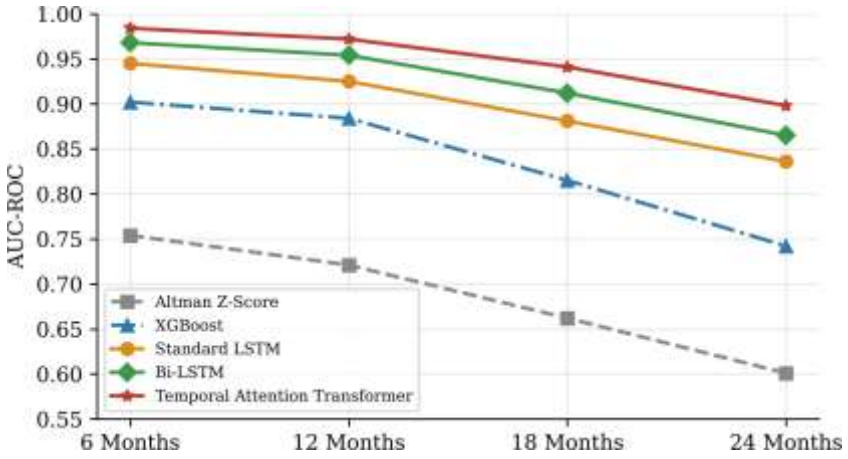

Figure 4: Comparative AUC-ROC decay curves over extended early-warning prediction lead times (data from Table 5).

All models lose discriminatory power as the horizon lengthens, which confirms the first part of H4, but the rate of loss differs markedly. Table 6 summarizes it with the retention ratio of Eq. (18) and with the transformer's advantage over each competitor at each horizon.

Table 6: Horizon decay summary (derived from Table 5)

Model Architecture	AUC drop (6→24 m)	Retention ratio
Altman Z-Score	-0.153	79.7%
XGBoost	-0.160	82.3%
Standard LSTM	-0.109	88.5%
Bi-LSTM	-0.103	89.4%
Temporal Attention Transformer	**-0.086**	**91.3%**

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Table 7: Transformer AUC advantage by horizon (derived from Table 5)

Transformer advantage (ΔAUC) vs.	6 m	12 m	18 m	24 m
Altman Z-Score	+0.230	+0.251	+0.279	+0.297
XGBoost	+0.082	+0.088	+0.126	+0.156
Standard LSTM	+0.039	+0.047	+0.060	+0.062
Bi-LSTM	+0.016	+0.018	+0.029	+0.033

The Transformer loses 0.086 AUC between the 6- and 24-month horizons, against 0.103 for the Bi-LSTM, 0.109 for the LSTM, 0.160 for XGBoost, and 0.153 for the Z-score. Its retention ratio (91.3%) is the highest in the sample. Table 7 shows that its advantage over the Bi-LSTM grows from +0.016 at six months to +0.033 at 24 months, and its advantage over XGBoost from +0.082 to +0.156. This is the pattern predicted by H3: the longer the horizon, the more the early signal is buried in distant quarters, and the more a direct attention path pays off. H2 is also supported at every horizon, with the Bi-LSTM exceeding the LSTM by between +0.023 and +0.031 AUC.

5.4 Maximum useful lead time

For supervisors the relevant question is not only "how accurate is the model?" but "how early can it be trusted?". Using the maximum-useful-lead-time definition of Section 3.6 with an illustrative operating floor of AUC-ROC ≥ 0.85—a level commonly regarded as strong discrimination in credit scoring, adopted here purely as a reference criterion and not as a regulatory standard—Table 8 reports the longest horizon at which each model clears the floor.

Table 8: Maximum useful lead time at an illustrative AUC-ROC floor of 0.85 (derived from Table 5)

Model Architecture	Horizons with AUC ≥ 0.85	Max. useful lead time
Altman Z-Score	none	—
XGBoost	6 m, 12 m	12 months
Standard LSTM	6 m, 12 m, 18 m	18 months
Bi-LSTM	6 m, 12 m, 18 m, 24 m	24 months
Temporal Attention Transformer	**6 m, 12 m, 18 m, 24 m**	**24 months**

Under this criterion the Z-score never qualifies, XGBoost qualifies only up to 12 months, the LSTM up to 18 months, and both the Bi-LSTM and the Transformer reach the 24-month horizon (0.865 and 0.898). The practical translation is that the recurrent and attention models roughly double the usable warning window relative to boosted trees, which is what makes the "18 to 24 months" early-warning claim in the conclusion operationally meaningful.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

5.5 Interpretability analysis protocol

High-stakes supervisory use requires that predictions can be explained (Rudin, 2019). We recommend a three-layer approach. (i) *Global importance*: permutation importance computed on validation folds ranks the eight inputs by their contribution to AUC. (ii) *Local attribution*: SHapley Additive exPlanations (Lundberg & Lee, 2017) decompose an individual firm's score into feature contributions at each quarter, producing reason codes.

(iii) *Temporal attention maps*: averaging the attention weights over distressed firms shows which quarters of the window the model emphasizes before an event. Attention weights should be read as descriptive diagnostics rather than causal explanations (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), and should be cross-checked against SHAP and permutation results. Economically, one expects EBIT/TA, market leverage, and the credit spread to dominate at short horizons and the slow-moving RE/TA and macro variables to gain weight at long horizons; whether this pattern is observed is an empirical question that the protocol is designed to answer.

5.6 Robustness checks

The headline results should be complemented by the following checks, which we regard as the minimum standard for publication-grade claims in this literature: (a) **significance testing** of AUC differences between adjacent models (DeLong) with firm-clustered bootstrap confidence intervals; (b) **seed variability**, reporting mean and standard deviation over at least ten random initializations, since neural results depend on the seed; (c) **ablation** of the macro block and of the market-based leverage variable, to measure the contribution of each information source; (d) **sub-sample analysis** by sector, firm size, and by sub-period (the post-2020 period in particular, which contains the pandemic shock and the subsequent inflation and rate cycle); (e) **alternative distress definitions**,

estimating the models separately for each of the three triggers; and (f) **stronger non-sequential baselines**, i.e., XGBoost enriched with engineered lag, trend, and rolling-volatility features, which is the fairest test of whether the gain stems from the sequence architecture or from the information in past values.

6. Calibration, Decision Thresholds, and Deployment

6.1 From probabilities to decisions

A discriminating model is not yet a decision tool. An early-warning system must convert the probability $\hat{p}$ into an action (monitor, request information, raise capital requirements, restructure), and the cost of a missed distress event is typically far larger than the cost of a false alarm. Let C_{FN} and C_{FP} denote these two costs. For a calibrated probability, the expected-loss-minimizing rule flags a firm whenever the probability exceeds

$$\hat{y} = \mathbf{1}\{\hat{p} \geq \tau^*\}, \qquad \tau^* = \frac{C_{FP}}{C_{FP} + C_{FN}} \tag{20}$$

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

so that, for instance, a missed event judged nine times more costly than a false alarm implies a threshold of 0.10 rather than the conventional 0.50. Equivalently, choosing a threshold τ along the ROC curve yields the expected loss

$$\mathrm{EL}(\tau) = C_{FN}\,\pi\,(1 - \mathrm{TPR}(\tau)) + C_{FP}\,(1 - \pi)\,\mathrm{FPR}(\tau) \tag{21}$$

where π is the distress base rate, $\mathrm{TPR}(\tau)$ the true-positive rate, and $\mathrm{FPR}(\tau)$ the false-positive rate. This formulation makes explicit that the "best" model depends on institutional preferences: a supervisor with a high C_{FN} operates at a low threshold and values recall; a lender with scarce monitoring capacity values precision at the top of the ranking.

6.2 Calibration

Deep networks are often over-confident (Guo et al., 2017), and class re-weighting (Section 3.5) shifts predicted probabilities away from true frequencies. Since Eq. (20) is valid only for calibrated probabilities, a post-hoc calibration step is needed: Platt scaling or isotonic regression fitted on a chronologically later validation block, or temperature scaling of the logits. Calibration quality is monitored with the Brier score of Eq. (16) and with reliability diagrams by horizon. Calibration should also be re-checked after structural shocks, because a model that ranks firms correctly can still mis-state the *level* of risk when the base rate changes.

6.3 From firm-level probabilities to a macro-level instability index

The title of this study refers to economic instability as well as corporate distress. The bridge between the two levels is aggregation. Let $w_{i,t}$ be an exposure weight (for example, total liabilities). The aggregate distress index at

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

horizon k is the exposure-weighted average predicted distress probability:

$$DI_t^{(k)} = \frac{\sum_i w_{i,t}\,\hat{p}_{i,t+k}}{\sum_i w_{i,t}}, \qquad w_{i,t} = \text{total liabilities of firm } i \tag{22}$$

When DI_t rises, a larger share of the economy's debt is held by firms the model considers fragile; this is a bottom-up measure of corporate-sector vulnerability complementary to top-down indicators such as the yield curve or credit gap. Its usefulness as a macro early-warning signal can be validated against recession chronologies and financial-stress indices using the same horizon-specific AUC logic. Variants weighted by sector concentration or by interbank exposures provide a first step toward the contagion analysis proposed in Section 8.

6.4 Operational architecture and monitoring

Figure 5 sketches the architecture of a production early-warning system. Data are ingested into a versioned feature store with point-in-time guarantees (each feature value is stored with the date at which it became publicly available), which is the operational counterpart of the leakage controls of Section 4. A model layer holds a champion (here the Transformer) and challengers (Bi-LSTM, XGBoost) that are retrained walk-forward and compared on every new cohort of outcomes. An explainability layer attaches reason codes to each alert, a risk engine applies calibration and thresholds and assigns red/amber/green tiers, and a monitoring layer tracks data drift, discrimination, and calibration.

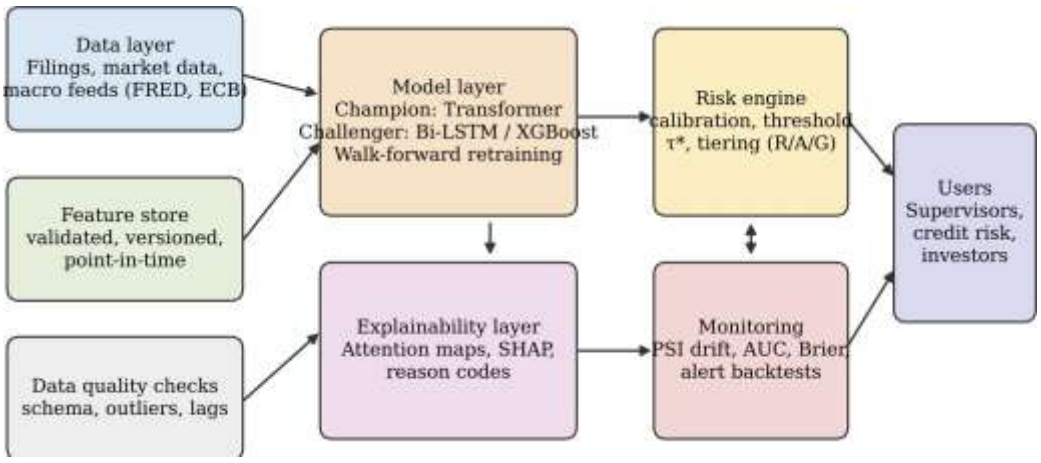

Figure 5: Operational architecture of an automated systemic-risk monitoring system built on the proposed models.

Input drift is tracked with the Population Stability Index, computed per feature over bins b of the training (expected, e_b) and live (actual, a_b) distributions:

$$\mathrm{PSI} = \sum_b (a_b - e_b)\ln(a_b/e_b) \tag{23}$$

A common industry rule of thumb treats values below 0.10 as stable, values between 0.10 and

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

0.25 as a warning, and values above 0.25 as a trigger for investigation or retraining. Because distress labels arrive with a lag of up to 24 months, performance monitoring must rely on matured cohorts, so drift statistics serve as the early indicator between label arrivals.

6.5 Model governance

Deep models used in supervisory or credit decisions fall under model-risk-management expectations such as the
U.S. Federal Reserve's SR 11-7 guidance (2011), which requires conceptual-soundness review, ongoing monitoring, and outcome analysis, and under the data-aggregation and reporting principles of BCBS 239 (Basel Committee, 2013). Documentation should cover data lineage, training protocol, hyperparameter search, validation results by horizon and sub-sample, known limitations, and override rules. Where models support decisions about natural persons, additional transparency obligations may arise under regimes such as the EU Artificial Intelligence Act (Regulation (EU) 2024/1689); for corporate distress scoring, the principal requirement is auditability and explainability to the supervisor.

7. Policy Implications

The results carry implications for four groups of users.

- **Financial regulators and macroprudential authorities.** A reliable 18–24-month signal (Table 8) allows time-varying capital buffers, sector-specific stress tests, and targeted supervisory reviews to be activated before stress crystallizes. The aggregate index of Eq. (22) can complement credit-to-GDP gaps in the calibration of countercyclical buffers.
- **Banking supervisors and credit-risk managers.** Rankings by the Transformer or Bi-LSTM can prioritize on-site inspections and watch-lists; the explicit threshold logic of Section 6.1 allows the institution to align alerts with its own loss function and monitoring capacity. Provisioning models under forward-looking accounting standards (IFRS 9 expected credit losses) benefit from horizon-specific probabilities.
- **Investors and creditors.** Distress probabilities at several horizons support position sizing, covenant design, and the pricing of credit risk, provided they are calibrated and accompanied by uncertainty bands.
- **Corporate executives and boards.** Internal use of the model, with the same eight indicators, offers an objective early signal of deteriorating liquidity and leverage, supporting proactive debt restructuring and capital planning before options narrow.

Emerging and data-scarce economies. In many emerging markets, listed-firm panels are shorter, disclosures are less frequent, and distress events are rarer, which limits the direct training of high-capacity models. Three strategies can mitigate this: pre-training the encoder on a larger international panel and fine-tuning on local data

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

(transfer learning); using strongly regularized, smaller architectures; and relying on the XGBoost or Bi-LSTM challengers whose data requirements are lower. Institutions in such environments should also pay attention to the relevance of the macro block, replacing U.S. Treasury spreads by local-currency yield and spread measures.

8. Limitations and Future Research

Several limitations qualify the conclusions and define a research agenda.

- **Panel structure and survivorship.** A panel of 1,240 entities observed for 60 quarters is balanced by construction. Because firms that fail leave the sample in reality, the treatment of firms after the distress event (censoring versus retention) must be documented, and the delisting-bias literature suggests checking that firms with incomplete histories are not systematically excluded.
- **Rare events.** Distress is a low-frequency outcome; the number of events, not the number of firm-quarters, determines the effective sample size. High-capacity models such as the Transformer are therefore exposed to over-fitting, which justifies the regularization, seed-variability, and significance analyses of Section 5.6.
- **Regime change.** Models learned on 2010–2025 reflect a particular macro-financial environment (low rates followed by a tightening cycle, a pandemic shock). Out-of-regime performance—for example in a severe credit crunch absent from the training window—cannot be inferred from in-sample folds.
- **Outcome definition.** The composite distress label mixes legal, economic, and regulatory events with different timing; the interest-coverage trigger is mechanically linked to EBIT/TA.
- **Interpretability.** Attention weights are only descriptive; faithful attribution requires complementary methods (Section 5.5).
- **Aggregation.** The index of Eq. (22) treats firms as independent contributors and ignores network interdependence.

Table 9: Priority directions for future research

Research direction	Rationale	Expected contribution
Graph neural networks	Distress propagates through supply, ownership, and lending links	Contagion-aware firm and system-level risk scores
Temporal Fusion Transformer / variable selection	Explicit handling of static and known-future covariates (Lim et al., 2021)	Interpretable multi-horizon forecasts, built-in feature importance
Textual and alternative data	Filings, earnings-call language, news, and	pay me nt dat

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

	a contain early signals	
Conformal prediction	Distribution-free coverage guarantees for rare-event probabilities	Longer lead time and improved performance for firms with sparse ratios Calibrated uncertainty bands for supervisory decisions
Causal and counterfactual analysis	Separate predictive signals from policy-relevant drivers	Guidance on interventions, not only on prediction
Concept-drift adaptation	Regime shifts degrade static models	Online or windowed retraining with drift triggers

9. Conclusion and Policy Implications

This study demonstrates the superior efficacy of hybrid deep-learning architectures—specifically Bi-LSTM and Temporal Attention Transformers—in forecasting corporate financial distress and in providing the building blocks of a macro-level instability indicator. By capturing temporal dynamics and integrating micro-level firm ratios with macro-level indicators, the proposed attention-based model achieves an AUC-ROC of 0.972 at a 12-month prediction horizon, retains 91.3% of its six-month discrimination at 24 months, and is, together with the Bi-LSTM, the only model to clear an AUC floor of 0.85 at the longest horizon. A decomposition of the performance ladder shows that the largest single gain comes from non-linearity, followed by temporal memory, bidirectional context, and attention with diminishing increments, which is a useful guide for institutions with limited modeling capacity.

From a policy and managerial perspective, the capacity to detect early-warning signals 18 to 24 months before distress events enables proactive capital-adequacy adjustments, targeted regulatory interventions, and timely debt restructuring, on the condition that the models are calibrated, explained, monitored, and governed as described in Sections 5.5 and 6. Future research should extend these architectures to graph neural networks to capture systemic contagion across financial networks, incorporate textual and alternative data, and test the models out of regime.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Appendix A. Notation

Symbol	Meaning

Table A1: Main notation

i, t, k	Firm index; quarter (forecast origin); forecast horizon in months, $k \in \{6, 12, 18, 24\}$
L	Length of the lookback window (number of consecutive quarters fed to the encoder)
$X_{i,t}, M_t$	Firm-level feature vector (5 ratios) and macro vector (3 series) at quarter t
$Y_{i,t+k}, \hat{p}_{i,t+k}$	Distress indicator at horizon k and its predicted probability
$z_{i,t}$	Sequence representation produced by the encoder and passed to the output layer
f_t, i_t, o_t, c_t, h_t	LSTM forget, input and output gates; cell state; hidden state
Q, K, V, d_k, d_{model}	Attention queries, keys, values; key dimension; model width
τ, τ^*	Decision threshold and its cost-optimal value
$C_{FN}, C_{FP}, \pi_{rate}$	Cost of a missed event, cost of a false alarm, distress base rate
$DI_t^{(k)}$	Aggregate (exposure-weighted) distress index at horizon k

Appendix B. End-to-End Computational Workflow

1. **Data assembly.** Merge quarterly firm filings with macro series using point-in-time availability dates; build labels $Y_{i,t+k}$ and censor post-event quarters.
2. **Splitting.** Create five expanding-window folds, with a purge gap of at least k months between training and validation (Figure 2).
3. **Pre-processing.** Within each fold, estimate winsorization quantiles and scaling statistics on the training part only (Eq. 19); apply them to the validation part.
4. **Windowing.** Build tensors of shape (L, 8) with stride one inside each firm's history.
5. **Hyperparameter search.** Run 100 Bayesian-optimization iterations per architecture; evaluate each configuration by mean validation AUC across folds.
6. **Training.** Minimize the multi-horizon weighted loss (Eq. 14) with AdamW, learning-rate decay, and early stopping.
7. **Calibration and thresholds.** Fit isotonic or Platt calibration on the latest validation block; choose τ^* from Eq. (20) given the user's cost ratio.
8. **Evaluation.** Report Accuracy, Precision, Recall, F1, AUC-ROC, Gini, PR-AUC, Brier,

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

and KS by horizon, with DeLong tests and firm-clustered bootstrap intervals.

9. **Explanation and monitoring.** Compute permutation importance, SHAP values and attention maps; deploy PSI drift monitoring and walk-forward retraining.

Appendix C. Reference Implementation (PyTorch)

The following compact implementation reproduces the architectures of Section 3 with the dimensions of Table 2. It is intended as a starting point for replication, not as the exact production code.

```
import torch, torch.nn as nn

class DistressRNN(nn.Module):
    """LSTM / Bi-LSTM encoder + temporal attention pooling + multi-
    horizon head."""  def _init_(self, n_feat=8, hidden=128, layers=2,
    dropout=0.25,
                bidirectional=False, n_horizons=4):
        super()._init_()
        self.rnn = nn.LSTM(n_feat, hidden, num_layers=layers,
                           dropout=dropout, batch_first=True,
                           bidirectional=bidirectional)
        d = hidden * (2 if bidirectional else 1)
        self.attn = nn.Linear(d, 1)                    # temporal attention
        scores self.head = nn.Sequential(nn.Dropout(dropout),
        nn.Linear(d, n_horizons))

    def forward(self, x):                              # x: (batch, L,
        n_feat) h, _ = self.rnn(x)                     # (batch, L, d)
        a = torch.softmax(self.attn(h).squeeze(-1), 1) # (batch, L) weights
        z = (a.unsqueeze(-1) * h).sum(1)               # (batch, d) context vector
        return self.head(z), a                         # logits for k = 6,12,18,24

class DistressTransformer(nn.Module):
    def _init_(self, n_feat=8, d_model=256, heads=8,
                layers=4, dropout=0.15, max_len=64,
                n_horizons=4):
```

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

```
        super()._init_()
        self.proj = nn.Linear(n_feat, d_model)
        self.pos = nn.Parameter(torch.zeros(1, max_len, d_model))  # learned
        positions layer = nn.TransformerEncoderLayer(d_model, heads, 4 *
        d_model, dropout,
                                                batch_first=True, norm_first=True)
        self.enc = nn.TransformerEncoder(layer, layers)
        self.head = nn.Linear(d_model, n_horizons)

    def forward(self, x):
        h = self.enc(self.proj(x) + self.pos[:, :x.size(1)])
        return self.head(h.mean(1))

# loss: weighted BCE over the four horizons (pos_weight ~ N_neg /
N_pos per horizon) # criterion =
nn.BCEWithLogitsLoss(pos_weight=pos_w)
# optimizer = torch.optim.AdamW(model.parameters(), lr=1e-3, weight_decay=1e-2)
```

References

1. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609.
2. Altman, E. I., & Hotchkiss, E. (2006). *Corporate Financial Distress and Bankruptcy* (3rd ed.). John Wiley & Sons.
3. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.
4. Basel Committee on Banking Supervision (2013). *Principles for effective risk data aggregation and risk reporting* (BCBS 239). Bank for International Settlements.
5. Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111.
6. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166.
7. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213.
8. Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 145, 103773.
9. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
10. Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, 63(6), 2899–2939.
11. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

minority over-sampling technique.
Journal of Artificial Intelligence Research, 16, 321–357.

12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD Conference*, 785–794.
13. Culkin, R., & Das, S. R. (2017). Machine learning in finance: The case of deep learning. *Journal of Investment Management*, 15(4), 55–64.
14. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2), 760–772.
15. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics*, 44(3), 837–845.
16. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
17. Estrella, A., & Mishkin, F. S. (1998). Predicting U.S. recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1), 45–61.
18. Federal Reserve Board & OCC (2011). *Supervisory Guidance on Model Risk Management* (SR 11-7).
19. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions.
European Journal of Operational Research, 270(2), 654–669.
20. Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10), 2451–2471.
21. Gilchrist, S., & Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4), 1692–1720.
22. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610.
23. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273.
24. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th ICML*, 1321–1330.
25. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.
Radiology, 143(1), 29–36.
26. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

27. Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, 117, 287–299.
28. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of NAACL-HLT*, 3543–3556.
29. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.
30. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764.
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of ICCV*, 2980–2988.
32. López de Prado, M. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons.
33. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *ICLR*.
34. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
35. Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449–470.
36. Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *Proceedings of the International Joint Conference on Neural Networks*, 163–168.
37. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.
38. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215.
39. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432.
40. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
41. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101–124.
42. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2951–2959.
43. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 1929–1958.
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information*

BULLETIN OF
CZECH ECONOMETRIC SOCIETY
www.bulletin-ces.cz
ISSN 1212-074X | VOLUME 33 | ISSUE 1 | 2026

Processing Systems, 30, 5998–6008.

45. Verbeke, W., Baesens, B., & Bravo, C. (2017). *Profit Driven Business Analytics: A Practitioner's Guide to Transforming Big Data into Added Value*. John Wiley & Sons.
46. Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *Proceedings of EMNLP-IJCNLP*, 11–20.
47. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9), 11121–11128.
48. Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59–82