Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dependence-Aware Forecast Combination under Temporal Distribution Shift: European Insurer Solvency-Ratio Deterioration

Aleksander Wojtowicz
University of Rzeszów
Ewa Rak
erak@ur.edu.pl
University of Rzeszów
Rafał Rak
University of Rzeszów
Anna Denkowska
Krakow University of Economics
Stanisław Wanat
Krakow University of Economics

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11426541/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dependence-Aware Forecast Combination under Temporal Distribution Shift: European Insurer Solvency-Ratio Deterioration

Aleksander Wojtowicz[1], Ewa Rak[2*], Rafał Rak[3], Anna Denkowska[4*], Stanisław Wanat[4]

[1]Department of Artificial Intelligence, Faculty of Exact and Technical Sciences, University of Rzeszów, Pigonia 1, PL 35-310, Rzeszów, Poland.
[2]Institute of Mathematics, Faculty of Exact and Technical Sciences, University of Rzeszów, Pigonia 1, PL 35-310, Rzeszów, Poland.
[3]Institute of Physics, Faculty of Exact and Technical Sciences, University of Rzeszów, Pigonia 1, PL 35-310, Rzeszów, Poland.
[4]Department of Mathematics, Cracow University of Economics, Rakowicka 27, PL 31-510, Kraków, Poland.

*Corresponding author(s). E-mail(s): erak@ur.edu.pl; anna.denkowska@uek.krakow.pl;
Contributing authors: alwojtowicz@ur.edu.pl; rrak@ur.edu.pl; wanats@uek.krakow.pl;

Abstract

We evaluate probability forecast combination for one-year-ahead solvency-ratio deterioration using 9,239 labelled firm-years from European insurers. Model fitting, feature selection and calibration follow calendar order. Using three selected features in the primary evaluation, logistic regression and a linear support-vector machine achieve an out-of-time area under the receiver operating characteristic curve (AUC) of 0.727, versus 0.718 for Dempster fusion and 0.724 for the current solvency ratio alone. With all eight candidate features, Dempster fusion reaches AUC 0.735, exceeding the best single model by 0.008. Exploratory calibrated generalised least-squares pooling improves AUC by 0.01 over the validation-selected constituent; decision-focused pooling reaches AUC 0.733. Pooling results cover two test origins, with firm-clustered uncertainty conditional on those years. A counterexample and synthetic experiments illustrate how repeated evidence can produce overconfident probabilities. Ranking, calibration and operational review value require separate assessment. Independent temporal validation is needed to establish a general advantage for dependence correction.

Keywords: combining forecasts, forecast evaluation, probability forecasting, insurance, risk management, model selection

1 Introduction

Insurance supervision screenings are one-year-ahead ranking problems. Each year a cross-section of undertakings is scored from information known at the reporting date, and a limited inspection budget is allocated to the upper tail. Studies of corporate distress often add a combination layer based on voting, stacking or Dempster–Shafer (DS) evidential fusion, with gains reported against the constituent classifiers [1–5]. Whether those gains persist under a calendar-time forecast design remains an empirical question.

Random k-fold on pooled firm-years can place the same macroeconomic period, and the same firm in adjacent years, on both sides of a split [6, 7]. This design answers a cross-sectional interpolation question. It does not reproduce the information set available when the next reporting year arrives, especially after a rate, inflation, or pandemic shock [8, 9].

The study compares arithmetic and reliability-weighted voting, Dempster's rule [10, 11], Yager's conflict-to-ignorance rule [12], Murphy's dependence correction [13], an instance-adaptive discount, and non-negative linear pools. The pool analysis includes equal, inverse-Brier, shrinkage, dependence-aware and fixed-capacity decision-focused weights. This reflects the distinction between forecast-score and downstream-decision optimisation [14–18]. The benchmarks include each constituent classifier and the current solvency ratio (SR) alone. Evaluation is repeated under random firm-year folds, firm-grouped folds, two out-of-time windows and an expanding rolling origin. Feature selection is repeated within the available history for each temporal origin.

The contribution is an empirical audit of forecast combination under an explicit calendar protocol. Validation-error weighting is compared with established pooling methods, while a counterexample and synthetic simulations examine the overconfidence of repeated evidence. The pooling analysis is exploratory: it was developed after the primary test results were examined and does not provide independent confirmation. Additional simulations and robustness checks are reported in the supplementary material.

This question is related to the forecast combination puzzle: simple averages and univariate benchmarks can match or beat elaborate combination schemes [14, 16, 17, 19, 20]. Classifier fusion combines probability forecasts, so the benchmark set includes logistic regression, a rank based on the current solvency ratio, a nested ratio-only probability, lagged-event persistence, and training prevalence. Rank quality and probability quality are reported separately [21, 22].

The insurance design states the forecast clock explicitly: year t is the origin, the outcome occurs in $t+1$, and a lagged change is computed only from $t-1$ to t. Consecutive calendar years are required for both lag and lead. The event is called material solvency-ratio deterioration rather than failure, insolvency or distress. Country-year

changes in interest rates, inflation and GDP growth enter the candidate pool because duration and rate mismatch affect European solvency ratios [8, 23, 24]. A market-wide Herfindahl index is used only as a diagnostic because it is constant between firms within a year.

2 Related work

Insurer distress and failure prediction with accounting ratios, neural nets, and tree ensembles is a mature application, from the 1990s property-liability models [25, 26] through yield-curve and macro predictors [24, 27, 28]. Recent evidence indicates that machine-learning architectures can improve out-of-sample insurer-failure prediction relative to logistic regression and standard machine-learning benchmarks [29, 30]. Our study addresses a different question: whether evidential fusion improves transparent baselines under a strictly temporal Solvency II evaluation. Combining classifiers with Dempster's rule, voting or stacking is also established in listed-firm distress, and these papers typically report a gain over every base learner [1–5]. Random k-fold is not a forecast protocol for dependent panels [6, 7]. Dempster's rule assumes evidential independence that classifier scores trained on the same features do not satisfy [11, 31–33]. Probability quality is different from rank quality [22, 34]. Decision curve analysis converts scores into net benefit at an explicit action threshold [35].

The open question is whether a reported fusion gain remains when the validation design follows calendar time and when combination is compared with a strong univariate forecast. Table 1 summarises representative evidential and voting studies, their validation protocols and their reported comparison with the best single classifier. Several studies in Table 1 report gains under random partitioning or validation without an explicit temporal separation. In the broader forecast-combination literature, simple averages are often difficult to improve [16, 17].

Recent studies extend financial-risk prediction with multimodal attention, modified Dempster–Shafer fusion and soft-probability random forests [36–38]. These studies address corporate financial distress using different data and event definitions. V4FinBench compares standard tabular models, tabular foundation models and a large language model for corporate bankruptcy prediction [39]. Recent work also modifies Dempster–Shafer fusion to account for conflict between evidence sources [40]. Related forecast-combination studies optimise weights for a downstream decision loss or regularise covariance estimates under structural breaks [18, 41]. Other recent work addresses serial dependence in combined-forecast errors and adaptive ensemble weighting [42, 43]. Serial dependence concerns errors across forecast origins; the covariance used here describes validation errors across classifiers. Our analysis applies related weighting ideas across classifiers under a calendar-time insurance design.

The Solvency II Standard Formula is not a simple linear aggregation. Diversification uses prescribed correlation matrices and square-root formulae [44, 45]. We use that fact only as an institutional background. The forecasting object here is the one-year solvency-ratio change of solo undertakings, not a claim that a statistical screen replaces the Standard Formula.

Table 1 Validation protocol in representative evidential or voting fusion papers for credit, insurance or financial risk. Time-respecting means training years strictly precede test years. The final column records a qualitative reported claim, not a pooled effect estimate, because the datasets, targets, metrics and operating points differ.

Study	Domain	Fusion	Time-respecting	Reported gain?
Xiao et al. [1]	listed-firm distress	rough set + DS	no (random)	claimed
Zhu et al. [5]	firm financial risk	DS-RF	not time-blocked	claimed
Wang and Chi [2]	listed-firm distress	cost-sensitive stack	mixed	claimed
Liu and Li [3]	financial-risk EWS	DS ensemble	not time-blocked DS	claimed
Zhang et al. [4]	listed manufacturers	TD-DS-RF	80:20; split years not specified	claimed
Sun et al. [36]	listed-firm distress	multimodal attention	pooled market data	claimed
Song et al. [37]	SME financial crisis	improved DS + SHAP	not calendar walk-forward	claimed
Qiu et al. [38]	listed-firm distress	soft-probability RF	pooled market data	claimed
This paper	Solvency II SR drop	vote / DS / Yager / Murphy	yes; random control	tested, not assumed

3 Data, clock and labels

3.1 Panel

Firm-years come from commercial Solvency II solo Quantitative Reporting Templates covering composite, general (non-life), health, and life undertakings. Public SFCR/QRT templates are the regulatory source; the panel itself is a vendor extract from Insurance Risk Data. We keep single reporters with positive assets, positive basic SCR, and solvency ratios in $(10\%, 2000\%)$. We do not require complete SCR submodule reporting, which would favour large composites. After requiring a lag and a lead for each firm in consecutive calendar years, the labelled panel has 9,239 rows on 1,979 undertakings, with forecast origins 2017- 2023 (Table 2). Table 3 records the exclusions. The commercial files contain 16,809 solo firm-years from 2016 to 2024. 14,717 survive the asset, basic-SCR and solvency-ratio bounds. Further rows 5,478 lack the required consecutive lag, lead, or country-year macro change. The reporting year 2024 has no observable $t+1$ outcome, while 2016 cannot provide a lag for a firm's first observation.

Internal-model adoption is rare (mean 0.0019; 18 of 9,239 rows). The dummy is withheld because it is nearly constant. A year-level Herfindahl asset index is computed for diagnostics and omitted from the primary candidate set. Supplementary Table S1 summarises the eight candidates and diagnostic series. Large changes in assets and ratios can reflect mergers, portfolio transfers, or recapitalisations. In every experiment, scaling and the 1st and 99th percentile winsorisation check are fitted on model-fit observations only, before the validation year.

Table 2 Labelled firm-years by forecast origin t. 2016 drops out (no consecutive $t-1$); 2024 drops out (no consecutive $t+1$). Outcomes are realised in $t+1$.

Forecast origin t	2017	2018	2019	2020	2021	2022	2023
Outcome year $t+1$	2018	2019	2020	2021	2022	2023	2024
n	991	1,215	1,329	1,424	1,454	1,471	1,355

Table 3 Sample selection. The commercial files are restricted to solo reporters from 2016 to 2024.

Step	n
Solo firm-years in the vendor extract, 2016-2024	16,809
Positive assets, positive basic SCR, solvency ratio in $(10\%, 2000\%)$	14,717
Consecutive $t-1$ lag, consecutive $t+1$ lead, and observed macro changes	9,239
of which unique undertakings	1,979

Table 4 Technical chronology. Feature selection and preprocessing use model-fit years. The validation year supplies reliability, thresholds, calibration and combination weights. The last column reports observed events in the test period.

Experiment	Model fit	Validation	Test	n_{fit}	n_{val}	n_{test}	$n_{\text{event,test}}$
Out-of-time W1	2017-2020	2021	2022-2023	4,959	1,454	2,826	660
Out-of-time W2	2017-2018	2019	2020-2021	2,206	1,329	2,878	740
Rolling origin	through $t-2$	$t-1$	year t	varies	varies	varies	varies

3.2 Clock

Every modelling row is indexed by forecast origin t. For a firm-year t:

$$\text{feature change:} \quad \Delta R_{i,t} = R_{i,t} - R_{i,t-1}, \qquad Y_{i,t} = \mathbf{1}\{R_{i,t+1} - R_{i,t} \leq -20\}, \tag{1}$$

with both differences defined only when the adjacent calendar years exist. Window 1 uses the 2017-2020 origins for model fitting and feature selection, 2021 for reliability, thresholds, calibration, and stack weights, and 2022-2023 for testing. The complete development sample contains 6,413 rows: 4,959 model-fit and 1,454 validation observations. Window 2 uses 2017-2018 for model fitting and feature selection, 2019 for validation, and 2020-2021 for testing. No test label is used for preprocessing, selection, tuning, or calibration, and no firm-year appears on both sides of an out-of-time split. Table 4 summarises the fitting, validation and test periods.

3.3 Labels

A 20 percentage-point drop in the solvency ratio is not insolvency, an SCR breach, a supervisory intervention, or even a low ratio. An undertaking at 400% can record such

a drop and remain comfortably capitalised. We therefore call $Y_{i,t}$ a material solvency-ratio deterioration event. The prevalence is 25.4% overall and 23.4% in Window 1. The test prevalence is not, by itself, evidence of greater macroeconomic stress than the training window. Sensitivity uses absolute drops of 10, 20 and 30 pp, relative drops of 10% and 20%, and 20 pp drops restricted to cases ending below 200% or 150%.

4 Methods

4.1 Feature selection

Three selectors are applied only to model-fit years: L1 penalised logistic regression [46], recursive feature elimination (RFE) with random forests [47, 48], and mutual information [49]. RFE and mutual information retain k features. The number of retained features is evaluated over $k \in \{3, 4, 5, 6, 8\}$ by expanding-year inner folds. The scaling is refitted inside each fold. The selected value is the smallest k whose mean inner AUC is within one standard error of the maximum. The resulting Window 1 set is solvency ratio, solvency-ratio change ($t-1$ to t) and Δ inflation. It is fitted on 2017-2020 and remains fixed for validation in 2021 and testing in 2022-2023. Window 2 and each rolling origin repeat selection using only earlier model-fit years. Random and grouped cross-validation repeat selection within each training fold. Selector inclusion counts, stability and the all-feature check are reported in Supplementary Section S2. Supplementary Figures S1 and S2 show the selector inclusion counts and bootstrap selection stability, respectively.

4.2 Classifiers

Logistic regression, random forest [48], XGBoost [50] and a linear SVM with Platt scaling [51, 52] are trained on the consensus features. Balanced class weights are used for logit, forest and SVM; XGBoost receives a positive-class weight. These choices target class imbalance ranking but distort raw probabilities, so the paper separates AUC from Brier and log scores and includes temporal recalibration. Full configurations and validation checks are given in Supplementary Table S2. Forest depth, boost depth, and logistic C are checked on the validation year without touching the test labels. The RBF SVM is a disclosed diagnostic. Scalers and winsorisation bounds use model-fit observations only. The validation year supplies source reliability and calibration. Forest and XGBoost are repeated in five seeds in Window 1. The main seed is 42.

Three naive forecasts sit outside the classifier stack. The *current-ratio* rank uses $R_{i,t}$ itself, with the high-ratio orientation implied by an unrestricted 20 pp drop. A univariate logistic regression on $R_{i,t}$ alone provides the nested probability forecast used in the Clark-West test [53]. The opposite, traditional distress orientation (low ratio, high score) is reported as a check. The *lagged-event* forecast uses $Y_{i,t-1}$ when it exists and training prevalence otherwise. The *climate* forecast is the constant rate of training events. An unweighted logistic regression on the consensus features is stored only for probability evaluation; the primary logit keeps balanced class weights because the screen is a ranking problem.

4.3 Combination rules

Let $p_i(x)$ be classifier i's predicted event probability and let $r_i \in [0.5, 0.99]$ be its AUC on the held-out training year. Soft voting is $\bar{p} = \frac{1}{I}\sum_i p_i$. Weighted voting uses $w_i = \max(r_i - 0.5, \varepsilon)$ and $\sum_i w_i p_i / \sum_i w_i$. For evidential rules the frame is $\Theta = \{N, A\}$ and the mass assignment is

$$m_i(\{A\}) = p_i r_i, \qquad m_i(\{N\}) = (1 - p_i) r_i, \qquad m_i(\Theta) = 1 - r_i. \tag{2}$$

The Dempster combination of two masses is

$$m_{1\oplus 2}(C) = (1 - K)^{-1} \sum_{X \cap Y = C} m_1(X) m_2(Y), \quad C \neq \emptyset, \qquad K = \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y), \tag{3}$$

with $m_{1\oplus 2}(\emptyset) = 0$ and $1 - K$ bounded below by 10^{-12} [11]. Yager's rule assigns K to $m(\Theta)$ instead of renormalising [12]. Murphy's rule averages the I mass functions and then applies the Dempster combination $I - 1$ times [13]. The adaptive variant multiplies r_i by a novelty factor $d(x) = (1 + 0.5\,\bar{\delta}_{15}(x))^{-1}$, where $\bar{\delta}_{15}$ is the mean Euclidean distance to the 15 nearest training points. The pignistic probability of A is $m(\{A\}) + \frac{1}{2} m(\Theta)$ [54]. Combination order is sequential in the fixed sequence logit, forest, boosting. Primary fusion excludes SVM. An ablation repeats soft voting and Dempster combination with SVM included. Tree-only fusion repeats soft voting and Dempster combination in forest and boosting scores alone.

Dependence boundary.

Let every source be a deterministic monotone transform of one latent score, $p_j(x) = g_j\{s(x)\}$, and suppose that the resulting fusion score can be written as $F(x) = h\{s(x)\}$ for a strictly increasing h. Then F and s induce the same ordering (apart from ties), and consequently have identical receiver operating characteristic (ROC) curves and AUC. Thus repeated monotone views of the same information cannot create a rank gain merely by being combined. Proper-score performance can be worse even without drift. For example, two identical calibrated Bayesian masses with $r = 1$ and event probability p yield, under Dempster's rule, $q = p^2/[p^2 + (1 - p)^2]$. For $p \notin \{0, 1/2, 1\}$, $q \neq p$; strict propriety of the Brier and logarithmic scores then implies a larger expected loss than the original calibrated probability. This counterexample does not assert that every dependent fusion must fail. It identifies why evidential independence and incremental information must be demonstrated rather than inferred from the number of fitted classifiers.

4.4 Inference and utility

Out-of-time AUC intervals and pairwise differences are based on 1,000 paired bootstrap resamples of undertakings. Each draw retains all test observations for a sampled undertaking, and all models are evaluated on the same resample. This is the main uncertainty summary. Giacomini–White, Clark–West and model-confidence-set results are reported in Supplementary Section S4. Non-negative Bates-Granger weights that

sum up to one are estimated in the validation year and then applied to the test years [14]. Rolling-origin evaluation uses years through $t-2$ for model fitting, $t-1$ for validation, and year t for testing, from 2019 to 2023. Five annual observations are too few for a useful two-sided Wilcoxon decision, so annual means are descriptive. Companion metrics include the area under the precision–recall curve (PR-AUC) [34], log score, calibration error, precision in fixed review shares, alert burden, and net benefit from decision-curve [35]. Isotonic maps are fitted on the validation year [21]. The Dempster conflict K is used only to compare low- and high-conflict subsets.

The exploratory combination analysis calibrates each source and estimates pool weights on the 2021 origin. It compares equal and inverse-Brier pooling, a constrained simplex, shrinkage toward equal weights, logistic stacking, fixed-capacity decision-focused pooling, and generalised least-squares weights from validation-error covariance. The GLS weights are clipped at zero and renormalised to sum to one. This procedure does not solve the constrained quadratic optimisation problem. The GLS contrast uses a constituent comparator selected by validation Brier rather than test loss. The pooling intervals use 2,000 paired bootstrap draws that resample entire firms and retain every observed year. The delayed-feedback pool repeats feature selection using only the model-fit history at each origin.

The main test contains two forecast origins, and the rolling evaluation contains five. Firm-clustered intervals describe uncertainty across undertakings conditional on the observed years and fitted models; they do not quantify variation across future calendar regimes. The Giacomini–White, Clark–West and model-confidence-set results are therefore treated as conditional diagnostics. The exploratory pools were developed after inspection of the primary 2022–2023 test results, and their intervals do not adjust for method selection or multiple comparisons.

5 Results

5.1 Selection and SVM

The one-standard-error rule selects $k = 3$ features: solvency ratio, solvency-ratio change ($t-1$ to t) and Δ inflation. With all eight candidates, Dempster AUC rises to 0.735 [0.715, 0.757], with an increment of 0.008 [0.001, 0.014] over the best single model. This result comes from the all-feature sensitivity analysis; the primary specification uses the three features selected by the one-standard-error rule. The linear SVM and logistic regression both achieve AUC 0.727 to three decimal places, compared with 0.622 for the RBF SVM. Full selection, SVM, HHI and tuning checks are reported in Supplementary Sections S2–S3.

5.2 Naive forecasts and proper scores

Ranking firms by current solvency ratio, with the high-ratio orientation implied by an unrestricted 20 pp drop, yields Window 1 AUC 0.724. Logistic regression with balanced weights reaches 0.727. The clustered logit-minus-ratio difference is 0.002 [-0.002, 0.006]. The current solvency ratio alone provides a strong ranking benchmark; the paired confidence interval does not establish an AUC improvement from logistic

Table 5 Window 1 rank and probability scores. Isotonic maps are fitted on the last training year. Current-ratio AUC uses the high-ratio orientation. Climatology is the training event rate.

Method	AUC	Raw Brier	Raw log loss	Isotonic Brier
Current solvency ratio	0.724	n/a	n/a	n/a
Lagged event	0.513	n/a	n/a	n/a
Climatology	0.500	0.180	0.546	n/a
Logistic regression	0.727	0.218	0.643	0.163
Random forest	0.710	0.218	0.621	0.166
SVM (linear, Platt)	0.727	0.164	0.506	0.163
Soft voting	0.717	0.213	0.614	0.164
Dempster rule	0.718	0.215	0.618	0.164

regression. Balanced logistic regression has the largest AUC in 63.4% of firm-clustered bootstrap samples, while Dempster fusion leads in 2.9%. The full resampling comparison is shown in Supplementary Figure S3. The opposite ratio orientation has an AUC 0.276, showing that the label is not a proxy for the proximity to the regulatory threshold. The lagged-event forecast has AUC 0.513.

Ranking performance and probability accuracy are evaluated separately. Table 5 reports the rank and probability scores for Window 1. An unweighted logit on the selected features has Window 1 AUC 0.723 and Brier score 0.164, compared with 0.165 for the ratio-only probability model. Balanced class weights raise rank performance to AUC 0.727 but produce raw Brier 0.218, compared to 0.180 for climatology. After isotonic calibration on 2021, logistic and Dempster Brier scores are 0.163 and 0.164. Loss tests and the model confidence set are reported in Supplementary Section S4.

5.3 Optimal combination and the model confidence set

Non-negative weights on logistic regression, forest and boosting assign 1.00 to logit and 0.00 and 0.00 to trees, for Window 1 AUC 0.727. When the ratio-only probability model is added, it receives weight 1.00 and the fitted classifiers receive zero, giving AUC 0.724. Equal weighting of the probability of the balanced and ratio-only logit gives AUC 0.727. In these two specifications, the fitted linear pool assigns all weight to a single source [14, 16].

The model confidence set retains 3 models: logistic regression, the current-ratio probability model and the linear SVM. It does not identify a unique best model. Details are reported in Supplementary Section S4.

5.4 Dependence-aware and decision-focused pools

The GLS pool assigns weights of 0.74, 0.23 and 0.03 to logistic regression, random forest and boosting, respectively. It achieves AUC 0.731, PR-AUC 0.432, Brier score 0.162 and log loss 0.490. Relative to the constituent selected by validation Brier, its paired AUC difference is 0.010 [0.0005, 0.0203], its Brier reduction is 0.001 [0.0001, 0.0020], and its log-loss reduction is 0.008 [0.0019, 0.0189]. The paired intervals compare GLS only with the validation-selected constituent. Table 6 gives descriptive comparisons

Table 6 Exploratory modern pooling audit on the 2022–2023 origins. Every source calibration and pool weight is fitted on 2021.

Method	AUC	PR-AUC	Brier	Log loss
Equal pool	0.722	0.423	0.163	0.493
Constrained simplex	0.731	0.433	0.162	0.490
Shrinkage simplex	0.723	0.424	0.163	0.493
Dependence-aware GLS	0.731	0.432	0.162	0.490
Decision-focused pool	0.733	0.437	0.162	0.491
Logistic stacking	0.728	0.430	0.163	0.498
Dempster rule	0.718	0.432	0.215	0.618

among the pools. Decision-focused pooling has the highest reported AUC (0.733) and PR-AUC (0.437) among the pools in Table 6. Comparing the calibrated pools with Dempster scores changes both source calibration and the combination rule, so it does not isolate the effect of accounting for source dependence.

For the primary 20 pp event, current SR alone reaches AUC 0.724, the selected specification with SR 0.727, and the same selected specification without SR only 0.607. The current solvency ratio is observed at forecast issuance and is a key predictor of the defined deterioration event. The delayed-feedback pool achieves AUC 0.697 when predictions for 2019–2023 are evaluated together.

In the synthetic experiment, higher correlation between source errors reduces the benefit of equal pooling and increases Dempster's Brier loss relative to the best single source. The design and results are reported in Supplementary Section S5 and Supplementary Table S5. These results illustrate a mechanism and do not estimate its effect in the insurer panel.

For next-year solvency-ratio change, the univariate ridge model based on the current solvency ratio has the lowest RMSE among the evaluated models: 84.1, compared with 85.2 for the full ridge model. Full results are reported in Supplementary Section S6. This supports the use of the current ratio as the main simple benchmark.

5.5 The protocol contrast

The pipeline is evaluated under random firm-year five-fold, firm-grouped five-fold and out-of-time Window 1. Figure 1 compares AUC across the three validation protocols for selected classifiers and combination rules. Table 7 gives the AUC estimates and out-of-time confidence intervals. The reported increment is the fusion AUC minus the best single-model AUC. Random forest is the strongest single model under random folds (0.729) and grouped folds (0.730), while logistic regression and the linear SVM both reach AUC 0.727 out of time to three decimal places. In the primary feature specification, Dempster's AUC increment over the best single model is 0.000 under random folds, -0.001 under firm-grouped folds and -0.009 out of time. The firm-clustered interval for the out-of-time increment is $[-0.019, 0.001]$. The random and grouped protocols use the pooled eligible panel, whereas Window 1 evaluates the

Table 7 AUC under three validation protocols. Feature selection and preprocessing are training-fold specific. Out-of-time Window 1 fits on 2017-2020, validates on 2021 and tests on 2022-2023. Δ is fusion minus the best single model. Random and grouped folds use the pooled eligible panel, so their differences from Window 1 do not isolate the effect of the split.

Method	Random k-fold	Grouped-by-firm	Out-of-time W1 [95% CI]
Logistic regression	0.708	0.714	0.727 [0.705, 0.748]
Random forest	0.729	0.730	0.710 [0.688, 0.732]
Gradient boosting	0.722	0.723	0.692 [0.668, 0.715]
SVM (linear, Platt)	0.706	0.712	0.727 [0.707, 0.747]
Soft voting	0.729	0.730	0.717 [0.695, 0.739]
Dempster rule	0.729	0.730	0.718 [0.695, 0.740]
Yager rule	0.729	0.729	0.715 [0.692, 0.737]
Murphy DS	0.729	0.730	0.718 [0.696, 0.740]
Adaptive DS	0.729	0.729	0.717 [0.695, 0.741]
Dempster (trees only)	0.728	0.728	0.706 [0.683, 0.729]
Dempster Δ vs best single	0.000	-0.001	-0.009

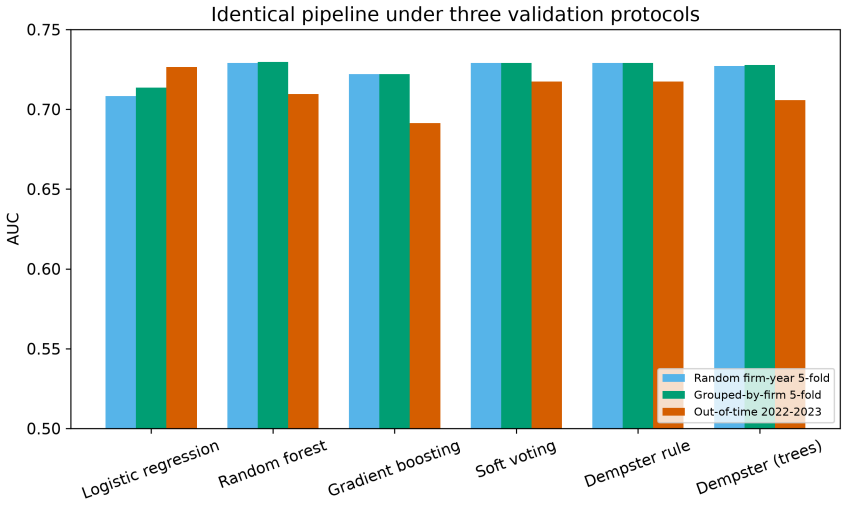


Fig. 1 Protocol contrast for the main classifiers and combination rules. The values summarise three evaluation designs; the out-of-time bar also uses a later evaluation period.

2022–2023 origins. The observed differences therefore reflect both the split design and the evaluation period.

Window 2 (COVID origins 2020-2021, validated on 2019) is a second non-overlapping test: forest 0.692, Dempster 0.694, logistic 0.697. Reliability weights estimated on the Window 1 validation year are 0.662 (logit), 0.655 (forest), 0.642 (boosting) and 0.662 (SVM). The five-seed Window 1 forest AUC has a mean 0.710 (sd 0.000).

Fixed-configuration CatBoost, LightGBM, Extra Trees and histogram gradient boosting provide an additional comparator check. With the selected three features,

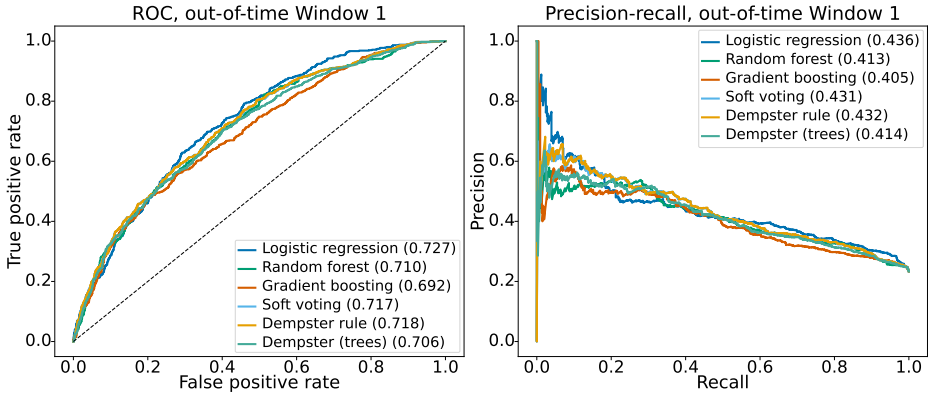


Fig. 2 Out-of-time Window 1 ROC and precision-recall curves for the main scores.

these additional classifiers have lower AUC than logistic regression in both temporal windows. Full results for both temporal windows and both feature sets are reported in Supplementary Tables S3–S4.

5.6 Source dependence and conflict

On Window 1, the Spearman correlation is 0.91 between random forest and boosting scores and 0.84 between logistic regression and random forest scores. The high correlation is not a formal proof that Dempster independence is violated, but it shows that the sources share information: they are trained on the same observations and the same features [31, 32]. The Yager and Murphy rules give AUC values close to that of the random forest. We split Window 1 at the median Dempster conflict K. The logistic AUC is 0.793 in the low-conflict half and 0.622 in the high-conflict half; Dempster tracks the same split (0.783 versus 0.593). At a 10% review budget, restricting selection to low-conflict observations gives precision 0.539, compared with 0.544 using the full Dempster ranking. The current solvency ratio has a population stability index (PSI) of 0.014, while policy-rate and inflation levels have PSI values of 5.02 and 5.69, respectively. Figure 2 shows the ROC and precision-recall curves for the main scores. PR-AUC is 0.413 for random forest and 0.436 for logistic regression at a test prevalence of 23.4%. The calibration plot is shown in Supplementary Figure S6.

Rolling-origin, alternative-label and winsorisation checks are reported in Supplementary Sections S6–S7. Supplementary Table S6 reports annual AUC values, and Supplementary Figures S4 and S5 show the classifier comparison without and with the current-ratio benchmark, respectively. Alternative-label results are summarised in Supplementary Table S7. Across 2019–2023, the current-ratio rank has mean AUC 0.718, compared with 0.717 for Dempster, 0.716 for random forest and 0.709 for logistic regression. Most reported fusion increments are negative. For a 20 percentage-point drop combined with a final solvency ratio below 200%, the best fusion rule gives an AUC increment of 0.004 over the best single model. Winsorising the predictors leaves the main conclusion unchanged.

Table 8 Window 1 operational summary. Thresholds and reliability are chosen without test labels.

Method	AUC	PR-AUC	Brier	Prec. at 5%	Prec. at 10%	n to 75% events
Logistic	0.727	0.436	0.218	0.582	0.495	1,426
Forest	0.710	0.413	0.218	0.518	0.526	1,487
Boosting	0.692	0.405	0.220	0.532	0.502	1,594
Soft vote	0.717	0.431	0.213	0.596	0.544	1,457
Dempster	0.718	0.432	0.215	0.596	0.544	1,468

5.7 Decision utility

Table 8 reports operational metrics on Window 1. Dempster has higher precision than logistic regression in the top 5% and 10% of scores. To capture 75% of events, logistic regression requires 1,426 reviews, compared with 1,468 for Dempster. Precision at a fixed review share and the number of reviews needed for a target recall answer different operational questions. Decision curves and selected net-benefit values are reported in Supplementary Figure S7 and Supplementary Table S8, respectively.

5.8 Worked score

The worked example uses an anonymised undertaking. Undertaking A receives predicted probability 0.998 for a 20 percentage-point deterioration event. It is not an insolvency finding, an SCR breach or a statement of financial difficulty. The reported logistic coefficients show how each predictor contributes to the model score. Coefficients, firm-clustered standard errors and odds ratios are reported in Supplementary Table S9.

6 Discussion

The current solvency ratio is a strong benchmark for the defined deterioration event. Logistic regression is competitive in the primary temporal test, but it does not have the highest mean AUC across rolling origins.

Using all eight candidates gives Dempster an AUC increment of 0.008 [0.001, 0.014] over the best single model. The primary three-feature specification does not show this gain. The fusion result therefore depends on the feature set.

The exploratory calibrated-pooling analysis also gives positive results. GLS improves AUC by 0.010 [0.0005, 0.0203] relative to the constituent selected by validation Brier, while decision-focused pooling has the highest reported pool AUC at 0.733. The comparison with Dempster changes both source calibration and the combination rule, so it does not isolate the effect of accounting for source dependence. The counterexample and the synthetic experiment illustrate how repeated evidence can produce overconfident probabilities.

The probability and rank findings differ. The unweighted logistic regression on the selected features has a lower reported Brier score than the ratio-only probability

model, even though its AUC is lower than the current-ratio rank. Class weighting improves the reported ranking but worsens raw probability accuracy. Temporal calibration reduces the Brier score. Dempster has higher precision at the 5% and 10% review budgets, while logistic regression requires fewer reviews to capture 75% of events. A supervisory application should specify whether the output allocates a review budget or estimates event probability before choosing weights, calibration, and evaluation metrics.

Limitations

The label measures deterioration of the solvency-ratio rather than failure. Country-year macroeconomic changes do not measure firm-level duration gaps, and SCR submodules are omitted to limit missing-data selection. Hyperparameter grids are modest, and the RBF SVM and additional tree models are fixed-setting checks. Random and grouped protocols use a broader pooled panel than Window 1, so their differences also reflect the evaluation period. The main test covers two forecast origins and the rolling evaluation five; firm-clustered intervals condition on the observed years. The exploratory pools were developed after inspection of the primary test results; their intervals do not adjust for method selection or multiple comparisons. The panel does not provide an independent 2024–2025 test or external insurer sample.

7 Conclusion

Logistic regression and a linear SVM reach AUC 0.727 in the primary temporal test, compared to 0.718 for Dempster fusion. The current solvency ratio is also a strong benchmark. Using all eight candidates gives Dempster AUC 0.735 and a positive increment of 0.008 over the best single model. Calibrated pools provide positive exploratory results: GLS improves AUC by 0.010 relative to the validation-selected constituent, while decision-focused pooling reaches AUC 0.733. These findings do not establish a general advantage for any combination method: the intervals are conditional on the two test origins, and the exploratory pools were developed after inspection of the primary test results. Independent temporal evaluation is needed to assess performance in other calendar periods. Model choice should reflect ranking accuracy, calibrated probabilities and the review budget.

Data availability

The data used in this study were obtained from commercial Solvency II solo QRT extracts supplied by Insurance Risk Data (https://www.insuranceriskdata.com/). The raw data are not publicly available because the vendor license prevents redistribution of the raw files and may restrict sharing of row-level derived data. Therefore, in accordance with the applicable data license, the partial data used in the article were made available and published in the RODBUK UEK repository at:
https://uek.rodbuk.pl/dataset.xhtml?persistentId=doi:10.58116/UEK/KEC7GX.

Funding

This research was funded by the National Science Center (NCN), Poland, grant No 2024/53/B/HS4/03258 (A.D., S.W., E.R., A.W.).

Competing interests

The authors declare no competing interests.

References

[1] Xiao, Z., Yang, X., Pang, Y. & Dang, X. The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster–Shafer evidence theory. *Knowl.-Based Syst.* **26**, 196–206 (2012).

[2] Wang, S. & Chi, G. Cost-sensitive stacking ensemble learning for company financial distress prediction. *Expert Syst. Appl.* **255**, 124525 (2024).

[3] Liu, Z. & Li, D. Research of Dempster–Shafer's theory and ensemble classifier financial risk early warning model based on Benford's law. *Comput. Econ.* **65**, 3361–3389 (2025).

[4] Zhang, T., Chen, Q., He, Q. & Zhu, W. AI-driven financial risk early-warning using TD-DS-RF: a decision support model integrating multi-source data. *Decis. Mak. Appl. Manag. Eng.* **8**, 478–496 (2025).

[5] Zhu, W., Zhang, T., Wu, Y., Li, S. & Li, Z. Research on optimization of an enterprise financial risk early warning method based on the DS-RF model. *Int. Rev. Financ. Anal.* **81**, 102140 (2022).

[6] Bergmeir, C., Hyndman, R. J. & Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **120**, 70–83 (2018).

[7] López de Prado, M. *Advances in Financial Machine Learning* (Wiley, 2018).

[8] Berdin, E. & Gründl, H. The effects of a low interest rate environment on life insurers. *Geneva Pap. Risk Insur. Issues Pract.* **40**, 385–415 (2015).

[9] EIOPA. Financial stability report. Tech. Rep., European Insurance and Occupational Pensions Authority (2024).

[10] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967).

[11] Shafer, G. *A Mathematical Theory of Evidence* (Princeton University Press, 1976).

[12] Yager, R. R. On the Dempster–Shafer framework and new combination rules. *Inf. Sci.* **41**, 93–137 (1987).

[13] Murphy, C. K. Combining belief functions when evidence conflicts. *Decis. Support Syst.* **29**, 1–9 (2000).

[14] Bates, J. M. & Granger, C. W. J. The combination of forecasts. *Oper. Res. Q.* **20**, 451–468 (1969).

[15] Granger, C. W. J. Combining forecasts – twenty years later. *J. Forecast.* **8**, 167–173 (1989).

[16] Genre, V., Kenny, G., Meyler, A. & Timmermann, A. Combining expert forecasts: can anything beat the simple average? *Int. J. Forecast.* **29**, 108–127 (2013).

[17] Claeskens, G., Magnus, J. R., Vasnev, A. L. & Wang, W. The forecast combination puzzle: a simple theoretical explanation. *Int. J. Forecast.* **32**, 754–762 (2016).

[18] Stratigakos, A., Pineda, S. & Morales, J. M. Decision-focused linear pooling for probabilistic forecast combination. *Int. J. Forecast.* **41**, 1112–1125 (2025).

[19] Stock, J. H. & Watson, M. W. Combination forecasts of output growth in a seven-country data set. *J. Forecast.* **23**, 405–430 (2004).

[20] Timmermann, A. in *Forecast combinations* In *Handbook of Economic Forecasting* (Elliott, G., Granger, C. W. J. & Timmermann, A. (eds)), Vol. 1 135–196 (Elsevier, 2006).

[21] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).

[22] Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 625–632 (2005).

[23] Möhlmann, A. Interest rate risk of life insurers: evidence from accounting data. *Financ. Manag.* **50**, 587–612 (2021).

[24] Zhang, L. & Nielson, N. Solvency analysis and prediction in property-casualty insurance: incorporating economic and market predictors. *J. Risk Insur.* **82**, 97–124 (2015).

[25] Brockett, P. L., Cooper, W. W., Golden, L. L. & Pitaktong, U. A neural network method for obtaining an early warning of insurer insolvency. *J. Risk Insur.* **61**, 402–424 (1994).

[26] Brockett, P. L., Cooper, W. W., Golden, L. L. & Xia, X. A case study in applying neural networks to predicting insolvency for property and casualty insurers. *J. Oper. Res. Soc.* **48**, 1153–1162 (1997).

[27] Rauch, J. & Wende, S. Solvency prediction for property-liability insurance companies: evidence from the financial crisis. *Geneva Pap. Risk Insur. Issues Pract.* **40**, 47–65 (2015).

[28] Eling, M. & Jia, R. Business failure, efficiency, and volatility: evidence from the European insurance industry. *Int. Rev. Financ. Anal.* **59**, 58–76 (2018).

[29] He, C., Huang, D., Jia, R. & Wang, X. Machine learning and insurer failure prediction. *Int. J. Forecast.* **42**, 1163–1180 (2026).

[30] Koçer, G. B. & Selcuk-Kestel, S. The assessment of early warning for insurance company using machine learning methods. *Gazi Univ. J. Sci.* **38**, 492–504 (2025).

[31] Altınçay, H. On the independence requirement in Dempster–Shafer theory for combining classifiers providing statistical evidence. *Applied Intelligence* **25**, 73–90 (2006).

[32] Yager, R. R. On the fusion of non-independent belief structures. *Int. J. Gen. Syst.* **38**, 505–531 (2009).

[33] Shafer, G. The problem of dependent evidence. *Int. J. Approx. Reason.* **79**, 41–44 (2016).

[34] Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**, e0118432 (2015).

[35] Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).

[36] Sun, X. *et al.* Disentangling multi-modal fusion for financial distress prediction with multi-level attention mechanisms. *Appl. Soft Comput.* **196**, 115085 (2026).

[37] Song, M., Li, J., Gao, F. & Hong, W. An interpretable multi-class financial crisis early warning model based on D-S evidence fusion. *J. Syst. Manag.* **35**, 452–461 (2026).

[38] Qiu, J., Xiao, Z., Xu, W. & Zhou, Y. Soft probability based random forest for financial distress prediction. *Inf. Sci.* **729**, 122870 (2026).

[39] Kostrzewa, M. *et al.* V4FinBench: Benchmarking tabular foundation models, LLMs, and standard methods on corporate bankruptcy prediction. arXiv preprint arXiv:2605.10896 (2026).

[40] Li, H. *et al.* Robust Dempster-shafer evidence fusion with chaos-conflict measurement and historical-experience weighting. arXiv preprint arXiv:2608.13108 (2026).

[41] Lee, T.-H. & Seregina, E. Combining forecasts under structural breaks using Graphical LASSO. *Int. J. Forecast.* **42** (2026).

[42] Liu, C.-A. & Vasnev, A. L. Corrected forecast combinations. arXiv preprint arXiv:2601.09999 [econ.EM] (2026).

[43] Nuthanakaluva, P. R. & Gaddam, N. K. Generalized Gibbs ensemble weighting for forecast combination. arXiv preprint arXiv:2608.28116 [cs.LG] (2026).

[44] EIOPA. Solvency II Standard Formula: the underlying assumptions. *EIOPA Technical Report* (2014). EIOPA-14-322.

[45] Gatzert, N. & Martin, M. Quantifying credit and market risk under Solvency II: standard approach versus internal model. *Insur. Math. Econ.* **51**, 649–666 (2012).

[46] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).

[47] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).

[48] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

[49] Ross, B. C. Mutual information between discrete and continuous data sets. *PLOS ONE* **9**, e87357 (2014).

[50] Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).

[51] Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

[52] Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 61–74 (MIT Press, 1999).

[53] Clark, T. E. & West, K. D. Approximately normal tests for equal predictive accuracy in nested models. *J. Econom.* **138**, 291–311 (2007).

[54] Smets, P. Decision making in the TBM: the necessity of the pignistic transformation. *Int. J. Approx. Reason.* **38**, 133–147 (2005).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarySR.pdf