EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Artificial intelligence-driven financial risk assessment: A deep learning-based credit scoring method for manufacturing enterprises

Shiliang Chang*

School of Business Administration, Linzhou College of Architectural Technology, Anyang, 456500, Henan, China

Abstract

Financial risk assessment for manufacturing enterprises supports credit-risk screening and early warning. Existing methods often fuse multi-period financial and industry information through concatenation, shared representations, or unrestricted interactions, making it difficult to separate meaningful disturbance–sensitivity correspondences from irrelevant combinations. This study proposes an industry-disturbance-conditioned credit-scoring framework. Its originality lies in explicitly matching external disturbances with firm-level financial sensitivities rather than treating them as unrestricted features. The framework decomposes financial information into levels, intertemporal changes, and accounting divergences; constructs demand, cost, and production sensitivities; and estimates conditioned exposures through a correspondence matrix and conditional gates. A matching-consistency loss constrains disturbance–sensitivity relationships, while dual-path prediction retains exposure-related and firm-specific risk information. Using Moody's Orbis and Eurostat Short-Term Business Statistics (STS), the method achieves an AUPRC of 0.512 in the full out-of-time test, exceeding TabPFN, the strongest AUPRC baseline, by 1.4 percentage points. Its AUROC is 0.879, and its FNR of 0.276 is lower than 0.289 for HGNN and 0.291 for TabPFN. Among highly sensitive firms, the proposed model achieves an AUPRC of 0.489 and an FNR of 0.288, improving on HGNN by 1.5 and 2.3 percentage points, respectively. Statistical tests, ablation studies, and repeated runs support the matching and prediction mechanisms. Independent calibration further reduces probability error and calibration bias. The framework supports relative risk ranking and distress screening during industry disturbances, although validation across additional databases and disturbance settings remains necessary.

Keywords: financial risk assessment of manufacturing enterprises, industry disturbance; financial sensitivity, conditioned exposure, matching consistency

Received on 04 August 2026, accepted on 25 August 2026, published on 31 August 2026

doi: 10.4108/eetsis.14324

1. Introduction

Fluctuations in market demand, production costs, and industry output are key factors in manufacturing financial-risk assessment [1]. Manufacturing firms maintain substantial inventories, fixed assets, and working-capital commitments [2], making distress risk dependent on financial conditions and demand declines, cost increases, or industry contraction [3]. Artificial intelligence methods using multi-period financial and industry data can support risk screening and early warning. Traditional prediction uses financial ratios, discriminant analysis, logistic regression, and scoring functions [4], while support vector machines, random forests, and boosting capture nonlinear relationships. However, single-period or independent firm-

*Corresponding author. Email: changshiliang1985@163.com

year observations cannot adequately represent declining cash flow, inventory accumulation, rising receivables, or short-term debt. Historical information has been incorporated through lagged variables, trends, distributed-lag models, and temporal networks [5,6], but many methods compress yearly variables without distinguishing financial buffers, deterioration, and disturbance-specific sensitivity [7,8]. Accounting divergences among revenue, cash flow, receivables, inventory, fixed assets, and operating returns may also be weakened.

Other studies incorporate market, macroeconomic, and industry-cycle indicators through concatenation, explicit interactions, ensembles, or attention mechanisms [9,10]. Existing approaches often integrate firm-level and industry-level information through concatenation or unrestricted interactions, which may introduce irrelevant cross-type relationships. This study therefore introduces explicit disturbance–sensitivity correspondence by organizing multi-period information into current buffers, deterioration, and accounting divergences and constructing demand, cost, and production sensitivities. A conditioned exposure mechanism restricts each disturbance to its corresponding sensitivity, while matching consistency separates same-type from incorrect combinations. Independent calibration produces relative scores for ranking and stratification rather than financial institutions' internal ratings. Out-of-time, disturbance-window, sensitive-firm, mismatched-permutation, and ablation tests evaluate the framework. The area under the precision–recall curve (AUPRC), false-negative rate (FNR), Brier score, and expected calibration error (ECE) assess ranking, missed-risk identification, probability error, and calibration bias.

The contributions are as follows: (1) A disturbance-specific encoder structures financial buffers, deterioration, and accounting divergences into demand, cost, and production sensitivities rather than merely extending historical windows or using unconstrained projections. (2) A constrained exposure mechanism restricts interactions through predefined correspondences and uses matching consistency to distinguish correct from mismatched combinations. Industry information thus generates firm-specific adjustments through financial sensitivity instead of directly entering the risk layer. (3) Out-of-time, disturbance-window, sensitive-firm, permutation, ablation, and calibration evaluations distinguish structured-matching effects from those of additional variables, model capacity, or post-processing while assessing ranking, missed-risk identification, and calibration.

2. Related Works

2.1 Static Financial-Indicator-Based Distress Prediction

Classical financial-distress prediction was established through ratio-based statistical models, including Altman's discriminant Z-score [11] and Ohlson's probabilistic bankruptcy model [12]. Recent reviews of deep learning for tabular data further summarize the development of neural architectures for structured-data prediction [13]. Financial-distress studies have subsequently employed solvency, profitability, liquidity, and operating-efficiency indicators with discriminant analysis, logistic regression, and scoring functions [14], while decision trees, support vector machines, random forests, and neural networks capture more complex relationships [15]. These studies established the principal tasks, variables, and evaluation frameworks, while statistical, tree-based, and deep tabular models provide relevant comparison points for firm-level risk prediction.

A central limitation concerns the information represented by the inputs rather than the classifiers themselves. Single-period variables capture cross-sectional financial conditions but cannot adequately represent changes in cash flow, inventory, receivables, or short-term debt [16]. Statistical and tree-based models therefore remain useful baselines for testing whether nonlinear classification alone can represent conditional risk during manufacturing disturbances.

2.2 Dynamic and Multiperiod Financial Risk Modeling

Previous studies use lagged variables, growth rates, trends, distributed-lag models, and temporal networks to represent financial changes [17,18]. These approaches demonstrate the value of historical information, but multi-period modeling alone does not necessarily distinguish economically different forms of financial change.

Dynamic models often compress yearly information into an overall representation without distinguishing current buffers, deterioration, and disturbance-specific sensitivities [19]. Accounting divergences involving revenue, cash flow, receivables, inventory, turnover efficiency, fixed assets, and operating returns are also rarely linked to specific disturbances [20]. The relevant distinction is therefore whether temporal financial information is organized into economically differentiated sensitivity representations rather than simply extending the historical window or model complexity.

2.3 Macroeconomic, Industry, and Heterogeneous Risk Information

Market, macroeconomic, and industry variables supplement firm data through concatenation, multilevel models, explicit interactions, and attention mechanisms [21]. These approaches demonstrate the value of external information, but generic interaction structures rarely establish explicit correspondences between demand, cost, and production disturbances and firm-side conditions such as inventory pressure, gross-margin capacity, cash–short-term-debt gaps, or asset structure [22,23].

Unrestricted interactions can capture complex dependencies but may also mix economically distinct disturbance and financial signals [24]. Predefined connections reduce the interaction space [25,26], yet connection restrictions alone do not establish whether corresponding disturbance–sensitivity pairs are more discriminative than mismatched combinations. This motivates comparison among structured matching, direct concatenation, tree-based interactions, unrestricted interactions, and deliberately mismatched structures.

2.4 Imbalanced Learning and Probability Reliability

Because financial-distress cases are less frequent than normal cases, resampling, class weighting, cost-sensitive learning, and ensemble methods are commonly used to improve minority-class identification [27,28]. AUPRC, Recall, F1 score, and geometric mean complement overall accuracy, while calibration and risk stratification support model application.

However, resampling may alter test-set prevalence, and a high area under the receiver operating characteristic curve (AUROC) does not ensure agreement between predicted probabilities and observed event rates. Minority-class identification and probability reliability are rarely evaluated jointly across future years, disturbance windows, and highly sensitive firms [29,30]. This study preserves the original out-of-time risk proportion and reports discrimination, missed-risk identification, and calibration metrics. Calibration supports credit-scoring output rather than introducing a new calibration theory [31].

2.5 Research Gap and Positioning of This Study

Previous research has advanced multi-period modeling, industry-information integration, firm heterogeneity analysis, and imbalanced learning [32]. The remaining gap concerns whether external disturbances can be linked to firm financial states through explicit and testable correspondences rather than through additional variables or unrestricted interaction capacity. This study addresses this gap by constructing demand, cost, and production sensitivities from structured financial information, restricting disturbances to their corresponding sensitivity channels, and using matching consistency to distinguish correct from mismatched combinations. Out-of-time, disturbance-window, sensitive-firm, permutation, ablation, and calibration evaluations then distinguish structured-matching effects from those of additional variables, model capacity, reduced connectivity, or post-processing.

3. Methodology

3.1 Problem Formulation

Let idenote a manufacturing firm, t the prediction point, s(i) its subsector, and T the historical window. The input contains financial indicators over T periods, intertemporal changes, accounting divergences, and contemporaneous industry disturbances:

$$\mathcal{I}_{i,t} = \left[X_{i,t}, \Delta X_{i,t}, A_{i,t}, D_{s(i),t}\right], p_{i,t+1} = f_{\Theta}\left(\mathcal{I}_{i,t}\right). \quad (1)$$

Here, $X_{i,t} \in \mathbb{R}^{T\times p}$ contains solvency, profitability, liquidity, operating cash flow, inventory, accounts receivable, and asset-utilization indicators. $\Delta X_{i,t}$ represents intertemporal changes, while $A_{i,t}$ captures revenue–cash-flow, revenue–accounts-receivable, inventory–turnover-efficiency, and fixed-assets–operating-return divergences. The disturbance set is $D_{s(i),t} = \left[d^{dem}_{s(i),t}, d^{cost}_{s(i),t}, d^{prod}_{s(i),t}\right]$, representing demand contraction, cost pressure, and production contraction. Moreover, $y_{i,t+1} \in \{0,1\}$ is the next-period proxy distress label, and $p_{i,t+1}$ is its predicted probability.

Type-specific masks M_k provide each sensitivity with a defined information source:

$$u^{k}_{i,t} = M_k \odot u_{i,t}, v^{k}_{i,t} = \phi_k\left(u^{k}_{i,t}; \Theta_{v,k}\right), k \in \{dem, cost, prod\}. (2)$$

Here, $u_{i,t}$ represents financial levels, intertemporal changes, and accounting divergences. The masks constrain information availability for each disturbance type, while weights, effect directions, and sensitivity levels remain data-driven.

A correspondence matrix $C_{kr} \in \{0,1\}$ is defined, where $C_{kr} = 1$ permits disturbance k to match sensitivity r, and $C_{kr} = 0$ otherwise. Firm-specific conditioned exposure is

$$e_{i,t} = \|_{k,r}\, C_{kr}\, G_{kr}\left(d^{k}_{s(i),t}, v^{r}_{i,t}\right). \quad (3)$$

The matrix constrains disturbance – sensitivity information flow. Only same-type combinations enter prediction, while cross-type combinations serve as structural negative matches in consistency learning. This differs from direct concatenation and unrestricted interactions.

The method does not estimate Bayesian uncertainty but addresses class imbalance, temporal distribution shifts, and probability reliability. It jointly optimizes prediction, matching consistency, and regularization:

$$\Theta^{*} = \arg\min_{\Theta}\left[\mathcal{L}_{pred} + \alpha\mathcal{L}_{match} + \lambda \parallel \Theta \parallel_2^2\right]. \quad (4)$$

Here, $\mathcal{L}_{pred}$ is class-weighted binary cross-entropy, with weights derived from the training distribution. $\mathcal{L}_{match}$ is defined in Section 3.4, while α and λ are the matching-loss and regularization weights. Class weighting addresses imbalance, whereas matching consistency distinguishes same-type from incorrect cross-type combinations. Outputs are calibrated on an independent period and converted into relative credit-risk scores.

3.2 Overall Framework

This study proposes an industry-disturbance-conditioned financial-sensitivity credit-scoring framework to

distinguish valid correspondences from incorrect cross-type matches under additive inputs or unrestricted interactions. Using multi-period financial states, accounting divergences, and subsector-level demand, cost, and production information, it generates a baseline risk representation, conditioned exposure, and relative risk score. Figure 1 presents the architecture, while Figures 2 and 3 show the two core modules.

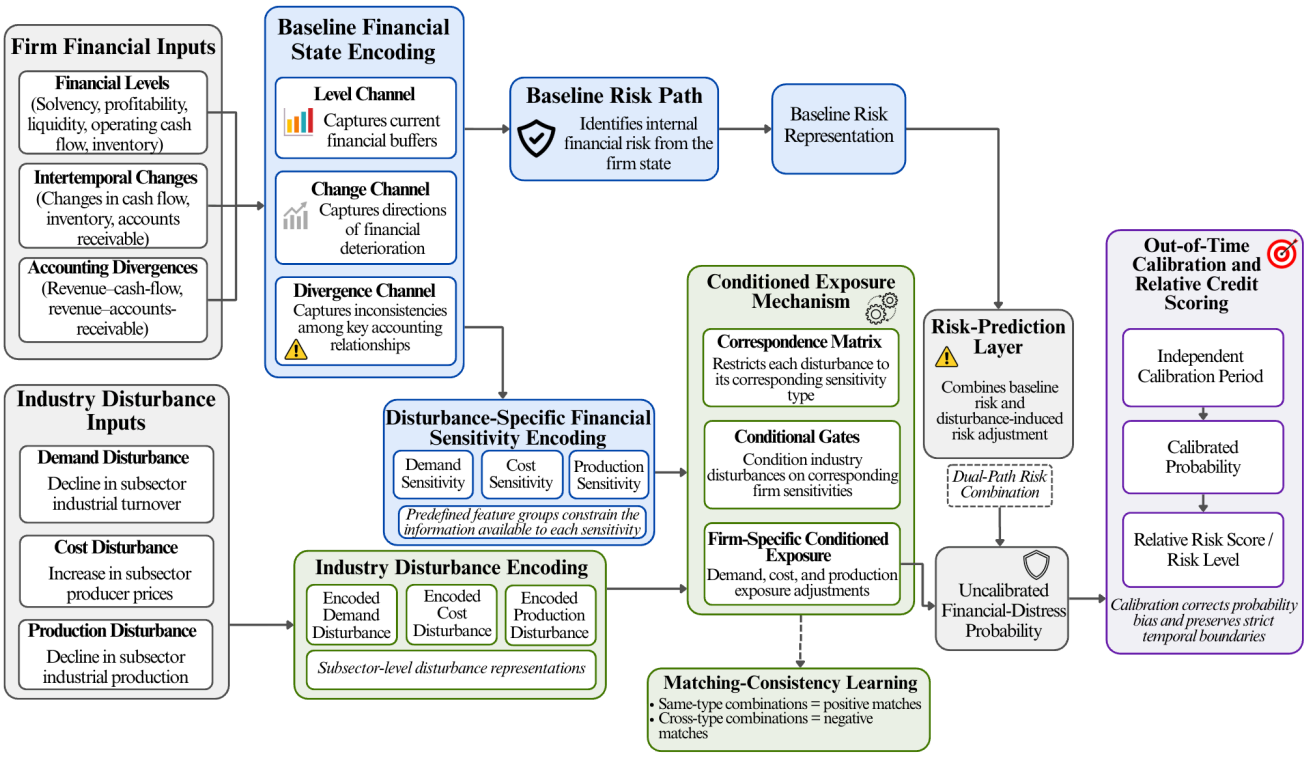

Figure 1. Overall architecture of the proposed credit-scoring framework

The framework contains baseline-risk and conditioned-exposure paths. The baseline path uses financial levels, intertemporal changes, and accounting divergences to represent current buffers, deterioration, and inconsistencies among key financial relationships. Based on predefined feature groups, the sensitivity encoder maps firm states into demand, cost, and production sensitivities.

The exposure path encodes three subsector-level disturbances and restricts each to its corresponding sensitivity block through the correspondence matrix. Industry information enters prediction only after being conditioned on firm sensitivity. Same-type and cross-type combinations act as structural positive and negative matches, and matching consistency strengthens their separation.

Finally, the prediction layer combines the baseline state and conditioned exposure to estimate uncalibrated distress probability. Independent calibration then corrects probability bias and generates the relative risk score. The dual-path structure distinguishes internal financial risk from external industry exposure while linking them through constrained information flow and matching-consistency learning.

3.3 Disturbance-Specific Financial Sensitivity Encoding

As shown in Figure 2, this module separates financial levels, intertemporal changes, and accounting divergences to construct type-specific sensitivities. The level and change channels are

$$h_{i,t}^{L} = f_L(X_{i,t}; \Theta_L), h_{i,t}^{\Delta} = f_{\Delta}(\Delta X_{i,t}; \Theta_{\Delta}). \quad (5)$$

Here, f_L and f_{Δ} are independent temporal encoders. The former represents solvency, profitability, liquidity, and operational buffers, while the latter captures changes in cash flow, inventory, accounts receivable, short-term debt, and gross margin.

Accounting divergences are encoded as

$$h_{i,t}^{A} = f_A(A_{i,t}; \Theta_A). \quad (6)$$

Here, $A_{i,t}$ contains revenue–cash-flow, revenue–accounts-receivable, inventory–turnover-efficiency, and fixed-assets–operating-return divergences. The baseline firm state is $u_{i,t} = h_{i,t}^{L} \| h_{i,t}^{\Delta} \| h_{i,t}^{A}$. To prevent the sensitivities from differing only through separate projection heads, fixed type-specific masks are applied before training:

$$\tilde{u}_{i,t}^{k} = M_k \odot u_{i,t}, v_{i,t}^{k} = \phi_k(\tilde{u}_{i,t}^{k}; \Theta_v^{k}), k \in \{dem, cost, prod\}. (7)$$

Specifically, M_{dem} retains inventory, accounts receivable, turnover efficiency, and fixed-asset intensity; M_{cost} retains gross margin, operating cash flow, and cost-related variables; and M_{prod} retains the cash–short-term-debt gap, liquidity, fixed assets, and operating leverage. These masks constrain information availability rather than predetermining variable weights, effect directions, or sensitivity levels, which remain learned from data.

The groupings reflect the financial channels through which each disturbance appears in firm accounts. Demand sensitivity uses inventory, receivables, turnover efficiency, and fixed-asset intensity to capture slower turnover, collection, and asset utilization. Cost sensitivity uses gross margin, operating cash flow, and cost-related variables to represent margin and cash-flow pressure. Production sensitivity uses the cash–short-term-debt gap, liquidity, fixed assets, and operating leverage to capture fixed commitments and short-term financing pressure during contraction. The masks define admissible information sources only; variable contributions and directions remain data-driven. Their dependence on manual grouping is further evaluated through mask-sensitivity experiments.

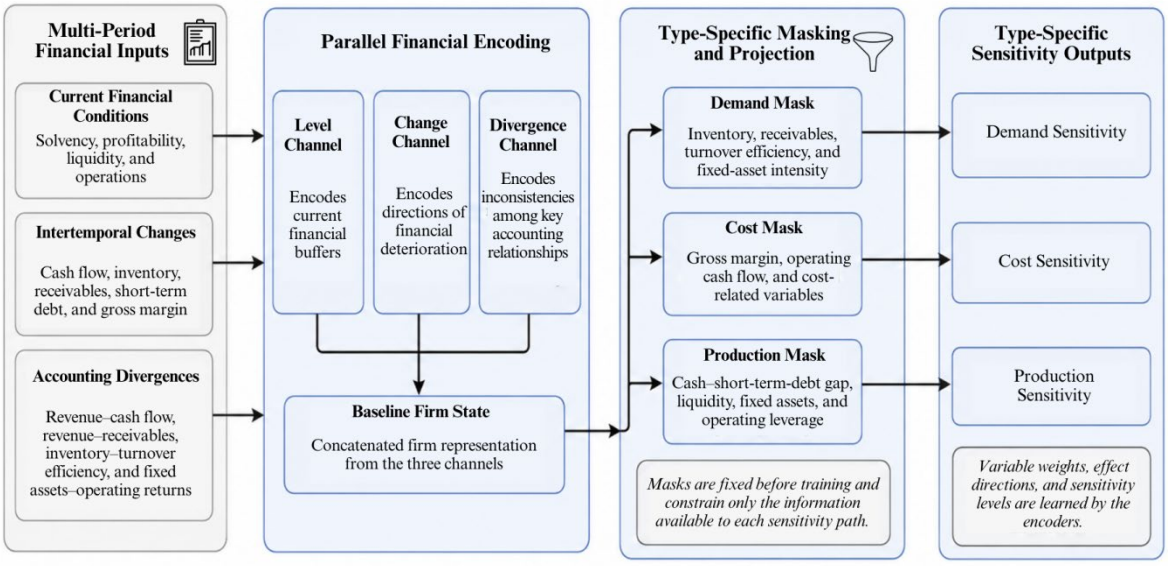

Figure 2. Disturbance-specific financial sensitivity encoding module

The module derives demand, cost, and production sensitivities from defined information sources rather than applying arbitrary projections to a shared financial state, providing firm-side representations for structured matching.

3.4 Disturbance-Conditioned Exposure Mechanism

As shown in Figure 3, this module constrains disturbance–sensitivity interactions through predefined correspondences and separates correct from incorrect combinations. Disturbance k is first encoded into a d_e-dimensional space:

$$q_{s(i),t}^{k} = \psi_k(d_{s(i),t}^{k}; \Theta_d^{k}), q_{s(i),t}^{k} \in \mathbb{R}^{d_e}. \quad (8)$$

Using the correspondence matrix C_{kr}, the matching term and element-wise conditional gate are

$$m_{i,t}^{(k,r)} = C_{kr}(q_{s(i),t}^{k} \odot v_{i,t}^{r}), g_{i,t}^{(k,r)} = \sigma\left(W_g^{(k,r)}\left[q_{s(i),t}^{k} \| v_{i,t}^{r} \| m_{i,t}^{(k,r)}\right] + b_g^{(k,r)}\right). \quad (9)$$

Here, $C_{kr} = 1$ permits a match, whereas $C_{kr} = 0$ excludes it. The main model retains demand–demand, cost–cost, and production–production matches, while cross-type combinations serve as incorrect-matching contrasts.

Because the correspondence matrix restricts pathways but does not ensure that same-type matches are more discriminative, the normalized matching score is defined as

$$s_{i,t}^{(k,r)} = \frac{(q_{s(i),t}^{k})^{\top} v_{i,t}^{r}}{\|q_{s(i),t}^{k}\|_2 \|v_{i,t}^{r}\|_2 + \varepsilon}. \quad (10)$$

The case $r = k$ is a structural positive match, while $r \neq k$ is a structural negative match. The matching-consistency loss is

$$\mathcal{L}_{match} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k}\sum_{r\neq k} \max\left(0, \gamma - s_{i,t}^{(k,k)} + s_{i,t}^{(k,r)}\right) (11)$$

Here, $\gamma > 0$ is the matching margin, and ε ensures numerical stability. The loss requires each same-type score to exceed incorrect cross-type scores by at least γ.

The final exposure representation is

$$e_{i,t} = \|_{k\in\{dem,cost,prod\}} \sum_{r}\left(g_{i,t}^{(k,r)} \odot m_{i,t}^{(k,r)}\right). \quad (12)$$

Industry disturbances therefore reach the prediction layer only after being matched with the corresponding firm sensitivities and modulated by conditional gates.

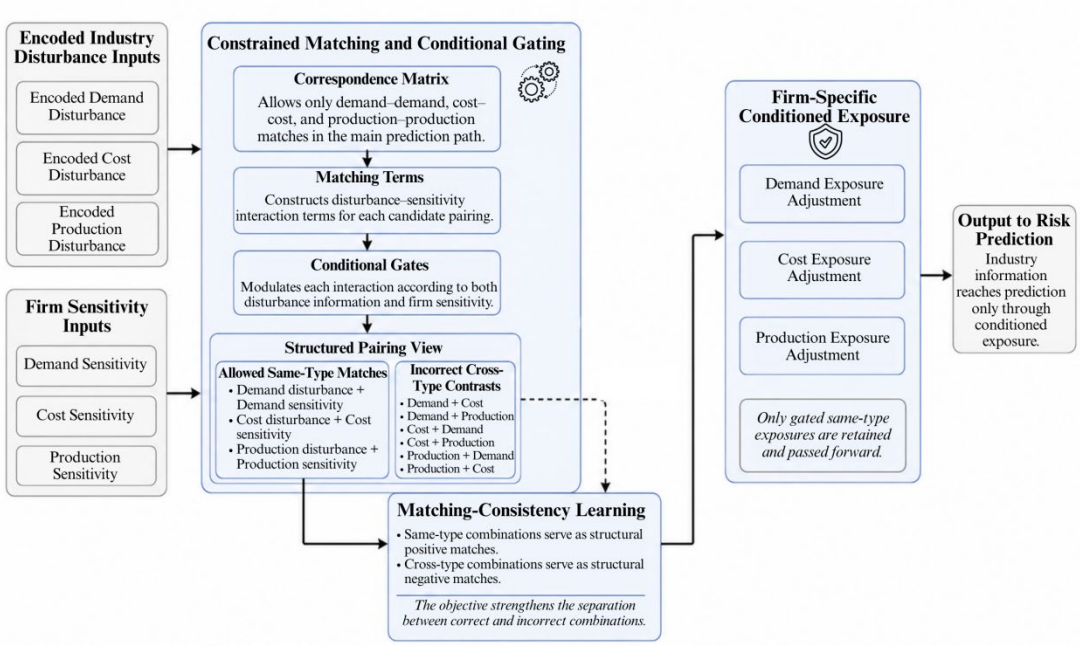

Figure 3. Industry-disturbance–firm-sensitivity conditioned exposure module

The contribution lies not in sigmoid gating alone but in constraining information flow through the correspondence matrix and separating same-type from cross-type combinations through matching consistency. This differs from direct concatenation, unrestricted interactions, and unconstrained attention-based fusion.

3.5 Joint Financial Risk Prediction and Optimization

The model uses baseline-risk and disturbance-adjustment paths to prevent conditioned exposure from being diluted by conventional fusion. The baseline path is

$$u_{i,t} = h^{L}_{i,t} \parallel h^{\Delta}_{i,t} \parallel h^{A}_{i,t}, \ell^{base}_{i,t} = w_b^{\top} f_b(u_{i,t}; \Theta_b) + b_b. \quad (13)$$

Here, $\ell^{base}_{i,t}$ represents internal financial risk. The exposure path uses only the conditioned exposure from Section 3.4:

$$\ell^{exp}_{i,t} = w_e^{\top} f_e(e_{i,t}; \Theta_e) + b_e, \hat{p}_{i,t+1} = \sigma\left(\ell^{base}_{i,t} + \ell^{exp}_{i,t}\right). \quad (14)$$

Thus, industry information adjusts risk only through structured matching and cannot bypass firm sensitivity. Ablation experiments evaluate both paths and their combination.

Given the limited distress cases, class-weighted binary cross-entropy is used:

$$\mathcal{L}_{pred} = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i \log \hat{p}_i + w_0(1-y_i)\log(1-\hat{p}_i)], w_c = \frac{N}{2N_c} \quad (15)$$

where N_c is the number of class-c training samples. The final objective is

$$\mathcal{L} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{match} + \lambda \parallel \Theta \parallel_2^2. \quad (16)$$

Here, α and λ control matching consistency and regularization, while Θincludes all model parameters. The objective jointly learns baseline risk, conditioned exposure, and disturbance–sensitivity correspondence.

3.6 Out-of-Time Calibration and Relative Credit Scoring

Strong ranking does not ensure agreement between predicted probabilities and observed distress rates. After training, Θ^* is fixed, and calibration parameters are estimated from an independent set $\mathcal{D}_{cal}$. Validation data support only early stopping, while calibration and test data do not update the model.

For model logit ℓ_i, temperature scaling gives

$$\tilde{p}_i = \sigma\left(\frac{\ell_i}{\tau}\right), \tau^* = \arg\min_{\tau>0}\left\{-\sum_{i\in\mathcal{D}_{cal}}[y_i \log \tilde{p}_i + (1-y_i)\log(1-\tilde{p}_i)]\right\} \quad (17)$$

The calibration-set temperature adjusts probability scale without changing rankings or using test information. Relative risk is then

$$R_{i,t+1} = 100\hat{F}_{cal}\left(\tilde{p}_{i,t+1}\right), R_{i,t+1} \in [0,100]. \quad (18)$$

Here, $\hat{F}_{cal}$ denotes the empirical cumulative distribution function estimated exclusively from calibrated probabilities in $\mathcal{D}_{cal}$. It is fixed before out-of-time testing, and no test observations are used to construct the percentile mapping or risk thresholds. Higher values indicate greater relative risk, with calibration-set percentile thresholds fixed during testing. This process separates risk learning, calibration, and stratification rather than proposing a new calibration method.

The operating threshold used for FNR evaluation is likewise selected on the calibration set and remains fixed during testing. Under a prevalence shift, the threshold is not automatically re-estimated, so the resulting FNR and precision may change with the test distribution.

Deployment under persistent prevalence changes therefore requires threshold monitoring and, when necessary, periodic recalibration using newly available historical data.

3.7 Optimization and Stability Analysis

The model's nonlinear encoding, constraints, gating, and matching consistency produce a nonconvex objective; global convergence is therefore not claimed. The analysis covers only exposure boundedness and local perturbations, not Bayesian posterior uncertainty.

Assume standardized and winsorized inputs satisfying $\| q^{k}_{s(i),t} \|_2 \le B_Q, \| v^{k}_{i,t} \|_2 \le B_V$. Because $0 < g^{k}_{i,t} < 1$,

$$\| e^{k}_{i,t} \|_2 = \| g^{k}_{i,t} \odot q^{k}_{s(i),t} \odot v^{k}_{i,t} \|_2 \le B_Q B_V. \quad (19)$$

Thus, the gate cannot amplify disturbance–sensitivity interactions without bound. If both prediction paths are locally Lipschitz,

$$| \ell(u + \delta_u, e + \delta_e) - \ell(u, e) | \le L_b \| \delta_u \|_2 + L_e \| \delta_e \|_2. \quad (20)$$

Bounded inputs and parameters therefore yield bounded local output changes. These results characterize local stability rather than constitute an independent theoretical contribution.

3.8 Training, Calibration, and Inference Algorithm

Algorithm 1 summarizes training, independent calibration, and out-of-time inference, including preprocessing, financial encoding, sensitivity construction, matching consistency, exposure computation, dual-path prediction, calibration, and scoring. The early-stopping set supports model selection, the calibration set determines temperature and risk thresholds, and the test set determines no parameters.

Algorithm 1. Training, Calibration, and Out-of-Time Inference of the Proposed Framework

1: Input: $D_{tr}, D_{es}, D_{cal}, D_{te}, \{M_k\}, C, E, P, w_0, w_1, \alpha, \gamma, \varepsilon, \lambda$
2: Output: $\hat{p}, \tilde{p}, R_{risk\_level}, e$
3: Fit preprocessing parameters on D_{tr} and apply them to all data splits
4: Initialize Θ, best_AUPRC $\leftarrow -\infty$, patience $\leftarrow 0$
5: for epoch $\leftarrow 1$ to E do
6: for each mini-batch $B \in D_{tr}$ do
7: Construct $X, \Delta X, A$; compute h_L, h_Δ, h_A, u
8: Compute $v^k \leftarrow \phi_k(M_k \odot u)$ and $q^k \leftarrow \psi_k(d^k)$
9: Compute $s^{(k,r)}$ and obtain L_{match} with margin γ
10: Compute $m^{(k,r)}, g^{(k,r)}, e$ according to C
11: Compute l_{base}, l_{exp}, and $\hat{p} \leftarrow \mathrm{Sigmoid}(l_{base} + l_{exp})$
12: $L \leftarrow L_{pred} + \alpha L_{match} + \lambda \| \Theta \|_2^2$; update Θ
13: end for
14: Evaluate AUPRC on D_{es}; update Θ_{best} and patience
15: if patience $\ge P$ then break
16: end for
17: Fit τ^* and fixed risk thresholds on D_{cal} using Θ_{best}
18: Infer $\hat{p}$ and e on D_{te}; compute $\tilde{p}, R_{risk\_level}$
19: return $\hat{p}, \tilde{p}, R_{risk\_level}, e$

Line 9 calculates same-type and cross-type scores. Cross-type combinations serve only as negative matches in $\mathcal{L}_{match}$ and do not enter prediction. Line 10 computes permitted exposures according to C.

4. Experiments and Analysis

4.1 Dataset and Experimental Settings

4.1.1 Dataset Construction

This study combines firm-level Moody's Orbis data with Eurostat Short-Term Business Statistics (STS). Orbis provides financial statements, firm status, event dates, and Statistical Classification of Economic Activities in the European Community Revision 2 (NACE Rev. 2) classifications for constructing financial features and next-year distress labels. STS provides turnover, producer price index (PPI), and production data for measuring demand, cost, and production disturbances. Table 1 summarizes both sources. Orbis data were obtained through authorized subscription-based access to the Moody's Orbis platform and used in accordance with the provider's licensing terms. Eurostat STS data were downloaded from the publicly accessible Eurostat Data Browser under its data reuse policy. Orbis records were extracted at the firm–fiscal-year level, while STS observations were extracted by country, two-digit NACE Rev. 2 industry, and month for 2012–2023. The two sources were used to construct firm-level financial states and distress labels and industry-level disturbance indicators, respectively.

Processing comprises six steps. First, Section C firms with at least three consecutive years of core records are retained. Duplicate firm-year statements are resolved by prioritizing consolidated and latest versions and deduplicating Orbis identifiers. Second, fiscal year-end t is the prediction point. A firm receives $y_{i,t+1} = 1$ if it first enters bankruptcy, insolvency, or compulsory liquidation in $t+1$; otherwise, $y_{i,t+1} = 0$. Voluntary liquidation, dormancy, unspecified deregistration, and previously abnormal firms are excluded from the positive class.

Third, only STS observations released before prediction are used. The latest 12 months are transformed into year-on-year disturbances: $d^{dem}_{s,t} = \max[0, -z_{tr}(\Delta_{12}\mathrm{Turnover}_{s,t})]$, $d^{cost}_{s,t} = \max[0, z_{tr}(\Delta_{12}\mathrm{PPI}_{s,t})]$, $d^{prod}_{s,t} = \max[0, -z_{tr}(\Delta_{12}\mathrm{Production}_{s,t})]$. Here, $z_{tr}(\cdot)$ uses training statistics. The training-period 75th percentile defines disturbance windows, while the 70th and 80th percentiles support sensitivity analysis. Later releases and revisions are excluded.

Fourth, records are matched by country, two-digit NACE industry, and fiscal year, using the primary industry for multi-industry firms. Fifth, monetary variables undergo $\log(1+x)$ transformation and training-set 1st–99th percentile winsorization. Single-period missing values use industry-year or training-period industry medians; firms missing over two consecutive periods are removed. Sixth, training-set standardization parameters remain fixed across later splits.

Highly sensitive firms are identified independently of the model. Demand, cost, and production indices use

inventory and receivables, gross margin and cash flow, and fixed-asset intensity and the cash–short-term-debt gap, respectively. Firms above the training-period 75th percentile of the composite index are classified as highly sensitive.

Table 1. Statistics of the two data sources

Item	Moody's Orbis	Eurostat STS
Data level	Firm–year	Country–industry–month
Study period	2012–2023	2012–2023
Study population	NACE Rev.2 Section C manufacturing firms	NACE Rev.2 two-digit manufacturing industries
Experimental sample	18,642 firms; 92,810 firm–year observations	24 industries; three disturbance indicators
Main variables	Revenue, cash flow, inventory, receivables, short-term debt, fixed assets, and gross margin	Turnover, producer prices, and industrial production
Data use	Financial states, intertemporal changes, accounting divergences, and distress labels	Demand, cost, and production disturbances
Distress-event share	4.7%	——

Note: Sample sizes are reported after screening, deduplication, and matching.

4.1.2 Baselines

Table 2 compares nine baselines covering statistical prediction, nonlinear fitting, temporal encoding, unrestricted interaction, and heterogeneous graph learning: Logistic Regression, Extreme Gradient Boosting (XGBoost), Gated Recurrent Unit–Industry (GRU-Industry), Deep & Cross Network Version 2 (DCN-V2), TabR, GAFormer, Modern Temporal Convolutional Network (ModernTCN), Tabular Prior-data Fitted Network (TabPFN), and a Heterogeneous Graph Neural Network (HGNN). The HGNN is implemented with HGT-style type-aware message passing [33], using firms and two-digit NACE industries as heterogeneous nodes. Firm nodes contain the same financial representations used by the other baselines, industry nodes contain the corresponding STS disturbance variables, and firm–industry edges represent primary-industry membership. Firm embeddings after message passing are used for distress classification. The proposed model instead uses disturbance-specific sensitivities, conditioned exposure, and matching consistency.

All methods use identical variables, temporal splits, and comparable validation-based tuning budgets. TabR retrieves only training observations, while GAFormer and ModernTCN use three-period sequences. The HGNN uses the same firm-year samples and train, validation, calibration, and test periods as the other methods, ensuring that its graph structure provides the additional relational inductive bias. Direct concatenation, unrestricted interaction, and correspondence-matrix removal are tested in Section 4.7.

Table 2. Baseline models and main characteristics

Method	Model Category	Main Characteristic
Logistic Regression [14,15]	Statistical model	Additive financial and industry inputs
HGNN (HGT-based) [33]	Heterogeneous graph model	Firm–industry type-aware message passing
XGBoost [34]	Tree-based model	Nonlinear and high-order interactions
GRU-Industry [17]	Recurrent temporal model	Financial-sequence encoding and industry concatenation
DCN-V2 [35]	Deep interaction model	Explicit feature crossing
TabR [36]	Retrieval-augmented model	Similar-firm retrieval
GAFormer [37]	Temporal Transformer	Data-driven variable grouping
ModernTCN [38]	Temporal convolution model	Convolutional temporal encoding
TabPFN [39]	Tabular foundation model	Pretrained tabular classification
Proposed Model	Structured risk model	Disturbance-specific sensitivity and matching consistency

4.1.3 Implementation Details

Experiments use an Intel i7-12700K processor, 32 GB memory, and an NVIDIA RTX 3060 graphics processor. Data from 2012–2018 support training, 2019–2020 validation, 2021 calibration, and 2022–2023 out-of-time testing. Training data determine preprocessing, validation data support model selection, and calibration data

determine temperature scaling and decision boundaries. Test data determine no parameters.

Country-held-out evaluation uses the same temporal boundaries. Each run excludes one country meeting sample-size and distress-event requirements, with results macro-averaged across countries. This evaluates cross-country performance within the Orbis–STS panel, not external-database generalization.

All deep models use Adam, class-weighted binary cross-entropy, and ten seeds. Results are mean ± standard deviation, with early stopping after 15 epochs without validation AUPRC improvement. The HGNN uses two heterogeneous message-passing layers, four attention heads, a hidden dimension of 64, and dropout of 0.2 under the same validation-based tuning protocol. Table 3 summarizes the settings.

For the proposed model, the trainable parameter count is 0.82 million and the serialized model size is 3.3 MB. On the stated RTX 3060 platform, mean training time is 18.4 ± 0.7 min per run, and average out-of-time inference latency is 0.18 ms per firm-year at batch size 128. These measurements are averaged over the same ten runs used for performance evaluation.

Table 3. Main experimental parameter settings

Parameter	Setting
Historical window (T)	3 years
Batch size	128
Optimizer	Adam
Learning rate	1×10^{-3}
Hidden dimension	64
Maximum epochs	150
Early-stopping patience	15
Dropout rate	0.2
Matching-loss weight (α)	0.1
Matching margin (γ)	0.2
Regularization coefficient (λ)	1×10^{-4}
ECE bins	10 equal-frequency bins
Random runs	10

4.1.4 Evaluation Metrics

AUPRC and FNR remain the primary metrics, while AUROC is additionally reported to facilitate comparison with prior financial-distress studies. AUPRC evaluates distressed-firm ranking under class imbalance, whereas AUROC measures the probability that a randomly selected distressed firm receives a higher risk score than a randomly selected non-distressed firm.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP+FN}}. \quad (21)$$

FNR measures the proportion of distressed firms classified as normal:

$$\text{FNR} = \frac{\text{FN}}{\text{TP+FN}}. \quad (22)$$

Higher AUPRC and AUROC and lower FNR indicate better discrimination and distress identification. FNR in Tables 4–6 uses the F1-maximizing calibration threshold, fixed during testing.

The Brier score and Expected Calibration Error (ECE) assess calibrated probabilities p_i:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \quad (23)$$

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\,|\,\text{acc}(B_m) - \text{conf}(B_m)\,|. \quad (24)$$

Here, B_m is an equal-frequency bin, while $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are its observed distress rate and mean predicted probability. Lower values indicate smaller probability errors and calibration deviations.

4.2 Main Experimental Results

Ten models are evaluated on the full Moody's Orbis–Eurostat STS out-of-time panel and under a country-held-out protocol, including the newly added HGNN baseline. Results are reported as mean ± standard deviation over ten identical random seeds. Tables 4–6 report the full-panel, country-held-out, industry-disturbance, and highly sensitive firm results. FNR thresholds are fixed on the calibration set, while subgroup boundaries use only training-period statistics.

Table 4 shows that the proposed model achieves an AUPRC of 0.512 and an AUROC of 0.879, compared with 0.498 and 0.867 for TabPFN and 0.493 and 0.862 for HGNN. Its FNR of 0.276 is also lower than 0.291 for TabPFN and 0.289 for HGNN. Because distressed observations account for only 4.7% of the sample, the AUPRC of 0.512 should be interpreted relative to this low positive prevalence and the strong out-of-time baselines rather than as an absolute performance level. A no-skill precision–recall ranking has expected precision close to the 0.047 event prevalence, while an all-negative decision rule has zero recall and an FNR of 1.000. The observed results therefore indicate meaningful ranking and missed-risk improvements under substantial class imbalance.

At the fixed calibration threshold, reducing FNR from 0.291 for TabPFN to 0.276 corresponds to approximately 1.5 fewer missed cases per 100 distressed firms. Under the country-held-out protocol in Table 5, the proposed model obtains an AUPRC/AUROC of 0.489/0.861, compared with 0.472/0.849 for ModernTCN and 0.468/0.846 for HGNN. The corresponding FNR decreases from 0.313 for ModernTCN to 0.295, equivalent to approximately 1.8 fewer missed cases per 100 distressed firms. TabPFN obtains slightly lower Brier scores and ECE values in both settings. The proposed model's advantage therefore lies primarily in ranking and missed-risk identification rather than uniformly dominating every calibration metric

Table 4. Overall performance on the Moody's Orbis–Eurostat STS full Out-of-Time panel

Model	AUPRC ↑	AUROC ↑	FNR ↓	Brier ↓	ECE ↓
Logistic Regression	0.421±0.009	0.801±0.006	0.362±0.012	0.089±0.003	0.041±0.004
XGBoost	0.468±0.008	0.838±0.005	0.327±0.010	0.081±0.002	0.033±0.003
GRU-Industry	0.476±0.010	0.845±0.006	0.315±0.011	0.080±0.003	0.031±0.003
DCN-V2	0.481±0.009	0.851±0.005	0.309±0.010	0.079±0.002	0.029±0.003
TabR	0.492±0.008	0.859±0.005	0.298±0.009	0.077±0.002	0.027±0.002
GAFormer	0.488±0.009	0.856±0.006	0.302±0.010	0.078±0.002	0.028±0.002
ModernTCN	0.495±0.007	0.864±0.004	0.294±0.008	0.076±0.002	0.026±0.002
TabPFN	0.498±0.008	0.867±0.005	0.291±0.009	0.074±0.002	0.023±0.002
HGNN (HGT-based)	0.493±0.009	0.862±0.005	0.289±0.010	0.076±0.003	0.026±0.003
Proposed Model	0.512±0.007	0.879±0.004	0.276±0.008	0.075±0.002	0.025±0.002

Table 5. Overall performance under the Country-Held-Out evaluation protocol

Model	AUPRC ↑	AUROC ↑	FNR ↓	Brier ↓	ECE ↓
Logistic Regression	0.398±0.011	0.783±0.007	0.389±0.014	0.095±0.004	0.047±0.005
XGBoost	0.442±0.010	0.817±0.006	0.352±0.012	0.087±0.003	0.039±0.004
GRU-Industry	0.451±0.011	0.826±0.007	0.341±0.013	0.086±0.003	0.037±0.004
DCN-V2	0.455±0.010	0.831±0.006	0.336±0.011	0.085±0.003	0.035±0.003
TabR	0.466±0.009	0.841±0.005	0.322±0.010	0.083±0.003	0.032±0.003
GAFormer	0.470±0.010	0.844±0.006	0.317±0.011	0.082±0.003	0.031±0.003
ModernTCN	0.472±0.009	0.849±0.005	0.313±0.010	0.081±0.002	0.030±0.003
TabPFN	0.469±0.010	0.847±0.006	0.319±0.011	0.079±0.002	0.027±0.002
HGNN (HGT-based)	0.468±0.010	0.846±0.006	0.316±0.011	0.082±0.003	0.031±0.003
Proposed Model	0.489±0.008	0.861±0.005	0.295±0.009	0.080±0.002	0.029±0.002

As shown in Table 6, industry disturbances generally reduce AUPRC and increase FNR across models. The HGNN provides a competitive relational baseline but remains below the proposed model under demand and production disturbances and among highly sensitive firms. The proposed model achieves higher AUPRC and lower FNR under demand disturbance, production disturbance, and the highly sensitive firm setting. Among highly sensitive firms, its AUPRC of 0.489 exceeds 0.474 for HGNN and 0.471 for ModernTCN, while its FNR of 0.288 is lower than 0.311 and 0.314, respectively. Relative to ModernTCN, this corresponds to approximately 2.6 fewer missed cases per 100 distressed firms. Under cost disturbance, TabPFN has a 0.3-percentage-point AUPRC advantage, whereas the proposed model achieves lower FNR. Its benefit in this setting is therefore mainly fewer missed distressed firms.

The screening value is therefore reflected in higher-risk ranking and fewer missed distressed firms at a fixed operating threshold. Because the available data do not contain loan-level exposure, loss-given-default, or realized credit-loss amounts, monetary loss prevention is not estimated directly; the operational comparison is expressed through AUPRC, AUROC, and missed-risk reduction.

Table 6. Performance under Industry-Disturbance and High-Sensitivity Settings on the full Out-of-Time panel

Model	Demand AUPRC ↑	Demand FNR ↓	Cost AUPRC ↑	Cost FNR ↓	Production AUPRC ↑	Production FNR ↓	High-Sensitivity AUPRC ↑	High-Sensitivity FNR ↓
Logistic Regression	0.382±0.012	0.411±0.015	0.371±0.013	0.426±0.016	0.365±0.012	0.438±0.015	0.398±0.011	0.397±0.014
XGBoost	0.421±0.011	0.374±0.013	0.414±0.012	0.389±0.014	0.405±0.011	0.401±0.013	0.438±0.010	0.356±0.012
GRU-Industry	0.432±0.012	0.361±0.014	0.421±0.013	0.377±0.014	0.417±0.012	0.389±0.014	0.447±0.011	0.344±0.013
DCN-V2	0.437±0.011	0.354±0.013	0.429±0.011	0.369±0.013	0.423±0.011	0.381±0.013	0.452±0.010	0.337±0.012
TabR	0.449±0.010	0.339±0.012	0.441±0.010	0.351±0.012	0.436±0.010	0.366±0.012	0.466±0.009	0.321±0.011

GAFormer	0.445±0.011	0.343±0.012	0.447±0.010	0.346±0.012	0.441±0.010	0.358±0.011	0.462±0.010	0.326±0.011
ModernTCN	0.454±0.009	0.331±0.011	0.449±0.009	0.338±0.011	0.448±0.009	0.349±0.010	0.471±0.008	0.314±0.010
TabPFN	0.451±0.010	0.335±0.011	0.455±0.009	0.341±0.011	0.445±0.010	0.353±0.011	0.468±0.009	0.318±0.010
HGNN (HGT-based)	0.456±0.010	0.328±0.011	0.450±0.010	0.343±0.012	0.452±0.009	0.344±0.011	0.474±0.009	0.311±0.010
Proposed Model	0.469±0.009	0.309±0.010	0.452±0.009	0.321±0.010	0.463±0.008	0.327±0.009	0.489±0.008	0.288±0.009

Figure 4 shows clearer precision–recall advantages under demand and production disturbances, with performance close to TabPFN under cost disturbance. Figure 5 confirms that precision and recall differences persist across a broad threshold range. In Figure 6, total and matching-consistency losses decrease progressively, while validation AUPRC stabilizes before early stopping. Calibration moves the reliability curve toward the ideal diagonal, although TabPFN remains slightly better calibrated.

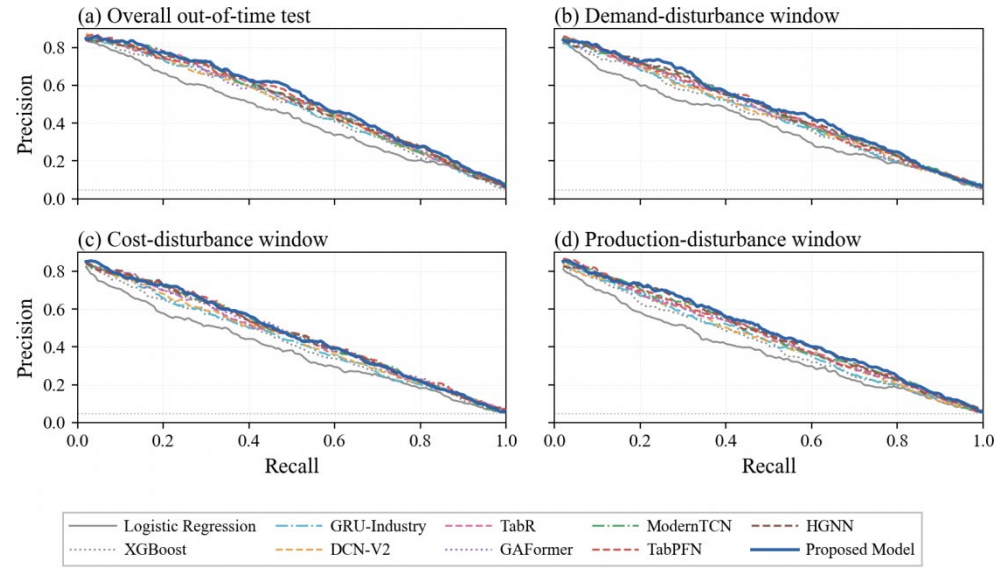

Figure 4. Precision–Recall curves of the proposed model and baseline methods on the 2022–2023 out-of-time test set under overall and industry-disturbance settings. (a) Overall test set; (b) demand-disturbance window; (c) cost-disturbance window; (d) production-disturbance window

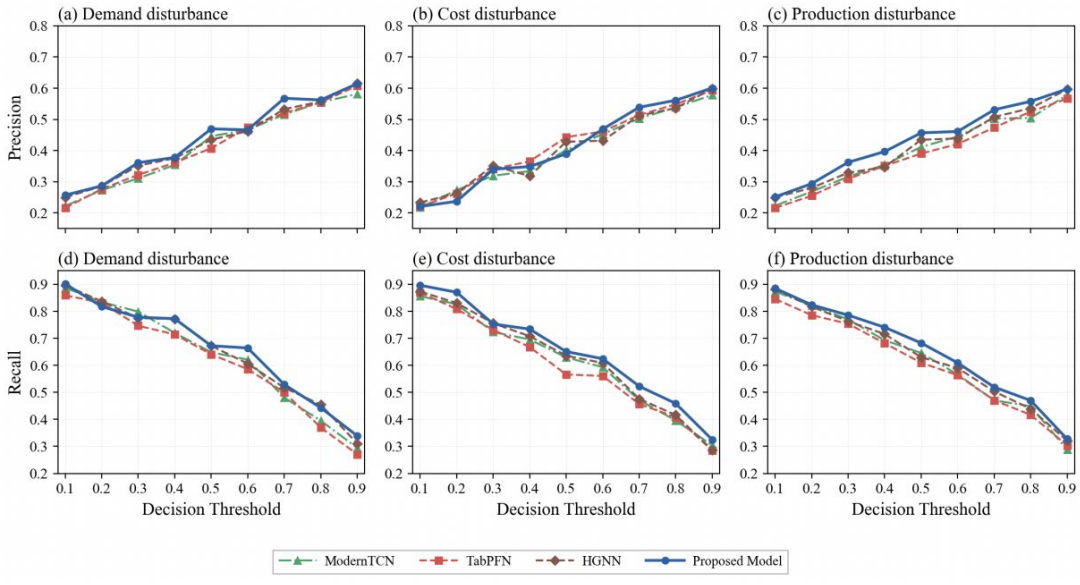

Figure 5. Threshold-dependent precision and recall of the proposed model and the strongest competing baselines under different industry disturbances on the 2022–2023 out-of-time test set. (a–c) Precision under demand, cost, and production disturbances; (d–f) recall under the corresponding disturbances

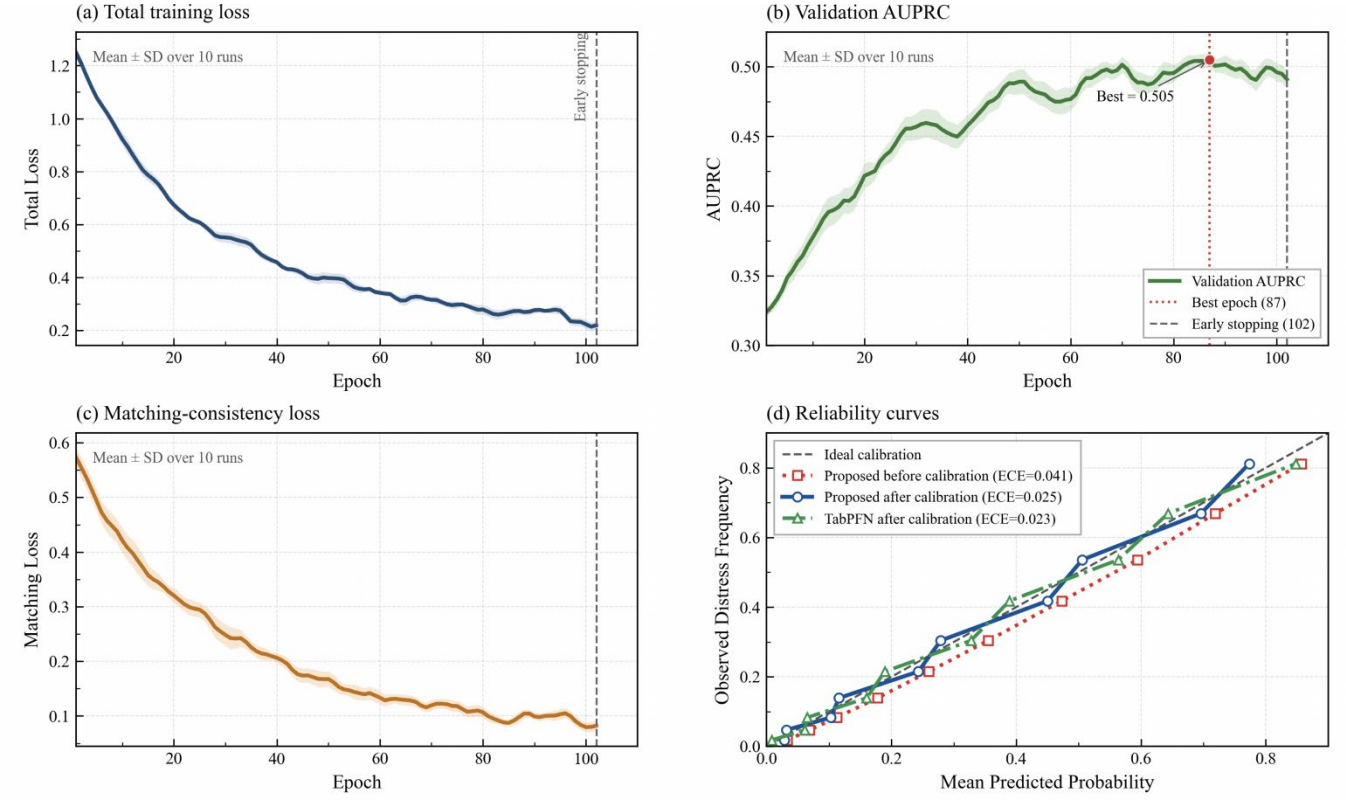

Figure 6. Training convergence, matching-consistency optimization, and probability calibration of the proposed model. (a) Total training loss; (b) validation AUPRC; (c) matching-consistency loss; (d) reliability curves before and after temperature calibration

4.3 Statistical Significance Analysis

The proposed model is compared with TabPFN, ModernTCN, and TabR to determine whether its AUPRC improvements reflect random variation. The same baselines, data splits, and ten random seeds are used across all settings. Run-level results are assessed using paired two-sided t-tests at α=0.05, with Holm correction across nine comparisons. Combined-disturbance AUPRC is the mean across demand, cost, and production disturbances. Table 7 reports paired mean differences, 95% confidence intervals (CIs), and test statistics computed from run-level results rather than summary means.

Table 7. Statistical significance against strong baselines

Comparison	Evaluation Setting	Mean Difference	95% CI	t-value	Cohen's d_z	Raw p	Holm-adjusted p
Proposed vs TabPFN	Full out-of-time test	0.014	[0.006, 0.022]	3.96	1.252	0.0033	0.0099
Proposed vs ModernTCN	Full out-of-time test	0.017	[0.009, 0.025]	4.81	1.521	0.0010	0.0060

Proposed vs TabR	Full out-of-time test	0.020	[0.011, 0.029]	5.03	1.591	0.0007	0.0049
Proposed vs TabPFN	Combined disturbance window	0.011	[0.003, 0.019]	3.11	0.983	0.0125	0.0125
Proposed vs ModernTCN	Combined disturbance window	0.011	[0.004, 0.018]	3.55	1.123	0.0062	0.0124
Proposed vs TabR	Combined disturbance window	0.019	[0.010, 0.028]	4.78	1.512	0.0010	0.0060
Proposed vs TabPFN	Highly sensitive firms	0.021	[0.012, 0.030]	5.28	1.670	0.0005	0.0045
Proposed vs ModernTCN	Highly sensitive firms	0.018	[0.009, 0.027]	4.52	1.429	0.0014	0.0060
Proposed vs TabR	Highly sensitive firms	0.023	[0.013, 0.033]	5.20	1.644	0.0006	0.0048

As shown in Table 7, all Holm-adjusted p-values are below 0.05 across the full out-of-time test, combined-disturbance window, and highly sensitive firm group. For highly sensitive firms, mean AUPRC improvements range from 0.018 to 0.023, consistent with the lower FNR values in Table 6. None of the CIs crosses zero, indicating a consistent improvement direction across ten paired runs. These findings support statistical significance and between-run consistency under the current protocol but do not establish distributional robustness.

Figure 7 presents the mean AUPRC differences and 95% CIs. All estimates lie to the right of zero, and no interval crosses the reference line, consistent with Table 7.

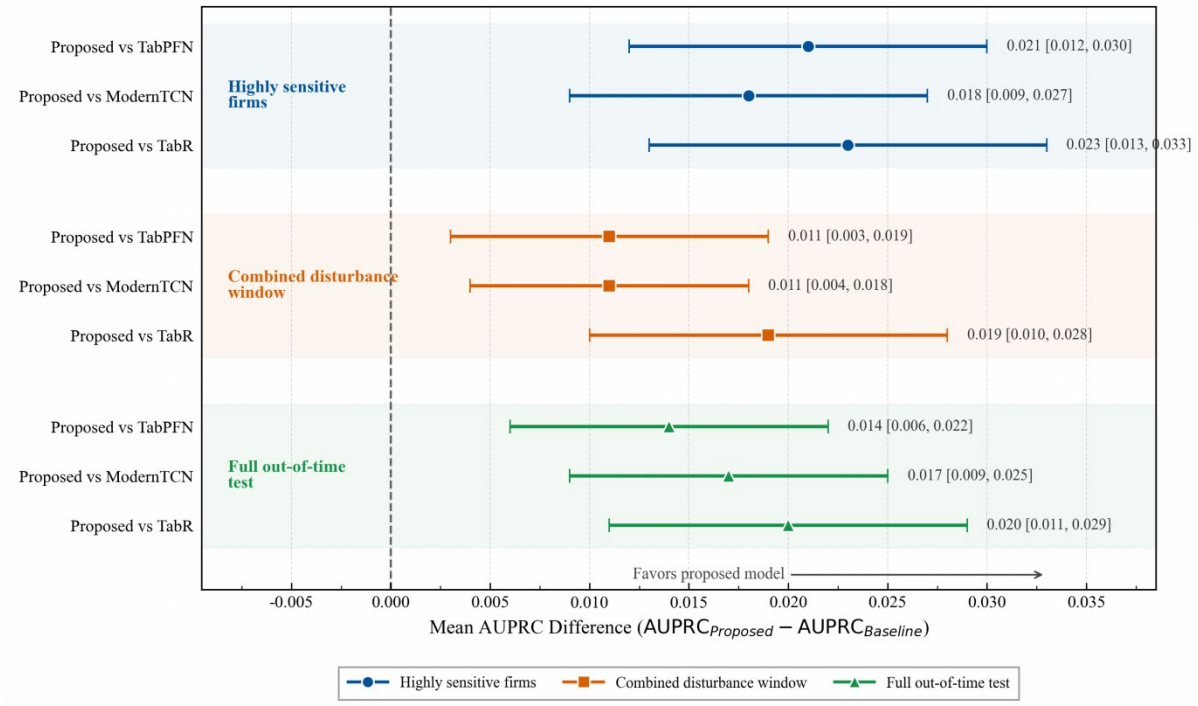

Figure 7. Confidence intervals of AUPRC improvements over strong baselines.

4.4 Representation and Feature Space Analysis

Figure 8 compares raw features, gated recurrent unit (GRU) representations, and the proposed representations using t-distributed stochastic neighbor embedding (t-SNE). All panels use the same stratified out-of-time sample, preprocessing, sample order, perplexity, iterations, and seed. Raw features overlap substantially, GRU representations show partial separation, and the proposed

representations form clearer distressed and highly sensitive clusters. However, t-SNE provides only two-dimensional visualization rather than independent evidence of predictive performance.

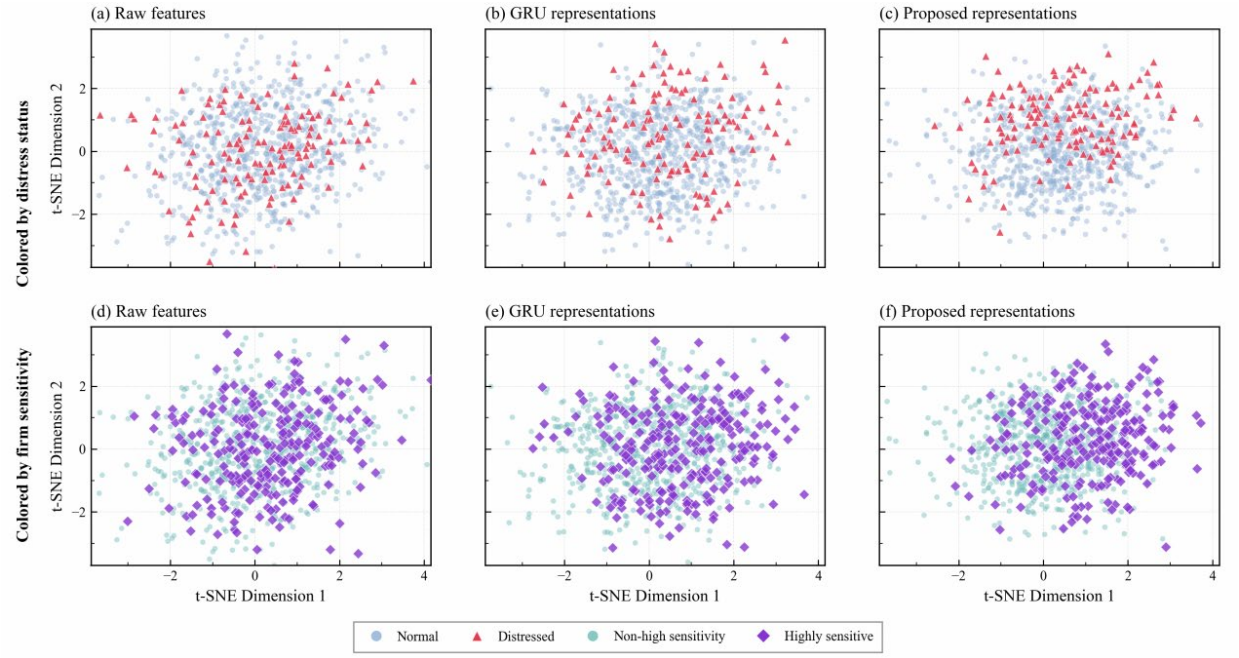

Figure 8. Representation space comparison.
(a–c) Raw features, GRU representations, and proposed representations colored by distress status; (d–f) the same representations colored by firm sensitivity

Figure 9 compares demand, cost, and production sensitivities by distress status and disturbance period. All differences are significant under the Mann–Whitney U test ($p<0.001$). Cliff's delta values are 0.43, 0.31, and 0.44 for distressed versus normal firms and 0.46, 0.17, and 0.41 for disturbance versus non-disturbance periods. Demand and production therefore show clearer shifts than cost sensitivity, although distributional overlap precludes deterministic classification.

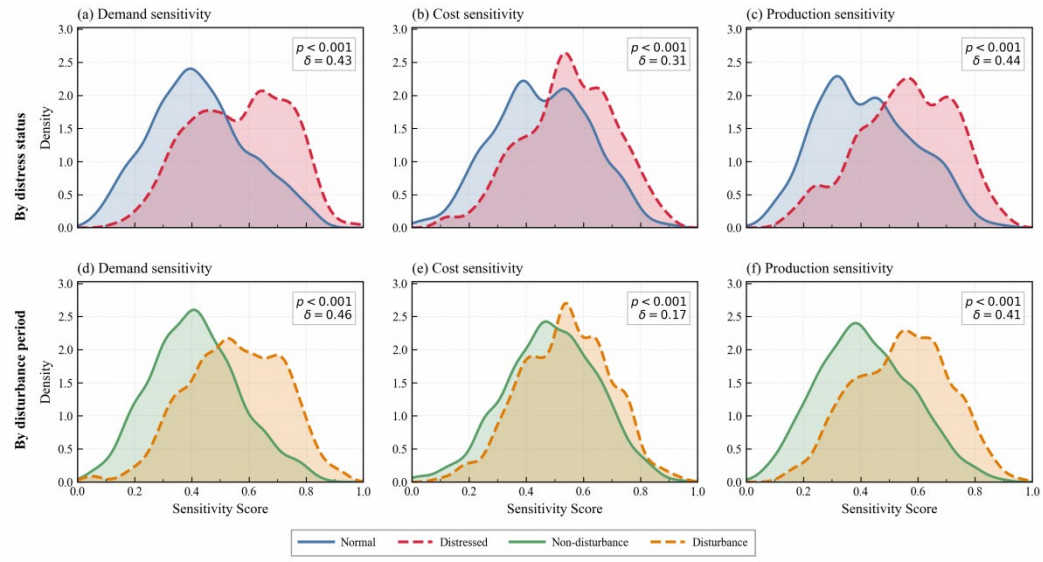

Figure 9. Distribution of disturbance-specific sensitivities.
(a–c) Sensitivity distributions by distress status; (d–f) sensitivity distributions during non-disturbance and disturbance periods

Figure 10 presents nine disturbance – sensitivity matching scores. Same-type scores are 0.626 for demand, 0.579 for cost, and 0.614 for production, each exceeding its cross-type scores. The matching margin is $\Delta s^{k} = s^{k,k} - \frac{1}{2}\sum_{r\neq k} s^{k,r}$. Demand, cost, and production margins are 0.151, 0.075, and 0.139, with bootstrap 95% CIs of [0.140, 0.161], [0.064, 0.090], and [0.125, 0.155], respectively. All intervals remain above zero, indicating consistent same-type score advantages within the test sample.

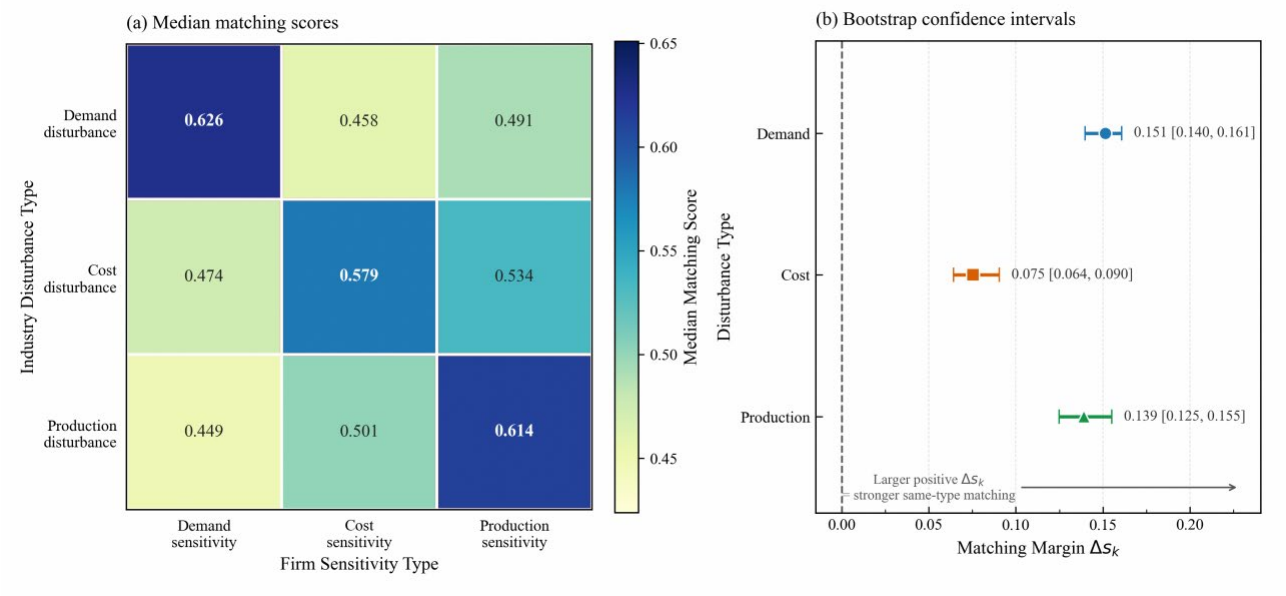

Figure 10. Matching-Score structure. (a) Median matching scores. (b) Bootstrap confidence intervals

Table 8 evaluates the original high-dimensional space. The silhouette score uses distress labels, inter-class distance measures the distance between class centroids, and intra-class distance measures the mean sample-to-centroid distance.

Table 8. Quantitative Representation-Space measures

Model	Silhouette ↑	Inter-class Distance ↑	Intra-class Distance ↓
Raw Features	0.118	1.42	1.08
GRU	0.163	1.58	0.99
GAFormer	0.176	1.66	0.95
Proposed Model	0.214	1.79	0.88

The proposed representation increases the silhouette score from 0.118 to 0.214 and inter-class distance from 1.42 to 1.79, while reducing intra-class distance from 1.08 to 0.88 relative to raw features. Figures 8–10 and Table 8 therefore support clearer sample separation and same-type matching, without establishing causality or cross-dataset effectiveness.

4.5 Interpretability and Exposure-Gate Analysis

4.5.1 Exposure-Gate Interpretation

Figure 11 compares conditional gates and exposure intensities across low-, medium-, and high-sensitivity firms. Groups are defined using the training-period indices in Section 4.1 and fixed during testing. For disturbance k, the mean gate and conditioned exposure are $\bar{g}^{k} = \frac{1}{N}\sum_{i=1}^{N} g_i^{k}$, $E_i^{k} = \| g_i^{k} \odot q_i^{k} \odot v_i^{k} \|_2$. Gate values generally increase with sensitivity, although their magnitudes differ across disturbances. Corresponding groups exhibit higher gates and exposures during demand contraction, producer-price pressure, and production decline. Figure 11 also shows that exposure depends jointly on industry disturbance and firm sensitivity; firms in the same industry-year may therefore receive different risk adjustments. These internal responses should not be interpreted as causal effects or direct measures of economic sensitivity.

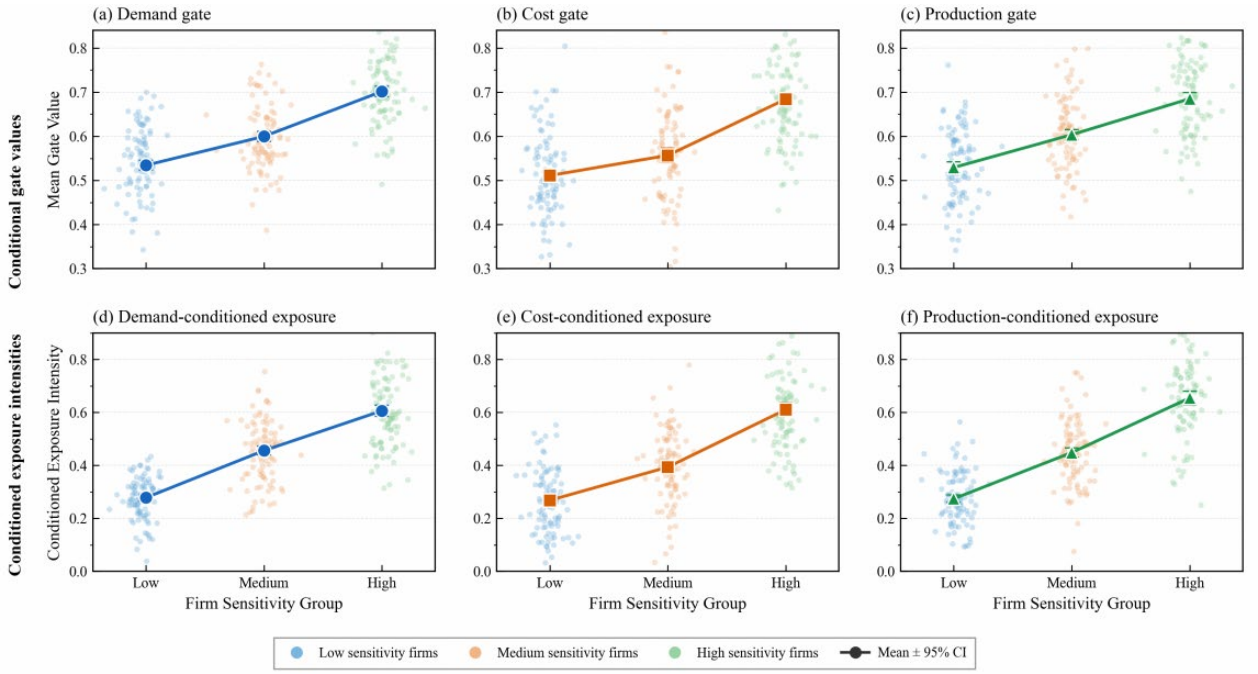

Figure 11. Gate and conditioned-exposure patterns.
(a–c) Demand, cost, and production gates; (d–f) demand-, cost-, and production-conditioned exposures across firms with low, medium, and high sensitivity

4.5.2 Stability across Random Runs

Figure 12 compares gate distributions across ten runs. Demand gates show the least variation, cost gates the greatest, and production gates an intermediate level. Spearman correlations compare rankings of the same out-of-time firms across 45 run pairs.

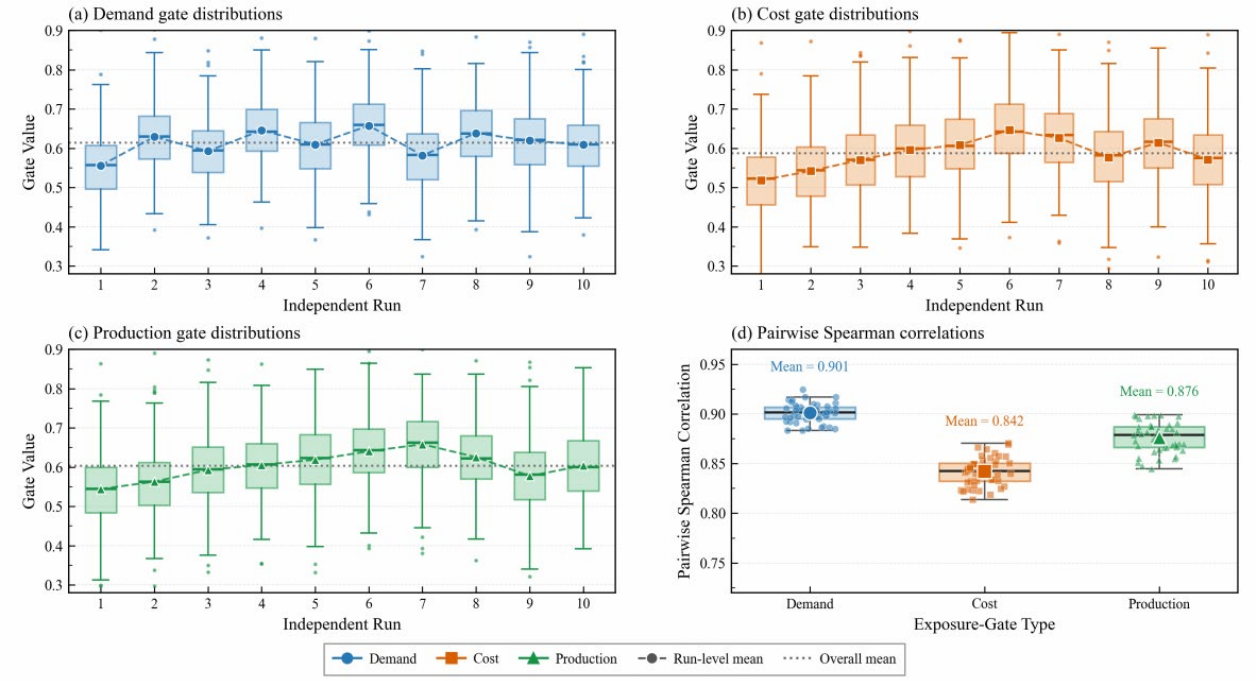

Figure 12. Stability of Exposure-Gate distributions across random runs.
(a–c) Distributions of demand, cost, and production gates; (d) pairwise Spearman correlations across runs

Table 9 confirms that demand gates have the lowest coefficient of variation and cost gates the highest. Mean Spearman correlations exceed 0.84 for all gate types, indicating relatively consistent firm rankings across repeated runs. This evidence concerns only the current split and training configuration and does not imply stability under other settings.

Table 9. Stability of Exposure-Gate interpretation

Exposure Type	Mean Gate	SD	Coefficient of Variation	Mean Spearman Correlation
Demand	0.614	0.031	0.050	0.901
Cost	0.587	0.039	0.066	0.842
Production	0.603	0.035	0.058	0.876

4.6 Error and Failure Case Analysis

This section examines 200 false-negative, false-positive, and high-calibration-error cases from the full out-of-time test. Cases with complete financial, industry, and status records are assigned to mutually exclusive categories based on their primary characteristics. Table 10 provides plausible explanations rather than causal attributions.

As shown in Table 10, delayed disclosure is the most frequent failure type, covering 48 cases (24.0%), followed by rapid post-disturbance recovery with 41 cases (20.5%). Together, they account for 44.5% of errors. Concurrent disturbances represent 35 cases (17.5%), indicating that the current type-specific matching mechanism does not explicitly represent joint exposure when multiple disturbances occur simultaneously. Coarse NACE classification and cross-industry operations account for 29 (14.5%) and 25 cases (12.5%), respectively, while legal or governance events account for the remaining 22 cases (11.0%).

Table 10. Error-Type Statistics

Failure Type	Count	Share	Typical Conditions	Possible Cause
Delayed financial disclosure	48	24.0%	Financial deterioration not yet reflected in annual statements	Lagged information availability
Rapid post-disturbance recovery	41	20.5%	Orders or liquidity recover rapidly after disturbance	Current exposure overstates persistent risk
Concurrent disturbances	35	17.5%	Demand, cost, and production disturbances occur simultaneously	Independent type-specific matching does not represent joint disturbance exposure
Coarse NACE classification	29	14.5%	Substantial heterogeneity within the same two-digit industry	Industry classification is too coarse
Cross-industry operations	25	12.5%	Firm revenue is generated across multiple industries	A single primary-industry label is incomplete
Legal or governance events	22	11.0%	Litigation, fraud, or abrupt governance changes	Financial and industry variables do not capture these events

The current correspondence structure performs type-specific matching for demand, cost, and production disturbances separately. A multi-label extension can retain these single-disturbance exposures while adding pairwise joint-exposure terms. Let E^k denote the exposure associated with disturbance type k. For simultaneously active types k and l, an interaction exposure can be defined as $\phi_{k,l}(D^k, D^l, S^k, S^l)$, where $\phi_{k,l}$ learns the joint effect of two disturbances and their corresponding financial sensitivities. The total exposure can then be extended as $\sum_k \mathbf{E}^k + \sum_{k<l} \mathbf{E}^{k,l}_{\text{int}}$.

The first term preserves the existing single-disturbance correspondences, while the second captures simultaneous shocks such as demand–cost or cost–production disturbances. The correspondence matrix can therefore be extended to a pairwise correspondence structure conditioned on a multi-hot disturbance vector. Higher-order interactions need only be introduced when sufficient concurrent-disturbance observations are available, avoiding unnecessary interaction growth. Figure 13 illustrates representative false-negative, false-positive, and high-calibration-error cases under the fixed calibration-set operating threshold.

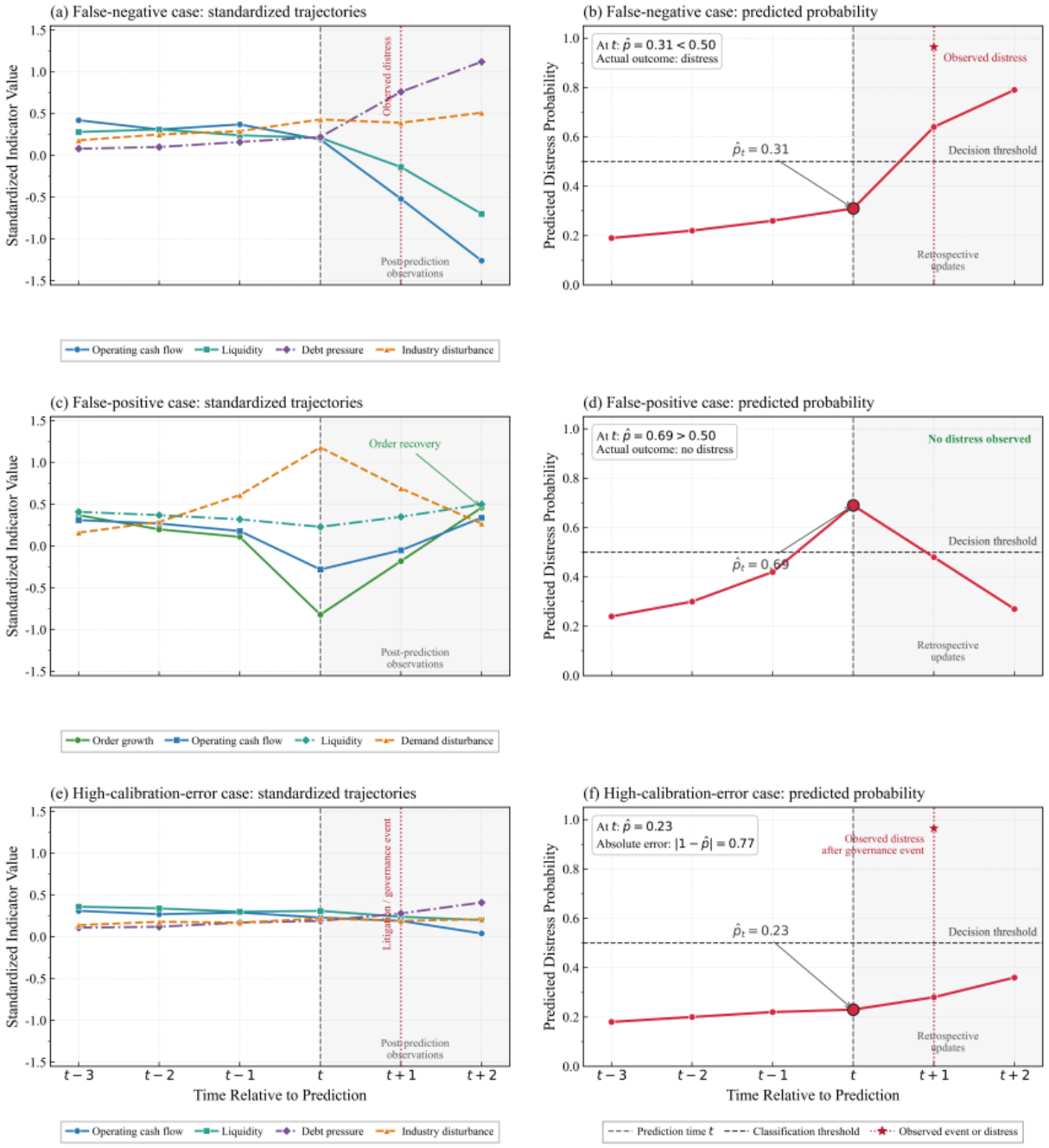

Figure 13. Representative failure cases from the 2022–2023 out-of-time test set. (a–b) False-negative case: financial and industry trajectories and predicted probability; (c–d) false-positive case; (e–f) high-calibration-error case. Predicted-risk classifications use the fixed calibration-set operating threshold

Future work can examine pairwise disturbance interactions together with timelier disclosures, finer industry and business-segment mappings, and legal or governance variables. Dynamic calibration using only historically available data may further address annual probability shifts.

4.7 Ablation Study

Ablations use the full out-of-time test and the combined-disturbance window averaged across demand, cost, and production disturbances. The full model includes three-channel sensitivity encoding, a correspondence matrix, conditional gates, matching-consistency loss, dual-path prediction, and temperature scaling. All variants follow Table 3 and use identical splits, training procedures, seeds, and calibration. Results are mean ± standard deviation over ten runs.

Table 11 shows that the full model achieves AUPRC/FNR values of 0.512/0.276 on the full test and 0.461/0.319 during disturbances. Removing the change channel causes the largest sensitivity-channel decline, reducing AUPRC by 0.014 and 0.018 and increasing FNR by 0.017 and 0.025, respectively. Shared Sensitivity Projection reduces disturbance AUPRC to 0.447. Direct Concatenation lowers full-test and disturbance AUPRC by 0.018 and 0.025, while Shuffled Matching produces the largest disturbance deterioration, decreasing AUPRC by 0.030 and increasing FNR by 0.042. Among single-path variants, Base Path Only performs better on the full test, whereas Exposure Path Only performs better during disturbances. Single Fused Path remains below the full model. Removing accounting divergence lowers AUPRC despite improving the Brier score from 0.075 to 0.074. Removing temperature scaling leaves AUPRC and FNR unchanged but increases the Brier score from 0.075 to 0.081 and ECE from 0.025 to 0.041.

Table 11. Ablation Results

Variant	Full Test AUPRC ↑	Full Test FNR ↓	Disturbance AUPRC ↑	Disturbance FNR ↓	Brier ↓	ECE ↓
Full Model	0.512±0.007	0.276±0.008	0.461±0.009	0.319±0.010	0.075±0.002	0.025±0.002
w/o Level Channel	0.506±0.009	0.285±0.010	0.451±0.011	0.333±0.012	0.076±0.003	0.028±0.002
w/o Change Channel	0.498±0.010	0.293±0.009	0.443±0.012	0.344±0.014	0.080±0.002	0.030±0.004
w/o Accounting Divergence	0.507±0.006	0.282±0.011	0.454±0.009	0.330±0.010	0.074±0.003	0.026±0.003
Shared Sensitivity Projection	0.500±0.009	0.289±0.008	0.447±0.013	0.337±0.012	0.078±0.004	0.027±0.003
w/o Correspondence Matrix	0.496±0.011	0.296±0.012	0.439±0.010	0.349±0.015	0.081±0.003	0.031±0.004
Full Cross-Type Interaction	0.501±0.008	0.291±0.011	0.445±0.009	0.342±0.013	0.078±0.003	0.030±0.002
Direct Concatenation	0.494±0.012	0.300±0.010	0.436±0.014	0.353±0.011	0.082±0.004	0.031±0.003
w/o Conditional Gate	0.499±0.007	0.294±0.013	0.441±0.011	0.346±0.014	0.079±0.002	0.032±0.004
w/o Matching-Consistency Loss	0.503±0.010	0.286±0.009	0.446±0.012	0.339±0.010	0.077±0.003	0.029±0.002
Shuffled Matching	0.492±0.013	0.302±0.012	0.431±0.015	0.361±0.016	0.083±0.004	0.035±0.005
Fixed Matching Score	0.497±0.008	0.297±0.011	0.440±0.013	0.348±0.012	0.079±0.004	0.030±0.003
Base Path Only	0.498±0.009	0.295±0.010	0.437±0.012	0.351±0.014	0.080±0.003	0.031±0.004
Exposure Path Only	0.487±0.014	0.307±0.013	0.443±0.010	0.347±0.011	0.084±0.005	0.036±0.004
Single Fused Path	0.504±0.007	0.287±0.010	0.450±0.011	0.334±0.012	0.077±0.002	0.028±0.004
w/o Temperature Scaling	0.512±0.007	0.276±0.008	0.461±0.009	0.319±0.010	0.081±0.004	0.041±0.006

Figure 14 presents the AUPRC reductions and FNR increases after removing core components, supporting the complementary contributions of sensitivity encoding, constrained matching, and dual-path prediction.

4.7.1 Historical-Window Sensitivity

Figure 15 evaluates the matching-loss weight α, matching margin γ, hidden dimension, and historical window T. For $T \in 2,3,4,5$, the overall validation AUPRC values are 0.493, 0.505, 0.502, and 0.498, respectively, with other settings fixed. Thus, $T = 3$ provides the best overall validation result, while longer histories yield no consistent aggregate gain. The value $T = 3$ is selected on validation data before out-of-time testing.

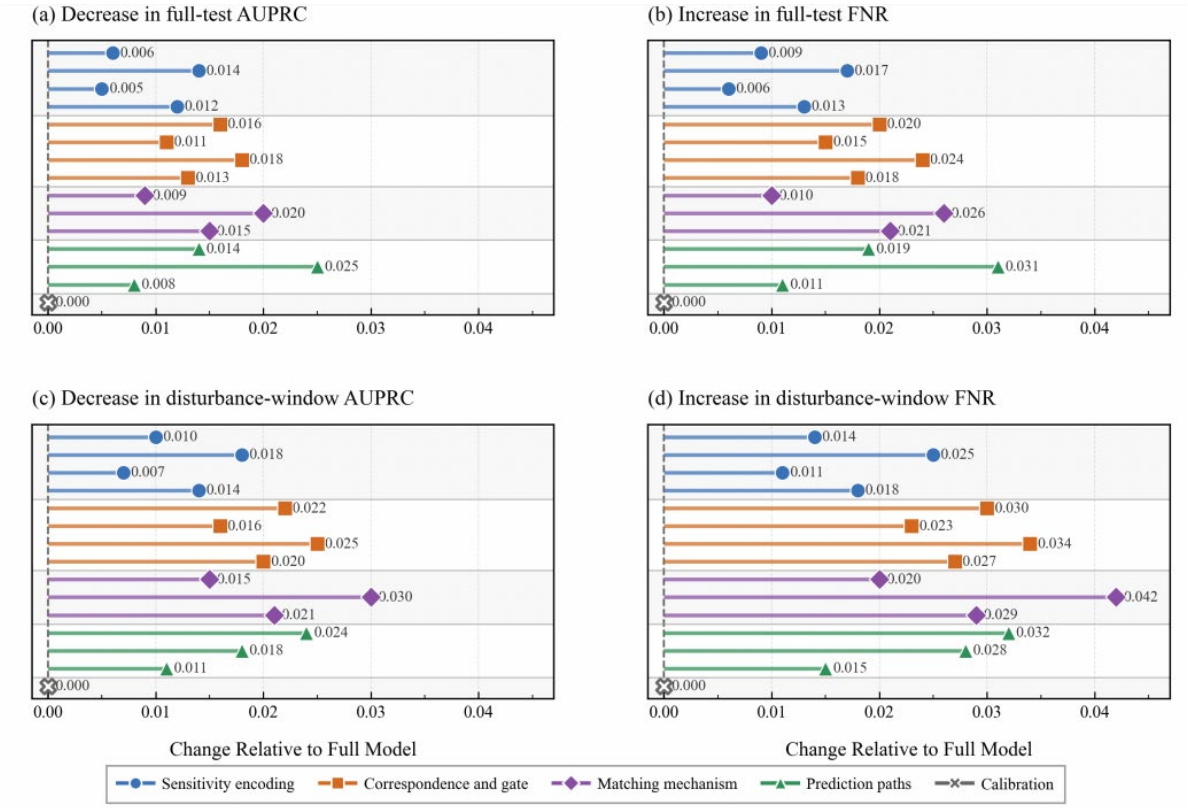

Figure 14. Performance changes after removing core model components.
(a) Decrease in full-test AUPRC; (b) increase in full-test FNR; (c) decrease in disturbance-window AUPRC; (d) increase in disturbance-window FNR

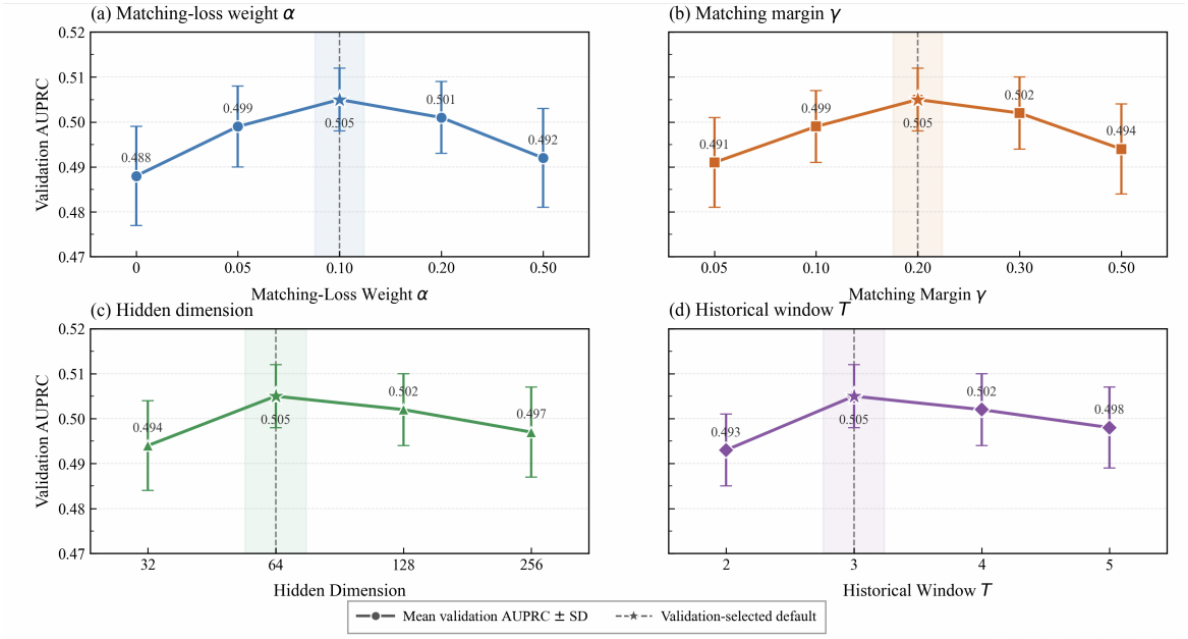

Figure 15. Sensitivity to matching weight, matching margin, hidden dimension, and historical window on the validation set
(a) Matching-loss weight α; (b) matching margin γ; (c) hidden dimension; (d) historical window $T \in 2,3,4,5$.

Table 12 further compares representative subsectors with different manufacturing cycles. Electronics performs best at $T = 3$, whereas machinery and heavy manufacturing obtain slightly higher AUPRC at $T = 4$. The small differences between $T = 3$ and $T = 4$ indicate that three years provide the best overall trade-off rather than a universally optimal window, while longer-cycle subsectors tolerate longer histories better.

Table 12. Historical-Window Sensitivity across Representative Manufacturing Subsectors

Subsector	T = 2	T = 3	T = 4	T = 5
Overall validation set	0.493±0.008	0.505±0.007	0.502±0.008	0.498±0.009
Electronics	0.479±0.011	0.498±0.009	0.493±0.010	0.487±0.011
Machinery	0.472±0.012	0.489±0.010	0.492±0.009	0.488±0.010
Heavy manufacturing	0.461±0.013	0.478±0.011	0.483±0.010	0.482±0.011

Note: Values are validation AUPRC (mean ± standard deviation over ten runs); subsector results assess window sensitivity without retuning the out-of-time test.

4.7.2 Mask Sensitivity Analysis

Mask robustness is evaluated using three perturbations. Remove-One-Variable deletes one retained variable from each type-specific mask; Cross-Group Perturbation moves one boundary variable to another group; and Shared Grouping makes the same financial representation available to all sensitivity heads. Other architecture and training settings remain unchanged.

Table 13 shows limited but consistent degradation under moderate perturbations. Remove-One-Variable reduces AUPRC from 0.512 to 0.506 and increases FNR from 0.276 to 0.284, while Cross-Group Perturbation yields 0.503/0.289. Shared Grouping further reduces AUPRC to 0.500. Thus, performance is not determined by a single exact assignment, but maintaining economically differentiated information groups remains beneficial.

Table 13. Type-Specific Mask Sensitivity on the Full Out-of-Time Test

Mask Setting	AUPRC ↑	FNR ↓
Original type-specific masks	0.512±0.007	0.276±0.008
Remove one variable from each mask	0.506±0.009	0.284±0.010
Cross-group perturbation	0.503±0.010	0.289±0.011
Shared grouping	0.500±0.009	0.289±0.008

5. Discussion

This study proposes an industry-disturbance-conditioned credit-scoring method for manufacturing financial-risk assessment. Unlike direct concatenation or unrestricted financial–industry interactions, it organizes multi-period information into financial levels, changes, and accounting divergences to construct demand, cost, and production sensitivities. A correspondence matrix, conditional gates, and matching consistency generate firm-specific exposure, allowing industry information to enter prediction only through financial sensitivity.

The three financial channels represent current buffers, deterioration, and accounting inconsistencies. Type-specific masks constrain the information available to each sensitivity without predetermining weights or effect directions. Mask-sensitivity results show limited changes under moderate assignment perturbations but clearer deterioration after removing type-specific grouping, indicating that the masks constrain information sources rather than fix learned financial effects.

The correspondence matrix controls which disturbance–sensitivity combinations enter prediction, while matching consistency separates same-type from incorrect combinations. Results for unrestricted interaction, direct concatenation, shuffled matching, and the HGNN baseline support this design. Matching-score and representation analyses also favor same-type relationships. However, these patterns reflect internal model behavior rather than economic transmission or causality.

5.1 Risk-Ranking and Screening Value

The AUPRC of 0.512 should be interpreted against the 4.7% distress-event prevalence and out-of-time setting rather than as a high absolute score. Relative to strong tabular and graph baselines, the gain is modest but accompanied by fewer missed risks. Reducing FNR from 0.291 to 0.276 corresponds to approximately 1.5 fewer missed cases per 100 distressed firms at the fixed threshold, with larger reductions among highly sensitive firms. The practical value therefore lies in improved high-risk ranking and missed-risk control. Because loan exposure, loss-given-default, and realized credit-loss data are unavailable, these results do not quantify monetary loss reduction.

TabPFN obtains slightly lower Brier scores and ECE values in the main evaluations, so the proposed model does not dominate all performance dimensions. Independent calibration improves probability reliability without changing ranking, while its main advantage remains distress identification.

5.2 Concurrent Disturbances and Interaction Effects

The current correspondence structure primarily represents demand, cost, and production through type-specific exposure pathways. Concurrent disturbances account for 17.5% of examined failures, suggesting that simultaneous shocks may produce non-additive effects not captured by independent matching. For example, combined demand contraction and cost increases may affect liquidity and margins differently from either shock alone. A multi-label extension could retain single-type exposures while adding pairwise interaction terms or a joint disturbance encoder and extending the correspondence matrix to higher-order structures. Such extensions should depend on sufficient concurrent-shock observations to limit unnecessary complexity.

5.3 Calibration under Prevalence Shift

The operating threshold and percentile mapping are estimated from the 2021 calibration period and fixed during the 2022–2023 out-of-time test. This prevents test information from entering threshold selection but does not ensure unchanged FNR or precision when future prevalence or macroeconomic conditions shift. Deployment therefore requires threshold monitoring and periodic recalibration or re-estimation using newly available historical observations. The present results establish performance under the specified out-of-time protocol rather than invariance to future prevalence changes.

The evidence supports predictive performance and between-run consistency only within the current firm–industry panel and protocol, not generalization across databases or domains. The scores may support screening and ranking but are neither loan-default probabilities nor financial institutions' internal ratings. Their decision value requires further business validation.

Continuous-record and industry-matching requirements may introduce sample-selection bias. Proxy labels may be affected by delayed disclosure and inconsistent event records, while two-digit and primary-industry classifications incompletely represent cross-industry operations. Predefined masks remain economically motivated design choices despite the sensitivity analysis, and country and industry samples are uneven. Future research should incorporate higher-frequency, business-segment, and governance data; improve joint disturbance modeling; evaluate external databases and subgroup performance; and examine dynamic recalibration under changing economic conditions.

6. Conclusion

This study addresses the difficulty of separating meaningful disturbance–sensitivity relationships from irrelevant cross-type interactions. Its principal originality lies not in adding multi-period financial or industry variables, but in converting external disturbances and firm sensitivities into explicitly matched variables. Unlike direct concatenation or unrestricted interaction, the correspondence matrix restricts candidate matches, matching consistency distinguishes same-type from incorrect cross-type combinations, and dual-path prediction prevents industry information from bypassing firm sensitivity. Financial levels, intertemporal changes, and accounting divergences provide the structured basis for demand, cost, and production sensitivities.

In the full out-of-time test, the method achieves an AUPRC of 0.512, exceeding TabPFN, the strongest AUPRC baseline, by 1.4 percentage points. Its FNR of 0.276 is lower than 0.289 for HGNN and 0.291 for TabPFN. Among highly sensitive firms, the proposed model reaches an AUPRC of 0.489 and an FNR of 0.288, improving on HGNN by 1.5 and 2.3 percentage points, respectively. It also generally achieves higher AUPRC or lower FNR under country-held-out and disturbance evaluations. Ablations support the complementary contributions of sensitivity encoding, constrained matching, and dual-path prediction, while calibration reduces the Brier score and ECE. These results support explicit matching for risk ranking and distress identification, particularly among highly sensitive firms.

The evidence remains limited to the selected databases, disturbance definitions, and experimental settings and does not establish general applicability. Future work should examine higher-frequency disclosures, interacting disturbances, finer business mappings, legal and governance variables, and calibration under changing prevalence. External validation, broader deployment-cost evaluation, and fairness assessment remain necessary.

References

[1] Jiang, Q., Wang, N., Jiang, B., & Bu, F. (2026). Relationship between supply chain disruption orientation and firm's resilience in China: the dynamic capabilities perspective. Asia Pacific Business Review, 32(1), 62-89.

[2] Flannery, M. J., & Öztekin, Ö. (2024). Working capital balances and financial policy. Journal of Corporate Finance, 87, 102618.

[3] Cheng, S. F., Vyas, D., Wittenberg-Moerman, R., & Zhao, W. (2025). Exposure to superstar firms and financial distress. Review of Accounting Studies, 30(2), 1355-1396.

[4] Lokanan, M. E., & Ramzan, S. (2024). Predicting financial distress in TSX-listed firms using machine learning algorithms. Frontiers in Artificial Intelligence, 7, 1466321.

[5] Wang, M., Huang, W., Xie, W., Hou, J., Gao, X., & Fu, X. (2026). TemFRC: Enterprise financial risk prediction with temporal folding and risk contrast. Information Processing & Management, 63(7), 104780.

[6] Beade, Á., Rodríguez, M., & Santos, J. (2024). Business failure prediction models with high and stable predictive power over time using genetic programming. Operational Research, 24(3), 52.

[7] Liu, J., & Jia, M. (2025). Financial distress prediction with annual reports-based deep textual feature extraction: A hybrid approach. Information Sciences, 686, 121318.

[8] Wu, C., Jiang, C., Wang, Z., & Ding, Y. (2024). Predicting financial distress using current reports: A novel deep

learning method based on user-response-guided attention. Decision Support Systems, 179, 114176.
[9] Abdelkader, N. A. M., & Wahba, H. H. (2024). A proposed multidimensional model for predicting financial distress: an empirical study on Egyptian listed firms. Future Business Journal, 10(1), 42.
[10] Che, W., Wang, Z., Jiang, C., & Abedin, M. Z. (2024). Predicting financial distress using multimodal data: An attentive and regularized deep learning method. Information Processing & Management, 61(4), 103703.
[11] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.
[12] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 18(1), 109-131.
[13] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 35(6), 7499–7519.
[14] Zhao, J., Ouenniche, J., & De Smedt, J. (2024). Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction. Machine learning with applications, 15, 100527.
[15] Nguyen, M., Nguyen, B., & Liêu, M. L. (2024). Corporate financial distress prediction in a transition economy. Journal of Forecasting, 43(8), 3128-3160.
[16] du Jardin, P. (2025). A Quantification Approach of Changes in Firms' Financial Situation Using Neural Networks for Predicting Bankruptcy. Journal of Forecasting, 44(2), 781-802.
[17] Thor, M., & Postek, Ł. (2024). Gated recurrent unit network: A promising approach to corporate default prediction. Journal of Forecasting, 43(5), 1131-1152.
[18] Wang, C., Gong, P., Li, J., & Wang, Z. (2025). Corporate financial distress prediction with multiperiod annual report data: A fusion deep neural network model. Plos one, 20(9), e0333064.
[19] Rahmi, A., Lu, C. C., Liang, D., & Fadilah, A. N. (2024). Splitting long-term and short-term financial ratios for improved financial distress prediction: Evidence from Taiwanese public companies. Journal of Forecasting, 43(7), 2886-2903.
[20] El Madou, K., Marso, S., El Kharrim, M., & El Merouani, M. (2024). Evolutions in machine learning technology for financial distress prediction: A comprehensive review and comparative analysis. Expert Systems, 41(2), e13485.
[21] Huang, M. N., & Lee, H. H. (2024). Inter-industry network and credit risk. International Review of Economics & Finance, 92, 598-625.
[22] Huang, Y., Wang, Z., & Jiang, C. (2024). Diagnosis with incomplete multi-view data: A variational deep financial distress prediction method. Technological Forecasting and Social Change, 201, 123269.
[23] Liu, Z., Luo, Y., & Duan, M. (2025). Macroeconomic factors, industrial enterprises, and debt default prediction: Based on the VAR-GRU model. Finance Research Letters, 78, 107122.
[24] Qiu, G., Kuang, D., & Goel, S. (2024, July). Complexity matters: feature learning in the presence of spurious correlations. In Forty-first International Conference on Machine Learning.
[25] Kim, C., Van Der Schaar, M., & Lee, C. (2024, July). Discovering features with synergistic interactions in multiple views. In Forty-first International Conference on Machine Learning.
[26] Wu, S. L., Du, L., Yang, J. Q., Wang, Y. A., Zhan, D. C., Zhao, S., & Sun, Z. X. (2024). RE-SORT: Removing spurious correlation in multilevel interaction for CTR prediction. In Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence (pp. 3816–3828).
[27] Song, Y., Jiang, M., Li, S., & Zhao, S. (2024). Class-imbalanced financial distress prediction with machine learning: Incorporating financial, management, textual, and social responsibility features into index system. Journal of Forecasting, 43(3), 593-614.
[28] Hu, Y. H., Tsai, C. F., & Wang, P. T. (2025). Combining multiple data resampling methods and classifier ensembles for better financial distress prediction: homogeneous and heterogeneous approaches. Annals of Operations Research, 353(2), 793-814.
[29] Guilbert, T., Caelen, O., Chirita, A., & Saerens, M. (2024). Calibration methods in imbalanced binary classification. Annals of Mathematics and Artificial Intelligence, 92(5), 1319-1352.
[30] Nassiri, V., Tekle, F., Tatikola, K., & Geys, H. (2024). Addressing class imbalance in Bayesian classification through posterior probability adjustment. Biometrical Journal, 66(8), e70004.
[31] Yang, M., & Bi, X. (2025). Cost-Aware Calibration of Classifiers. INFORMS Journal on Data Science, 4(2), 101-113.
[32] Gnip, P., Kanasz, R., Zoričak, M., & Drotar, P. (2025). An experimental survey of imbalanced learning algorithms for bankruptcy prediction. Artificial Intelligence Review, 58(4), 104.
[33] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020, April). Heterogeneous graph transformer. In Proceedings of the web conference 2020 (pp. 2704-2710).
[34] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[35] Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., & Chi, E. (2021, April). Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In Proceedings of the web conference 2021 (pp. 1785-1797).
[36] Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., & Babenko, A. (2024). TabR: Tabular deep learning meets nearest neighbors. In International Conference on Learning Representations.
[37] Xiao, J., Liu, R., & Dyer, E. (2024, May). GAFormer: Enhancing timeseries transformers through group-aware embeddings. In International Conference on Learning Representations.
[38] Luo, D., & Wang, X. (2024, May). ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In ICLR.
[39] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). Tabpfn: A transformer that solves small tabular classification problems in a second. In International Conference on Learning Representations.