

Article

Network Instability as a Signal of Systemic Financial Stress: An Explainable Machine-Learning Framework

Livia Valentina Moretti , Enrico Barbierato *  and Alice Gatti 

Dipartimento di Matematica e Fisica, Università Cattolica del Sacro Cuore, Via Garzetta 48, 25121 Brescia, Italy; liviavalentina.moretti01@icatt.it (L.V.M.); alice.gatti@unicatt.it (A.G.)

* Correspondence: enrico.barbierato@unicatt.it

Abstract

This paper develops a framework for monitoring and forecasting episodes of systemic financial stress using a combination of market information, macro-financial indicators, and measures derived from time-varying correlation networks, embedded in a sequential machine-learning setting. The contribution is not tied to a single modelling innovation, but rather to the way these ingredients are brought together under an evaluation protocol designed to mimic real-time supervisory use, and to an interpretability layer that makes the resulting predictions easier to inspect. Monthly data covering the period from 2006 to 2025 are used to construct evolving correlation structures and summary indicators of market co-movement. These features are combined with standard predictors and fed into logistic regression, random forest, and gradient boosting models, all estimated in expanding windows and assessed strictly on future observations. Predictive accuracy remains limited, which is consistent with the difficulty of anticipating stress regimes several months ahead at monthly frequency, although gradient boosting attains the highest average AUC across evaluation folds and displays noticeable variation over time. Inspection of SHAP values points to instability in correlation networks, volatility conditions, and short-horizon return behaviour as recurring drivers of the predicted stress probabilities, suggesting that the models draw on information that goes beyond individual market series. Taken together, the results indicate that recurrent statistical regularities and changes in market structure can be exploited for monitoring purposes when models are trained and tested in a sequential fashion. The overall design is intended to be usable in practice and to support supervisory analysis, while remaining transparent enough to allow scrutiny of the signals driving the forecasts.

Keywords: financial networks; systemic fragility; network-based early-warning systems



Academic Editor: Paolo Bellavista

Received: 13 December 2025

Revised: 25 January 2026

Accepted: 5 February 2026

Published: 9 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Anticipating episodes of systemic financial stress remains a central challenge for economists, policymakers, and supervisory authorities. Traditional econometric forecasting tools have often struggled in this context, as they typically abstract from nonlinear interactions, evolving dependencies, and structural interconnections across financial markets and macroeconomic conditions. As a result, major crises have frequently materialised without clear signals from standard early-warning indicators. In response, recent research has increasingly explored machine-learning techniques and network-based representations as complementary tools for systemic-risk monitoring, as they offer greater flexibility in

capturing complex and time-varying relationships that are difficult to model within linear frameworks [1,2].

From a conceptual perspective, it is important to distinguish between prediction and monitoring. Models are abstractions of real systems designed to represent structural relationships and dynamic behaviour, while predictions are inherently uncertain statements about future outcomes. Stochastic and probabilistic modelling provide a principled way to quantify such uncertainty, particularly when the underlying processes are driven by heterogeneous agents and feedback effects [3,4]. Financial crises represent extreme and persistent disruptions that impair the functioning of markets and institutions [5,6]. Although definitions vary, empirical studies consistently document recurrent pre-crisis conditions, including rapid credit expansion, excessive leverage, asset-price booms, weak supervisory oversight, and heightened co-movement across financial markets [7,8]. These regularities motivate the development of analytical frameworks aimed not at precise crisis forecasting, but at probabilistic monitoring and early detection of stress regimes.

Two strands of empirical literature are particularly relevant to this study. The first investigates the informational content of financial network structures. Network-based approaches model markets as systems of interconnected entities and show that changes in correlation patterns, connectivity, and synchronisation often accompany periods of heightened systemic stress. For instance, Peralta et al. [9] demonstrate that evolving partial-correlation networks contain useful information about emerging instability, although strong distributional assumptions and limited temporal resolution constrain interpretability and operational deployment. The second strand focuses on distress prediction at the firm or sectoral level. Studies such as Giannopoulos et al. [10] show that balance-sheet indicators and market-based measures can anticipate individual failures, but they typically rely on feature sets that abstract from macro-financial context and systemic interdependencies.

Building on these insights, this study proposes an integrated framework that combines market variables, macro-financial indicators, and dynamic correlation-based network features within a unified probabilistic monitoring and interpretability pipeline. Rather than aiming at deterministic crisis prediction or claiming large gains in forecasting accuracy, the framework is designed to assess whether recurrent statistical and structural patterns associated with systemic stress can be identified several months ahead when models are evaluated under realistic, real-time informational constraints. Two research questions guide the analysis. First, does the joint use of market, macro-financial, and network-derived features yield stable and informative signals about systemic stress regimes when assessed sequentially using expanding-window estimation? Second, can explainability tools clarify how these heterogeneous sources of information contribute to model-implied stress probabilities in a manner suitable for supervisory interpretation?

The methodological contribution of the paper lies in the simultaneous integration of three components that are rarely combined within a single empirical pipeline. The first is a dynamic representation of financial interdependencies, constructed from monthly correlation networks that capture changes in market synchronisation and structural fragility over time. The second is a nested expanding-window evaluation design using logistic regression, random forests, and gradient boosting, which explicitly avoids look-ahead bias and reflects the informational constraints faced by real-time monitoring systems. The third is a systematic interpretability layer based on SHAP values, applied to a representative nonlinear specification in order to decompose predicted stress probabilities into feature-level contributions and to assess the relative importance of market, macro-financial, and network indicators, rather than to assert the dominance of any single algorithm.

The framework is not intended to replace judgement or policy analysis, nor to deliver precise crisis forecasts. Instead, it provides a coherent way to organise heterogeneous infor-

mation sources, quantify uncertainty, and support decision-making in financial-stability surveillance. By explicitly combining sequential evaluation, network-based representations, and explainable machine learning, the study contributes to the growing literature on probabilistic tools for monitoring systemic financial stress under real-world constraints.

The paper is organised as follows. Section 2 recaps the scientific background. Section 3 comments on the scientific literature. Section 4 presents the models and the experiments, while Section 5 draws the conclusions.

2. Background

The task of predicting crises is framed as a supervised classification problem in which each month is assigned to either a stress or non-stress state. Labels are constructed from extreme negative returns and high volatility, using empirical quantiles to define the relevant thresholds, and are subsequently smoothed to capture the persistence that typically characterises episodes of systemic strain. To align the exercise with the horizons at which financial-stability authorities operate, the indicator is shifted three months ahead, so that models are asked to forecast rather than to contemporaneously identify stress. The focus is therefore on probabilistic early signals of deteriorating conditions, rather than on pinpointing isolated or externally triggered events.

Since the labels themselves are derived from market data, particular attention is paid to avoiding mechanical overlap between predictors and targets. The forecasting horizon is displaced forward relative to the available information set, and all data transformations and feature-engineering steps are carried out within expanding historical samples. Additional checks exclude subsets of return- and volatility-based variables to verify that predictive performance is not driven entirely by components that are closely tied to the crisis definition.

All models are estimated under a nested expanding-window scheme intended to approximate real-time use. At each point in the evaluation, only information that would have been available at that date enters the analysis. Hyperparameters are selected through an inner cross-validation loop within the training window, while performance is measured on the subsequent period. This setup mirrors the constraints faced in operational monitoring environments and guards against the inadvertent use of future information in either fitting or model selection.

A key ingredient of the framework is the construction of financial networks from rolling-window correlation matrices. These networks provide a compact description of how markets move together over time and of how tightly institutions appear to be linked. A large empirical literature documents that the run-up to major crises is often accompanied by rising synchronisation, a weakening of diversification benefits, and the growing influence of a small number of common factors, sometimes associated with funding pressures or fire-sale dynamics. From this perspective, correlation-based network measures are treated as statistical summaries of such latent forces rather than as structural representations of causal channels.

More concretely, an increase in the leading eigenvalue of the correlation matrix signals the growing dominance of a market-wide component, in line with flight-to-quality episodes or broad deleveraging. Network density and average absolute correlations track the overall tightening of linkages across assets and balance sheets, while dispersion in node degrees or entropy captures whether systemic importance is becoming concentrated in a limited set of institutions. Sudden movements in these indicators can therefore be read as early statistical warnings of mounting fragility, even though the economic triggers behind them may differ across episodes.

Three families of predictive models are examined. Logistic regression serves as a linear baseline, with crisis probabilities for an observation x given by

$$P(Y = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \beta^\top x)}}.$$

Its main attraction is the ease with which coefficients can be interpreted, though this comes at the cost of limited ability to represent nonlinearities.

Random forests build on decision trees by averaging across many trees estimated on bootstrap samples and random subsets of covariates. Denoting the output of tree t by $f_t(x)$, the ensemble prediction takes the form

$$F(x) = \frac{1}{T} \sum_{t=1}^T f_t(x).$$

This aggregation reduces the instability of single trees and allows the classifier to pick up interaction effects and threshold-type behaviour that are common in financial time series.

Gradient boosting proceeds differently, constructing an additive model by fitting successive trees to the negative gradient of a chosen loss function. Under logistic loss,

$$F_m(x) = F_{m-1}(x) + \nu h_m(x),$$

where h_m denotes the newly added tree and ν is a learning-rate parameter. Through this incremental refinement of the decision surface, boosting is well suited to settings marked by regime changes and shifts in dependence patterns.

Model performance is evaluated primarily through the Area Under the Receiver Operating Characteristic Curve, which summarises how well a classifier ranks crisis months above tranquil ones across all possible thresholds. The ROC curve traces the true positive rate $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$ against the false positive rate $\text{FPR} = \text{FP}/(\text{TN} + \text{FP})$. An AUC of 0.5 corresponds to random ordering, while values only moderately higher are often viewed as informative in macro-financial applications, where class imbalance and noisy signals are the norm. To convey stability, results are reported as averages and dispersions of out-of-sample AUC across evaluation folds, together with confidence intervals.

Because supervisors ultimately act on probabilities rather than rankings alone, calibration is also examined. Reliability diagrams compare predicted and realised crisis frequencies within probability bins, making it possible to assess discrimination and probability accuracy side by side and to highlight tensions between the two.

Interpretability is addressed through SHAP values, which express predictions as additive contributions of individual features derived from cooperative game theory. For a model f and an input x ,

$$f(x) = \phi_0 + \sum_{i=1}^d \phi_i$$

decomposes the output into a baseline term and feature-specific effects. Explanations are computed for the final model refitted on the full sample using the selected hyperparameters so that they correspond to the reported predictive performance. Aggregate summaries highlight the variables that most strongly influence estimated crisis probabilities, while dependence plots show how changes in key inputs translate into shifts in predicted risk. These results are interpreted as descriptions of the model's internal logic rather than as evidence of causal economic mechanisms.

3. Related Work

A growing literature on systemic risk emphasises the role of interconnectedness, contagion mechanisms, and structural fragility in shaping the dynamics of financial systems. Numerous studies investigate how distress spreads across institutions and how the configuration of financial linkages affects the scale and persistence of crises. Fasano et al. [11], for instance, analyse corporate default prediction using recurrent neural networks combined with SHAP-based explanations. They show that deep models can uncover complex temporal patterns in firm-level data while still offering insight into the variables associated with distress. Their contribution is relevant here not only because it illustrates the importance of nonlinear interactions, but also because it underscores the need for interpretability in early-warning applications. Although their empirical focus is microeconomic, the methodological message carries over to macro-financial monitoring.

Nyabera et al. [12] propose an adaptive neural network architecture for detecting distress in Kenyan microfinance institutions and argue that flexible models are particularly useful in environments subject to structural change. At the same time, their setting is constrained by limited sample sizes and by the absence of macroeconomic or network-based covariates, which points to the value of embedding firm-level information within a broader systemic perspective. The present study follows this line by combining market-wide indicators, macro-financial conditions, and correlation-based network measures in a single predictive framework.

Dastkhan et al. [13] examine idiosyncratic volatility in a network of firms listed on the Tehran Stock Exchange and document that firms exposed to such shocks tend to occupy central positions in the market network and to earn higher abnormal returns. Their findings draw attention to the role of network topology and spectral properties in shaping contagion. This motivates the use of statistics such as mean absolute correlation, network density, and movements in the leading eigenvalue as reduced-form indicators of collective stress, ideas that are taken up here through the construction of rolling correlation networks whose evolution is treated as a potential source of predictive information.

Another strand of research focuses more directly on forecasting systemic events using network-derived variables. Ren et al. [14] build spillover networks across Chinese industries and apply a gated graph neural network to classify crisis periods, showing that explicit modelling of sectoral interactions can improve predictive accuracy. The complexity of such models and the difficulty of interpreting their outputs, however, limit their immediate usefulness in supervisory contexts. In contrast, the present analysis relies on classifiers such as logistic regression, random forests, and gradient boosting, which allow network features to be incorporated while keeping the mapping from inputs to predictions comparatively transparent.

Evidence on the descriptive behaviour of financial networks around crises is provided by Chowdhury et al. [15], who document sharp increases in interconnectedness across Asian and global markets during major stress episodes. Their results support the view that changes in market structure often precede systemic instability, an observation that motivates the empirical strategy pursued here, where shifts in clustering, co-movement, and connectivity are examined as potential early-warning signals.

Related insights appear in Chowdhury et al. [16], who analyse co-exceedances and lead–lag relations in the Shanghai Stock Exchange using network representations to infer conditions of systemic risk. Although their approach relies on high-frequency data and does not incorporate macroeconomic drivers, it illustrates how latent channels of contagion can be extracted from patterns of dependence. The present study complements this perspective by working at a lower frequency but drawing on a richer set of covariates that includes network, macro-financial, and market-based information.

A further antecedent is Guritanu et al. [17], who introduce a topological machine-learning approach to crisis detection based on persistent homology, persistence diagrams, and landscape-based classifiers. They show that topological signatures computed over rolling windows of financial time series can act as early indicators of systemic stress, capturing changes in the global geometry of the data. At the same time, their framework relies on specialised constructions, offers limited economic interpretability, and does not incorporate macro-financial variables or explicit network structures. The present work draws inspiration from their emphasis on structural change but replaces topological summaries with correlation-network indicators and combines them with explainable machine-learning tools. Whereas the earlier contribution focuses on geometric anomalies in time series, the current analysis concentrates on shifts in co-movement patterns, eigenvalue dynamics, and macro-financial conditions, with the aim of producing signals that are easier to relate to supervisory concerns.

It can be seen that research points to several recurring trade-offs. Deep learning approaches often achieve strong predictive performance but provide little transparency; firm-level studies highlight heterogeneity but leave aggregate systemic dynamics in the background; network analyses document structural fragility yet are sometimes short on forecasting evidence; and topological methods uncover global patterns while offering limited guidance for policy interpretation. The framework proposed here is designed to sit across these literature by combining correlation-network measures, macro-financial indicators, and explainable machine-learning models. In doing so, it seeks to link predictive accuracy with structural diagnostics and to deliver an early-warning system that remains empirically grounded and usable in supervisory settings.

4. Models

The empirical setup is organised to resemble a real-time monitoring environment, where the information set at any point consists only of what has been observed up to that date and forecasts are revised sequentially as new data arrive. The dataset contains monthly observations from January 2006 to April 2025, a span that includes several widely recognised stress episodes, such as the Global Financial Crisis, the European sovereign debt crisis, and the COVID-19 shock. The market block includes institution-level series such as Credit Default Swap spreads, equity returns, and price dynamics for a broad set of internationally active intermediaries. The macro-financial block covers standard aggregates, including labour-market conditions, interest rates, inflation measures, and major equity indices. All variables are aligned to a common monthly calendar and stored in a panel structure, which simplifies feature engineering and supports expanding-window estimation. For reproducibility, the workflow is driven by a single configuration object that records data paths, random seeds, preprocessing choices, correlation thresholds, and model hyperparameters. Figure 1 depicts the proposed model structure discussed in this work.

Before model estimation, the raw data are harmonised and cleaned. Missing values are handled through Multivariate Imputation by Chained Equations, implemented as an iterative procedure in which each variable is modelled conditionally on the others and missing entries are updated until the process stabilises. Relative to mean imputation or ad hoc interpolation, this approach better preserves multivariate dependence and avoids systematically dampening volatility. Observations identified as anomalous via isolation-based detection are treated as missing and subsequently imputed. The aim is not to remove economically meaningful extremes, but to prevent a small number of outliers from mechanically dominating rolling correlations while keeping the time series continuous. Because stress episodes can themselves generate extreme observations that may be informative, we also assess sensitivity to this step by re-running the full pipeline with anomaly detection

disabled, keeping all observations before the imputation. The qualitative performance profile is unchanged, while the baseline specification yields more stable network estimates and less fold-to-fold variability. This pattern suggests that the anomaly-handling step mainly reduces spurious artefacts rather than erasing genuine stress signals.

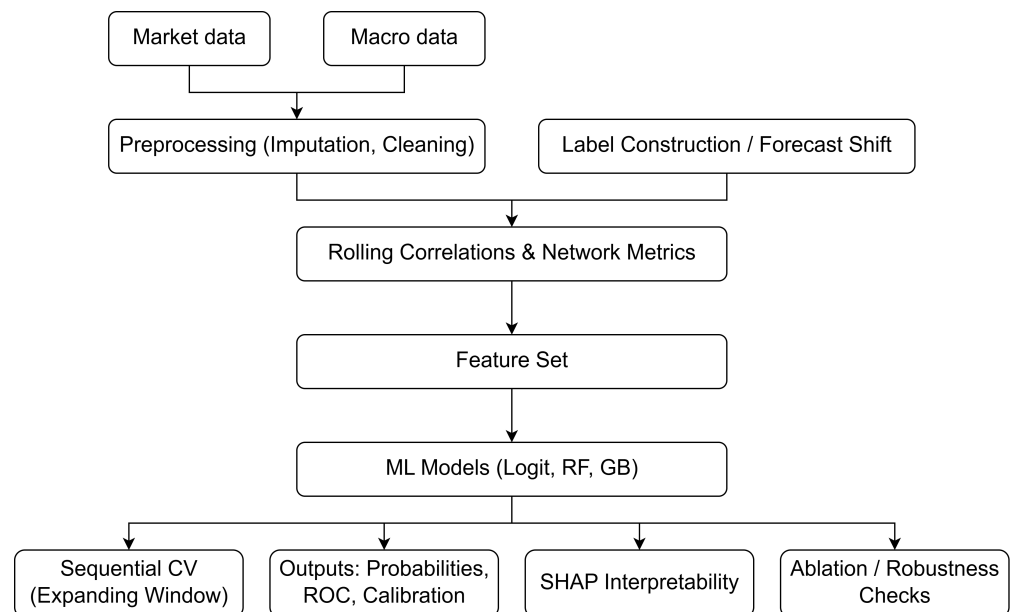


Figure 1. Model workflow.

After preprocessing, rolling correlation matrices are computed using Pearson coefficients. Pearson correlation is used for pragmatic reasons: it is easy to interpret, numerically stable at monthly frequency, and inexpensive to re-estimate repeatedly within expanding-window evaluation. To limit estimation noise and avoid a fully connected graph, correlations are sparsified using a fixed absolute threshold. The resulting matrices are visualised via hierarchical clustering. Figure 2 shows the clustered correlation heatmap after cleaning and thresholding. The visible block structure is consistent with sectoral grouping and time variation in co-movement, and it provides a compact representation of evolving market interdependence that motivates the extraction of network summaries.

4.1. Network Construction Choices and Robustness

The baseline network is built from Pearson correlations with a fixed absolute threshold, which offers a transparent and computationally stable rule for edge selection under monthly data and repeated re-estimation. Alternative constructions could, however, alter edge formation and therefore affect derived network metrics. Partial-correlation networks, for example, can reduce indirect dependence but typically require regularisation and may become sensitive to estimation error when the rolling window is short relative to the number of series. Shrinkage covariance estimators can stabilise correlations but introduce additional tuning decisions and make the mapping from data to edges less direct. Dynamic conditional correlation (DCC) and related multivariate volatility models provide parametric time-varying dependence estimates, yet they impose stronger modelling assumptions and are substantially more expensive when re-estimated sequentially. For these reasons, Pearson-threshold networks are adopted as a conservative baseline. Extending the framework to partial correlations, shrinkage-based correlations, or DCC-type estimates is a natural robustness direction, particularly to evaluate whether instability indicators remain informative when dependence is measured conditionally rather than marginally.

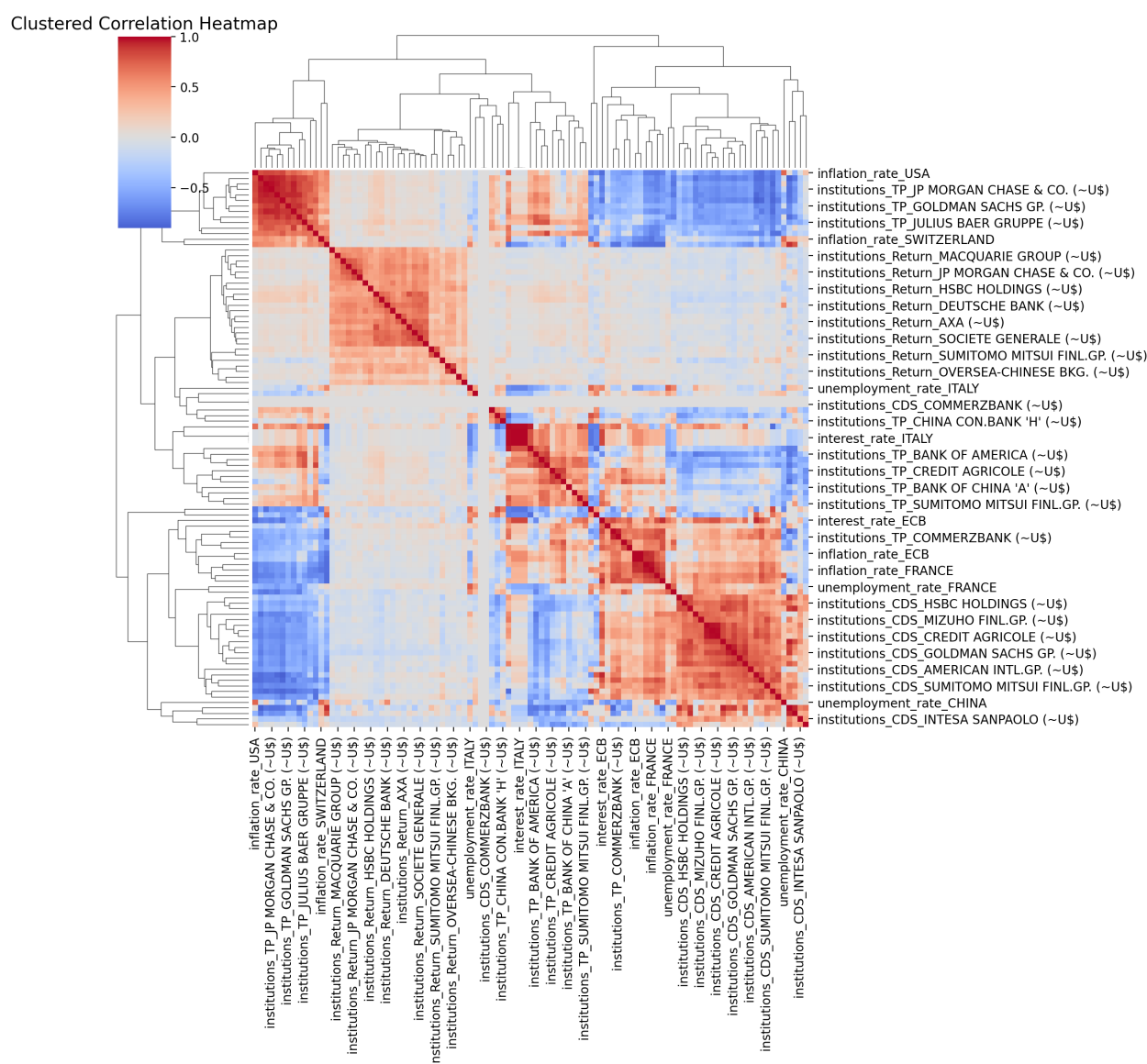


Figure 2. Hierarchically clustered correlation heatmap after data cleaning and thresholding. Block structures reveal latent sectoral clustering and evolving co-movement patterns.

Predictors are computed on rolling windows of twenty-four months and include standard return- and volatility-based measures, relative changes, and network-derived metrics summarising structural properties of the correlation graph. The network block includes mean absolute correlation, network density, degree entropy, the giant-component ratio, and the first difference in the largest eigenvalue of the correlation matrix. These variables are designed to track changes in synchronisation, the erosion of diversification, and abrupt reconfigurations in dependence patterns. Throughout, they are treated as descriptive proxies for evolving market structure rather than as causal drivers of systemic events.

Crisis labels are constructed using a joint-tail rule based on extreme negative returns and elevated volatility, calibrated through empirical quantiles and smoothed to reflect the persistence of stress regimes. Labels are then shifted forward by three months to match a forecasting horizon typical of early-warning practice. This produces a supervised sample of 207 monthly observations, with crisis months representing roughly 35% of the dataset. The exercise should therefore be read as forecasting the onset or continuation of stress regimes, not as predicting isolated, exogenous shocks.

Predictive modelling is carried out using three complementary specifications: regularised logistic regression, random forests, and gradient boosting. Logistic regression serves as a transparent benchmark, random forests allow for nonlinear interactions through ensemble averaging, and gradient boosting refines decision boundaries sequentially to accommodate complex dependencies. All models are trained and evaluated using a nested expanding-window cross-validation scheme. In each outer fold, the historical sample available at that point is split into a training window and a subsequent test window, mirroring a deployment setting. Within the outer training window, hyperparameters are selected through an inner expanding-window cross-validation loop by maximising the mean inner-fold Area Under the Receiver Operating Characteristic Curve (AUC). The model is then refitted on the full outer training window using the selected hyperparameters and evaluated on the held-out outer test window. This design is explicitly aimed at avoiding look-ahead bias and ensuring that tuning decisions depend only on information that would have been available at prediction time.

Figure 3 summarises model performance by reporting mean outer-fold AUCs across expanding-window splits, together with dispersion measured by the standard deviation across folds. Discrimination remains moderate and varies across windows, which is not unexpected given the limited number of crisis observations in early test periods and the broader difficulty of forecasting systemic stress at monthly frequency. In this setting, crises reflect layered interactions between markets and macroeconomic conditions, so large and stable gains from any single classifier are unlikely.

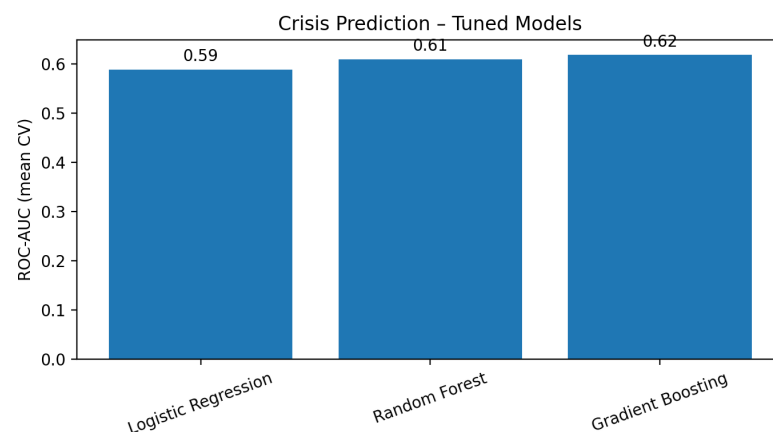


Figure 3. Mean outer-fold AUC (with dispersion across folds) for logistic regression, random forest, and gradient boosting under expanding-window evaluation.

Table 1 reports the same information in numeric form. For each outer test window, the AUC is computed on strictly out-of-sample observations. The mean AUC is the average of these fold-level values, while variability is summarised by their standard deviation. In practice, this dispersion is shaped by changes in sample size and class balance across time, with early windows particularly exposed to small-sample effects.

Table 1. Mean and variability of out-of-sample AUC across expanding-window evaluation folds. Variability is measured as the standard deviation of fold-level AUC values.

Model	Mean AUC	Std. Dev. (AUC)
Logistic Regression	0.43	0.39
Random Forest	0.56	0.25
Gradient Boosting	0.53	0.07

4.2. Statistical Comparison of AUC Across Classifiers

Since all classifiers are evaluated on the same sequence of expanding-window test periods, their ROC curves are correlated. To assess whether observed differences in AUC are statistically meaningful, pairwise DeLong tests for correlated ROC curves are applied using pooled out-of-sample predictions across outer folds. (Pooled out-of-sample predictions are obtained by concatenating predicted scores and true labels from the outer test windows. This preserves strict out-of-sample evaluation while yielding a single set of paired predictions per model for statistical comparison). Table 2 reports AUC estimates, pairwise differences, z -statistics, and two-sided p -values. None of the comparisons rejects equality at conventional significance levels. As a result, while ensemble methods achieve slightly higher point estimates, the evidence does not support strong claims of algorithmic dominance in this sample. For the remainder of the analysis, the emphasis is therefore placed on stability and interpretability rather than on ranking methods by small differences in average AUC.

Table 2. Pairwise DeLong tests for differences in AUC using pooled out-of-sample predictions across expanding-window folds.

Model 1	Model 2	AUC ₁	AUC ₂	Δ AUC	z	p -Value
Logistic Regression	Random Forest	0.471	0.553	−0.082	−1.77	0.077
Random Forest	Gradient Boosting	0.553	0.501	0.052	1.54	0.123
Logistic Regression	Gradient Boosting	0.471	0.501	−0.030	−0.51	0.613

4.3. Fold-Averaged AUC Versus Pooled Out-of-Sample AUC

Table 1 reports AUC as the arithmetic mean of fold-level values, which makes temporal variation visible because each test window corresponds to a distinct historical period with its own size and class balance. The DeLong tests in Table 2 instead operate on pooled out-of-sample predictions obtained by concatenating the outer-fold test predictions into a single score vector per model. Reporting both is useful in this application. The fold-averaged view highlights instability and small-sample effects, which are particularly pronounced in early windows, while the pooled view provides a single overall discrimination summary and supports formal testing under correlated ROC curves without violating the out-of-sample nature of each prediction. Unless otherwise stated, references to AUC below indicate explicitly whether they refer to fold-averaged or pooled computations.

To illustrate discrimination across thresholds, mean ROC curves aggregated over outer folds are reported in Figures 4–6. Fold-specific true-positive rates are interpolated on a common false-positive-rate grid and then averaged, producing a summary curve that remains consistent with the time-ordered evaluation protocol. Logistic regression provides a relatively smooth but moderate curve, consistent with linear decision boundaries. Random forests show greater fold-to-fold variability, reflecting sensitivity to data partitions and interaction effects. Gradient boosting exhibits a steeper initial rise, indicating comparatively stronger discrimination at low false-positive rates, a region often of interest in supervisory screening.

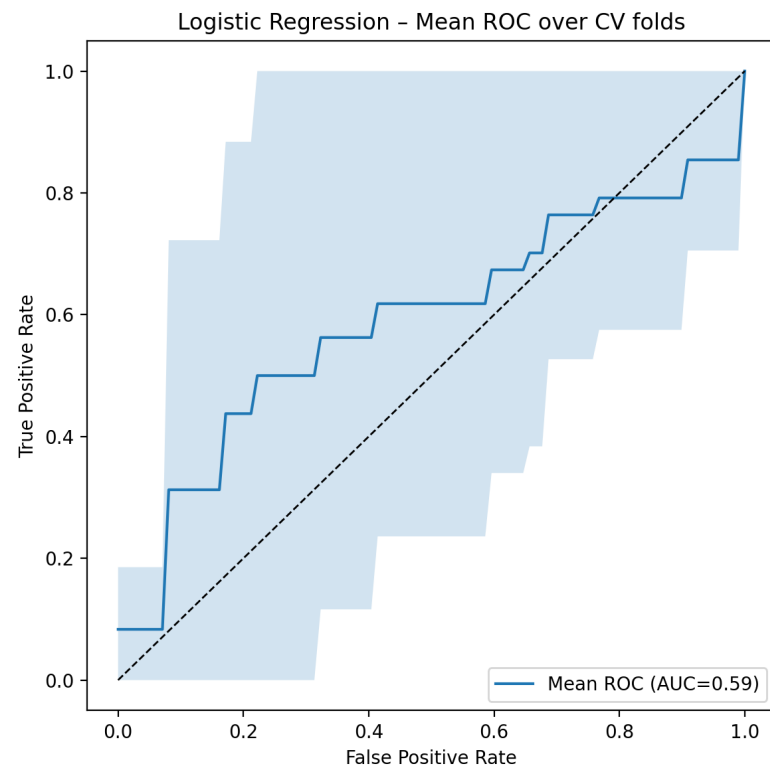


Figure 4. Mean ROC curve averaged across expanding-window out-of-sample folds for logistic regression.

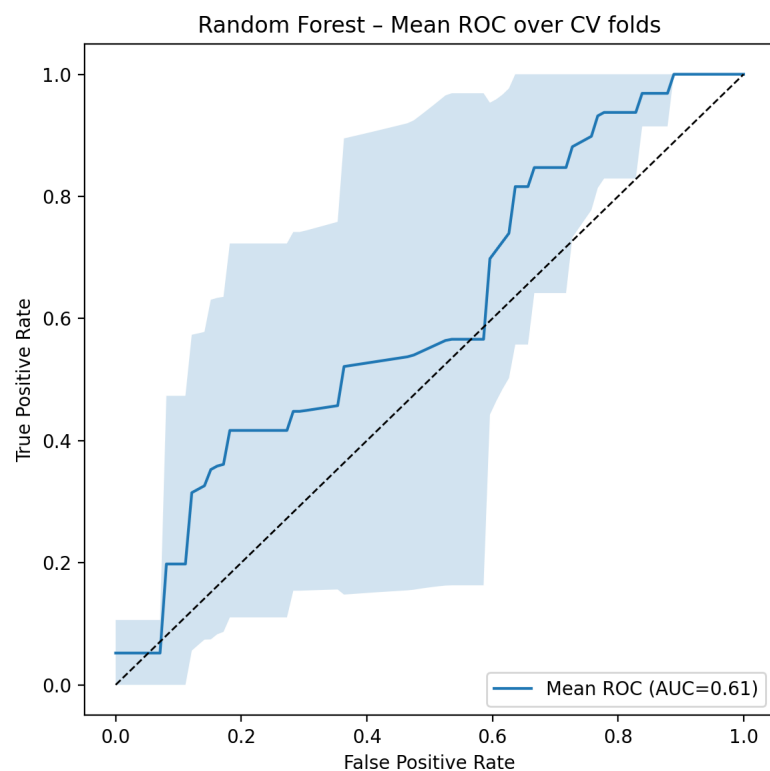


Figure 5. Mean ROC curve averaged across expanding-window out-of-sample folds for the random forest classifier.

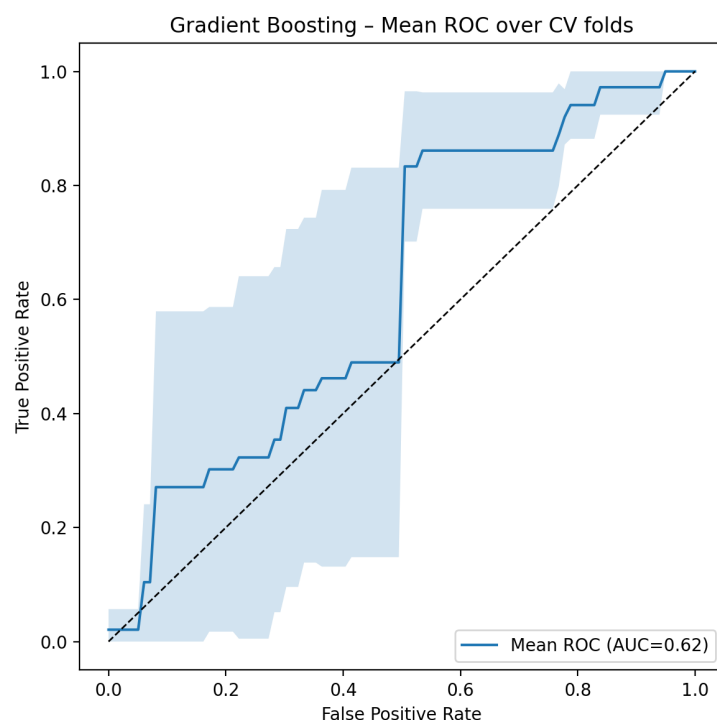


Figure 6. Mean ROC curve averaged across expanding-window out-of-sample folds for gradient boosting.

Calibration is assessed using the reliability diagrams in Figures 7–9. All curves are computed from strictly out-of-sample predictions: within each test fold, predicted probabilities are binned and compared to realised crisis frequencies, without applying any post hoc calibration procedure. Logistic regression tends to deliver comparatively well-calibrated probabilities over most of the support. Random forests display a tendency to overstate risk at high predicted probabilities, while gradient boosting sits between the two, consistent with its optimisation for ranking rather than probability calibration.

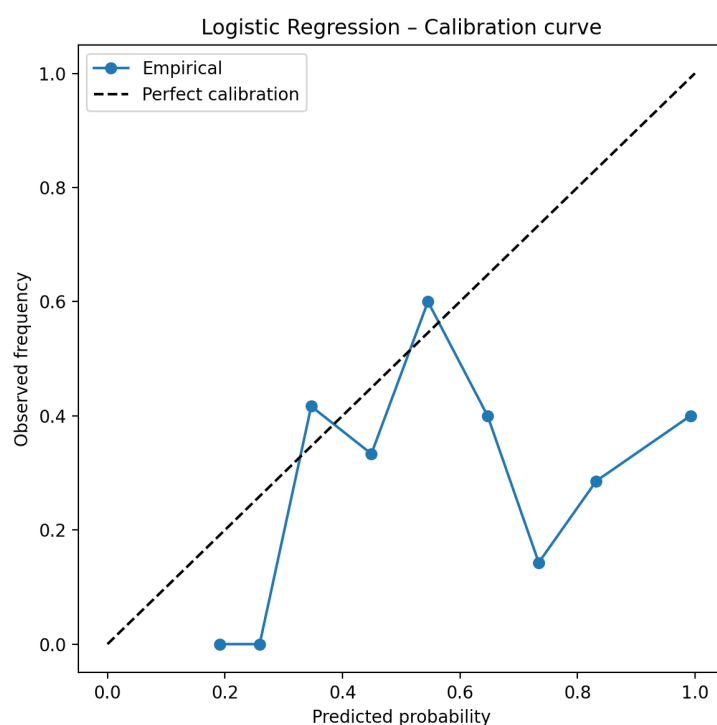


Figure 7. Calibration plot for logistic regression.

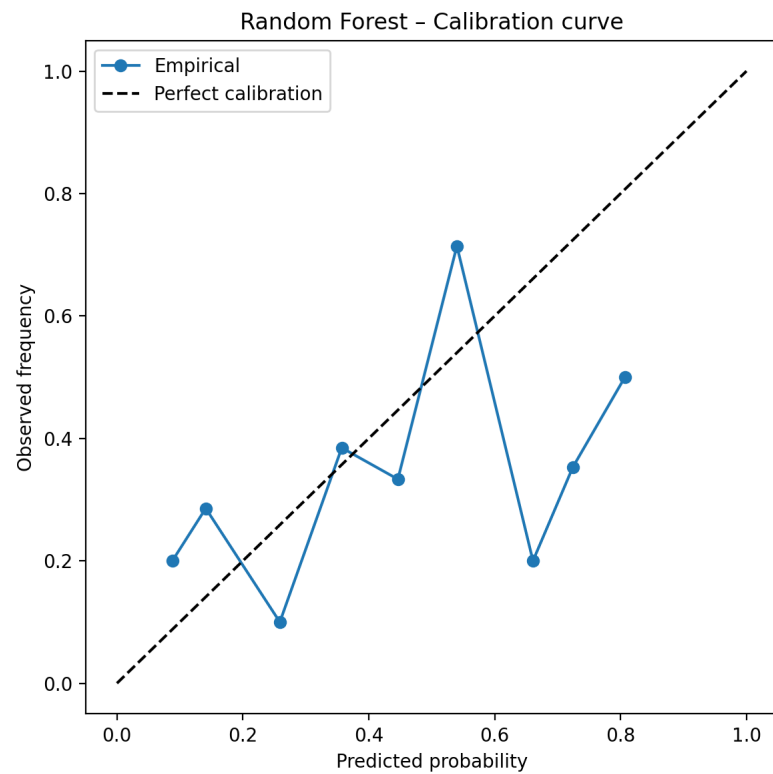


Figure 8. Calibration plot for the random forest classifier.

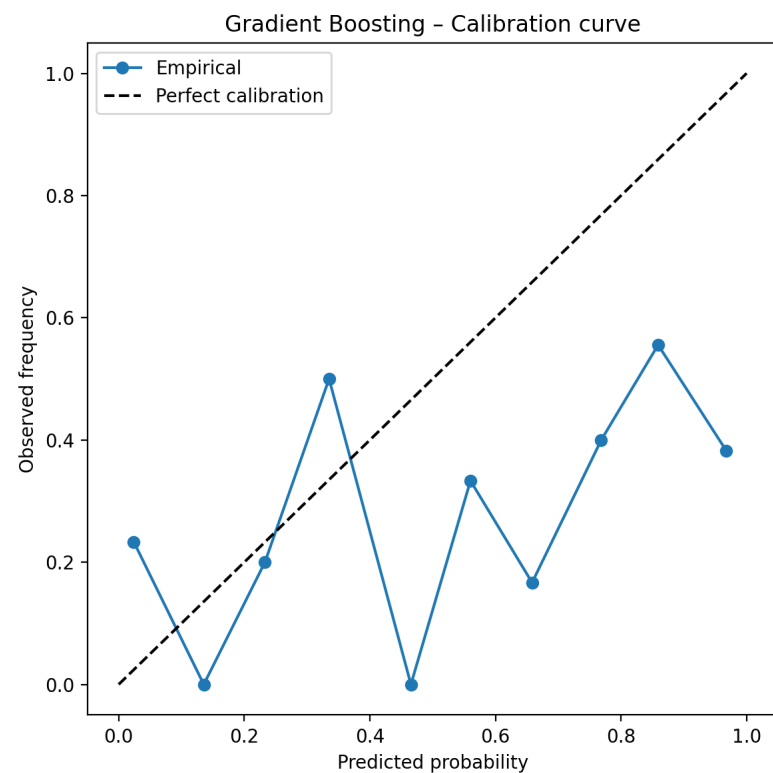


Figure 9. Calibration plot for gradient boosting.

The stability of feature importance is examined across outer folds. Figures 10–12 show persistent contributions from macro-financial indicators, while network-derived variables become more prominent in the nonlinear specifications. In the gradient-boosting model, predictive weight concentrates on a relatively small subset of predictors, suggesting that stress regimes, at this frequency, are associated with a handful of dominant signals rather than with diffuse information spread evenly across the feature space.

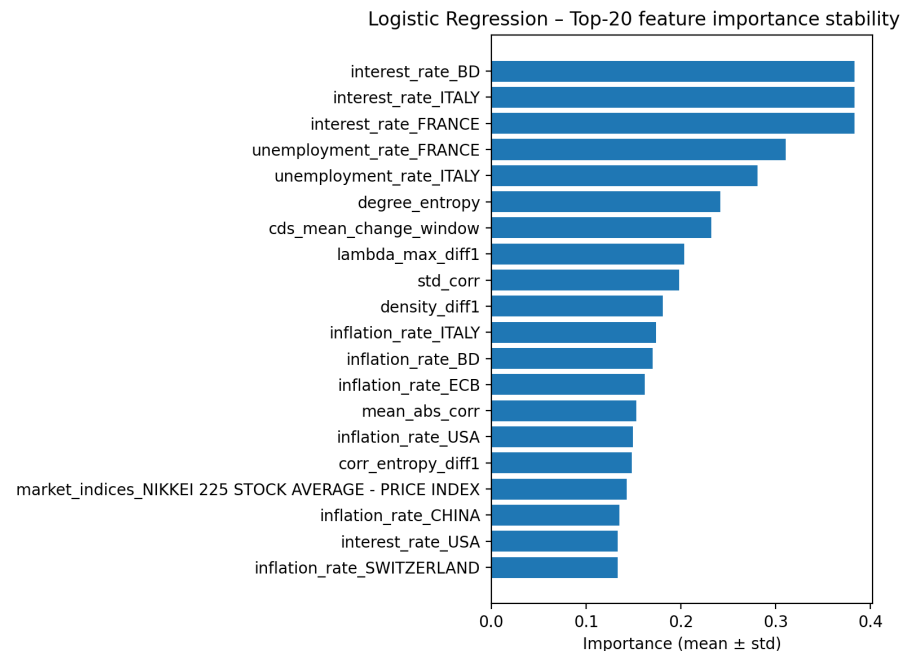


Figure 10. Feature-importance stability for logistic regression.

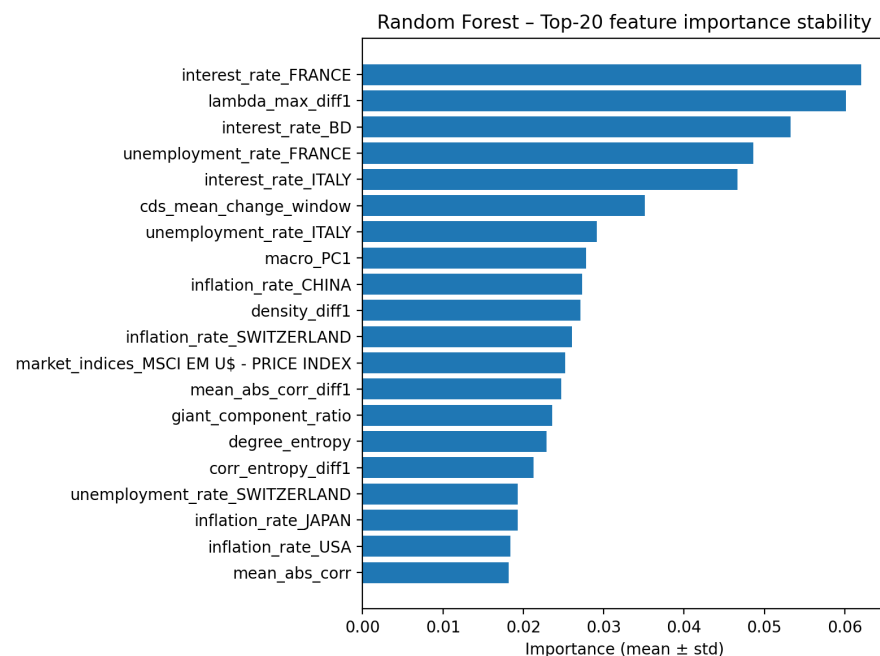


Figure 11. Feature-importance stability for the random forest.

In Figures 10–12, bar heights represent mean feature importance across expanding-window folds, while error bars denote one standard deviation, intended as a stability summary rather than as additional point estimates.

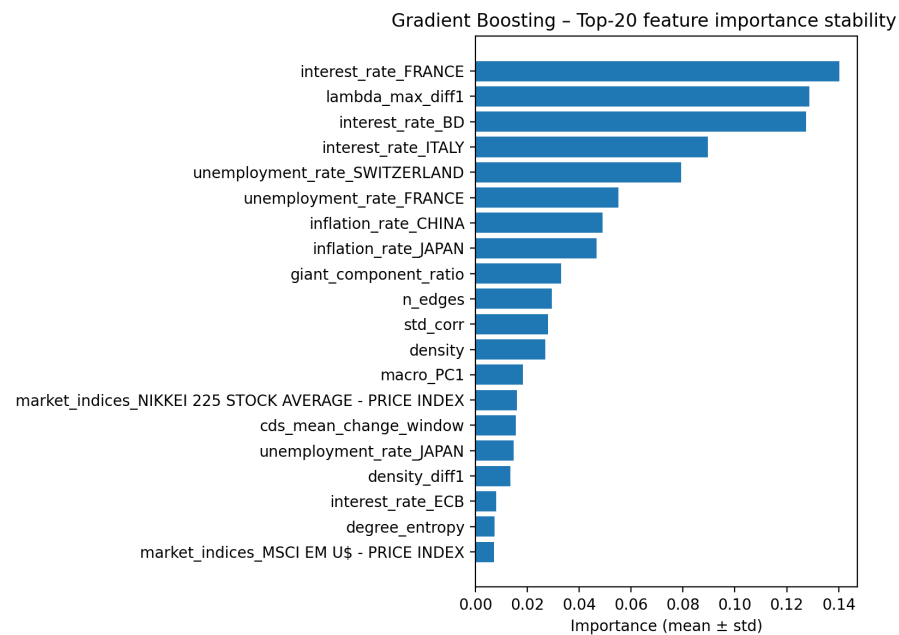


Figure 12. Feature-importance stability for gradient boosting.

4.4. Choice of Model for SHAP Analysis

SHAP-based interpretation is reported for the gradient-boosting classifier because it provides a flexible nonlinear mapping while exhibiting comparatively stable out-of-sample behaviour (Table 1). At the same time, Table 2 shows that differences in AUC across classifiers are not statistically significant in this sample, so the SHAP results should be read as an interpretability case study for a representative nonlinear model rather than as an argument that gradient boosting is uniquely superior. Analogous explanations can be produced for the other classifiers within the same pipeline but are omitted to keep the exposition focused.

Model interpretability is examined using SHAP values for the gradient-boosting classifier refitted on the full sample after applying the same preprocessing and feature-engineering steps used in the predictive experiments. Missing values are imputed before the SHAP computation using the same iterative procedure adopted during training. Global SHAP importance (Figure 13) highlights volatility-sensitive indicators, short-horizon return dynamics, and network-instability measures as influential inputs. The SHAP beeswarm plot (Figure 14) indicates that extreme values of these variables are associated with large positive contributions to predicted crisis probabilities. Figure 15 provides a more detailed view for a key network-instability measure, showing a nonlinear but broadly monotone relationship with predicted risk. These explanations describe how the model combines heterogeneous information sources; they are not interpreted as causal statements about the economic origins of crises.

4.5. Ablation Study: Contribution of Network Features

To examine what the network block contributes in predictive terms, an ablation study is conducted on the gradient-boosting specification. Three feature sets are evaluated under the same nested expanding-window protocol: (i) macro-financial features only, (ii) network-based features only, and (iii) the full specification combining macro-financial and network indicators. Table 3 reports mean out-of-sample AUC, fold-to-fold variability, and confidence intervals. In this sample, all three variants deliver identical discrimination. The natural reading is that, at monthly frequency, the correlation-network summaries largely encode information that overlaps with the macro-financial block rather than providing an independent source of incremental AUC. Importantly, adding network features does not

destabilise the model or degrade performance. Their role is therefore better understood as representational: they provide a structural description of co-movement and synchronisation that is useful for diagnosis and interpretation, even when it does not translate into measurable gains in out-of-sample ranking.

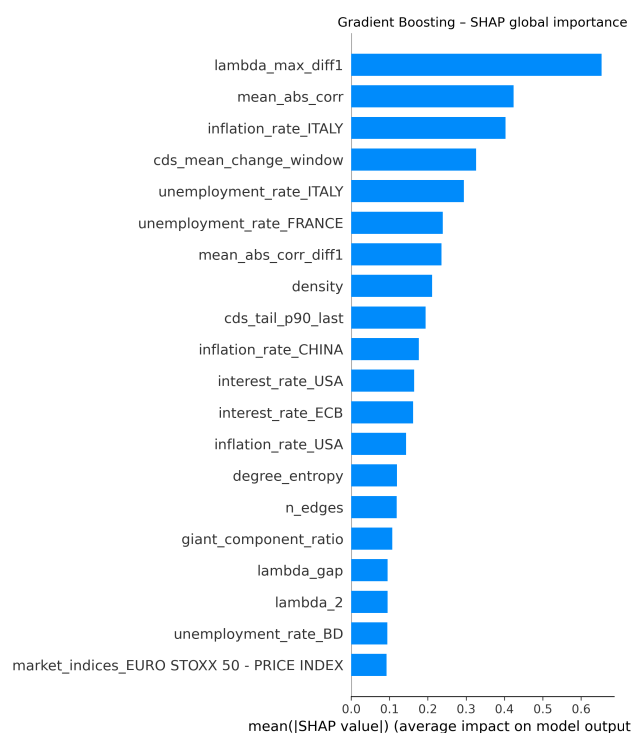


Figure 13. Global SHAP importance for gradient boosting.

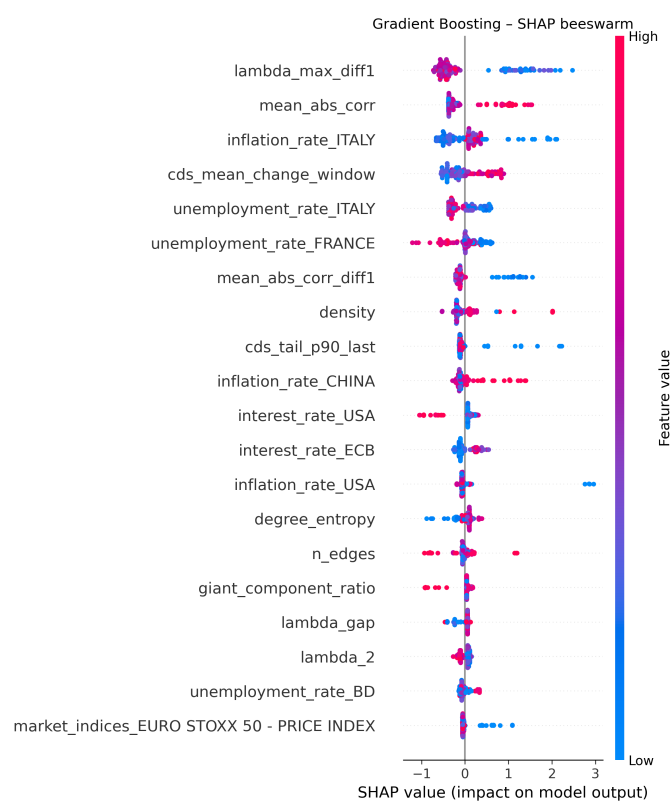


Figure 14. SHAP beeswarm plot for the most influential predictors.

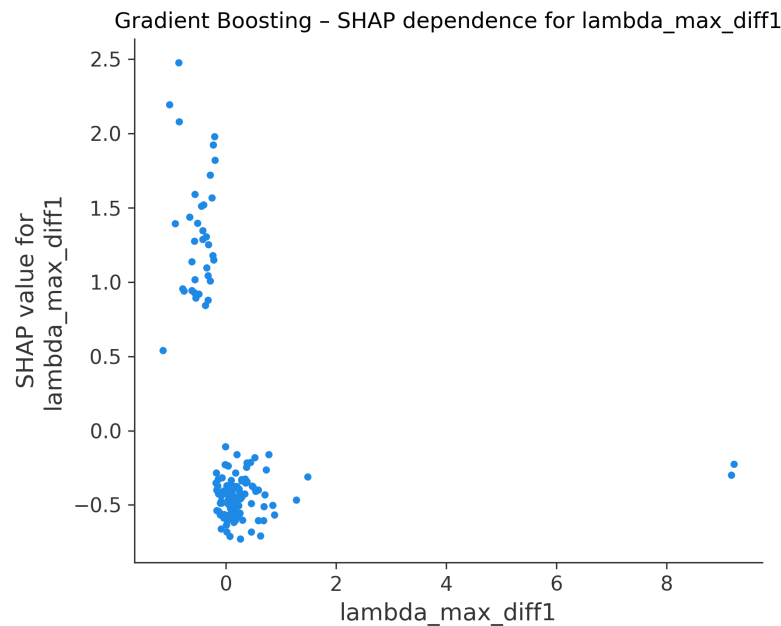


Figure 15. SHAP dependence plot for the most influential feature. The x-axis reports the feature value, while the y-axis shows the corresponding SHAP value, indicating the contribution of the feature to the model's predicted crisis probability relative to the baseline.

Table 3. Gradient boosting ablation study comparing feature specifications under expanding-window evaluation.

Feature Specification	Mean AUC	Std. AUC	CI _{low}	CI _{high}
Macro-financial only	0.526	0.066	0.481	0.591
Network only	0.526	0.066	0.481	0.591
Macro + Network (full)	0.526	0.066	0.481	0.591

4.6. Interpretation of the Ablation Result

Identical AUC values across the ablations do not imply that network measures are uninformative; they indicate predictive substitutability between two ways of encoding related signals. Network indicators such as mean absolute correlation, density, or eigenvalue-based instability summarise broad co-movement patterns that are also reflected, albeit indirectly, in credit spreads, volatility proxies, and aggregate market dynamics. When predictors are correlated, a model can shift weight between alternative representations with little change in ranking performance. This also helps reconcile the ablation result with the SHAP analysis. SHAP values describe attribution conditional on the feature set used by the fitted model; with correlated inputs, large attributions can be assigned to one representation even if another representation can replace it with little loss in AUC. In this setting, the practical value of network measures lies in how they organise information about synchronisation and interdependence into monitorable structural indicators that can be communicated alongside macro-financial variables.

While AUC is threshold-independent, early-warning systems are often judged by performance at specific operating points.

Table 4 reports Accuracy, Precision, Recall, and F1-score computed at a fixed threshold of $\tau = 0.5$, again within the expanding-window evaluation. All values are summarised as means and standard deviations across test folds. Recall is particularly relevant in a monitoring context, where missing a stress regime can be more costly than issuing false alarms. The results show that, despite moderate AUC values, the models retain non-trivial

sensitivity to crisis periods. This supports interpreting them as screening tools that can flag elevated risk, rather than as devices for precise point prediction.

Table 4. Out-of-sample classification performance across expanding-window folds (threshold = 0.5). Reported values are mean $\pm$ standard deviation across folds.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.59 $\pm$ 0.26	0.38 $\pm$ 0.39	0.56 $\pm$ 0.52	0.43 $\pm$ 0.41
Random Forest	0.59 $\pm$ 0.20	0.29 $\pm$ 0.39	0.44 $\pm$ 0.52	0.32 $\pm$ 0.39
Gradient Boosting	0.52 $\pm$ 0.09	0.28 $\pm$ 0.21	0.73 $\pm$ 0.49	0.40 $\pm$ 0.28

4.7. Label-Leakage Robustness

Because the crisis label is constructed from extreme returns and volatility, a natural concern is that predictors based on closely related quantities could inflate performance mechanically. To address this, an additional robustness exercise removes all features directly tied to short-horizon returns and volatility proxies before estimation. The remaining predictors consist of macro-financial indicators and network measures only. The full nested expanding-window procedure is then re-run without other changes. Table 5 reports the resulting out-of-sample AUC values and their variability.

Across all three model classes, performance is essentially unchanged relative to the baseline specification (differences across specifications are below 10^{-3} and therefore vanish at the reported precision). In particular, gradient boosting achieves a pooled out-of-sample AUC of approximately 0.62 in both the full-feature model and the restricted model. This stability suggests that the predictive content of the framework is not driven solely by variables that mechanically resemble the label definition but is instead linked to broader macro-financial conditions and to the structural properties of the correlation network.

Table 5. Leakage-robustness experiment: out-of-sample AUC for full-feature set and after removing return/volatility predictors.

Specification	Model	Mean AUC	Std.	95% CI
Full	Logistic Regression	0.589	0.358	[0.238, 0.940]
Full	Random Forest	0.610	0.218	[0.396, 0.824]
Full	Gradient Boosting	0.619	0.157	[0.465, 0.772]
No Return/Volatility	Logistic Regression	0.589	0.358	[0.238, 0.940]
No Return/Volatility	Random Forest	0.610	0.218	[0.396, 0.824]
No Return/Volatility	Gradient Boosting	0.619	0.157	[0.465, 0.772]

4.8. Computational Considerations

Although nested expanding-window tuning is more computationally intensive than a single static fit, the burden is limited here because the dataset contains 207 monthly observations and 51 predictors, and the hyperparameter grids are deliberately compact. More importantly, the design is chosen as a conservative evaluation protocol rather than as a requirement for deployment. In operational settings, hyperparameters could be fixed after an initial calibration period and models updated on a lower-frequency schedule (for example quarterly), preserving the expanding-window logic while reducing repeated search.

5. Conclusions

This paper asked a fairly practical question: whether combining market variables, macro-financial indicators, and correlation-network summaries can yield an early-warning tool for systemic stress that remains interpretable, and whether explainability methods can make clear how different information sources enter the model's risk assessments. The

empirical results suggest that this is workable under sequential, real-time-style constraints. At the same time, they also make plain that forecasting systemic stress is a noisy exercise, especially at monthly frequency, and that no modelling choice can remove that uncertainty.

On the predictive side, the nested expanding-window evaluation shows that the feature set contains information that helps separate stress from non-stress months, but only to a limited extent. Pooled out-of-sample AUC values lie at around 0.55–0.60, whereas fold-averaged AUC values are lower in some cases, most visibly for logistic regression, because early test windows are small and class composition changes over time. The effective sample is 207 months after accounting for rolling-window feature construction and the three-month forecasting shift. Performance at this level is consistent with the nature of the problem: systemic episodes tend to reflect layered interactions between markets and macroeconomic conditions, rather than clean signals that can be picked up reliably several months in advance. Differences between logistic regression, random forests, and gradient boosting are modest, which suggests that what predictive content exists is largely driven by the engineered predictors and the way information is summarised, not by the particular classifier.

From an implementation standpoint, the pipeline is designed so that it can be run in a monitoring environment without the full cost of repeated hyperparameter search. In practice, one can fix hyperparameters after an initial calibration period and then update the model on a regular schedule, preserving time-consistent estimation while keeping computation manageable.

Interpretability is where the framework adds more tangible value. The SHAP analysis provides a concrete account of which variables are used by the fitted model and in what direction they tend to move the predicted probabilities. In the gradient-boosting specification, volatility-sensitive indicators, short-horizon return dynamics, and network-instability measures repeatedly appear among the most influential inputs. Beeswarm and dependence plots show that higher stress proxies and stronger structural co-movement are typically associated with higher model-implied risk, often in a nonlinear way. These explanations describe the internal logic of the model conditional on the available feature set; they should not be read as causal claims about the economic origins of crises.

Overall, the results are consistent with the view that financial markets display recurring statistical regularities and shifts in dependence structure that can be exploited for monitoring. The proposed framework does not promise precise crisis prediction, and the evidence does not support that interpretation. Its contribution is instead that it combines market dynamics, macro-financial conditions, and evolving network summaries in a single workflow that can be evaluated sequentially and reproduced, while still allowing the analyst to inspect which signals matter for the model's output.

The main policy relevance lies in diagnosis and surveillance rather than in point forecasts. Correlation networks offer a compact way to track synchronisation and changes in market structure over time, and the interpretability layer helps link alerts back to observable indicators that can be discussed, audited, and communicated within supervisory processes. This makes it easier to justify why the system flags elevated risk, without resorting to opaque architectures that are difficult to validate.

Several limitations remain. The crisis label is necessarily rule-based, and even with careful shifting and robustness checks it may overlap with some predictor families; exploring alternative stress definitions and external crisis chronologies would therefore be valuable. The network layer could also be extended: higher-frequency data, different dependence estimators, or richer structures such as directed or multilayer networks may alter which instability signals are most informative. Finally, explicitly modelling cross-

border transmission channels is a natural next step, given the international scope of large institutions and the role of global linkages in modern systemic events.

Author Contributions: Conceptualisation, E.B. and L.V.M.; methodology, L.V.M.; software, L.V.M.; validation, A.G. and E.B.; data curation, L.V.M.; writing—original draft preparation, L.V.M.; writing—review and editing, E.B. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is available at <https://github.com/EBarbierato/Systemic-Crisis-Prediction/tree/main/data> (accessed on 24 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fromlet, H. Predictability of Financial Crises: Lessons from Sweden for Other Countries. *Bus. Econ.* **2012**, *47*, 262–272. [CrossRef]
2. Ahelegbey, D.F. The Econometrics of Bayesian Graphical Models: A Review with Financial Application. *J. Netw. Theory Financ.* **2016**, *2*, 1–33. [CrossRef]
3. Oxford University Press. *Oxford English Dictionary*; Oxford University Press: New York, NY, USA, 2025.
4. Bergomi, L. *Stochastic Volatility Modeling*; CRC Press: Boca Raton, FL, USA, 2015.
5. Gundel, S. Towards a new typology of crises. *J. Contingencies Crisis Manag.* **2005**, *13*, 106–115. [CrossRef]
6. Shaluf, I.M.; Ahmadun, F.-r.; Mat Said, A. A review of disaster and crisis. *Disaster Prev. Manag. Int. J.* **2003**, *12*, 24–32. [CrossRef]
7. Claessens, S.; Kose, M.A.; Laeven, L.; Valencia, F. *Understanding Financial Crises: Causes, Consequences, and Policy Responses*; Working Paper 05/2013; Centre for Applied Macroeconomic Analysis (CAMA), Australian National University: Canberra, Australia, 2013.
8. Claessens, S.; Kose, A. Financial Crises: Explanations, Types, and Implications. *IMF Work. Pap.* **2013**, *13*, 1–65. [CrossRef]
9. Peralta, G. Network-based Measures as Leading Indicators of Market Instability: The Case of the Spanish Stock Market. *SSRN Electron. J.* **2014**, *1*, 91–122.
10. Giannopoulos, G.; Sardar, S.A.; Salti, R.; Sykianakis, N. Analyzing Insolvency Prediction Models in the Period Before and After the Financial Crisis: A Case Study on the Example of US Firms. *Int. J. Financ. Insur. Risk Manag.* **2022**, *12*, 23–45. [CrossRef] [PubMed]
11. Fasano, A.; La Rocca, M. The Dilemma of Accuracy in Bankruptcy Prediction: A New Approach Using Explainable AI Techniques. *SSRN Electron. J.* **2023**, *28*, 1–22. [CrossRef]
12. Nyabera, A.M.; Ikamari, C.; Mocheche, G. Prediction of Financial Distress Using Dynamic Artificial Neural Network for Early Warning System. *Asian J. Probab. Stat.* **2024**, *26*, 151–159. [CrossRef]
13. Dastkhan, H.; Rad, H.S. Idiosyncratic Shocks and Financial Links: A Network Representation. Available at SSRN 4178143. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4178143 (accessed on 24 January 2026).
14. Ren, Y.; Zhu, Y.; Zhu, H.; Jiang, J. *Prediction of Financial Crisis Events Using Graph Neural Network Model: Based on Inter-Industry Spillover Information*; SSRN Elsevier: Rochester, NY, USA, 2024.
15. Chowdhury, B.; Dungey, M.; Kangogo, M.; Sayeed, M.A.; Volkov, V. The changing network of financial market linkages: The Asian experience. *Int. Rev. Financ. Anal.* **2019**, *64*, 71–92. [CrossRef]
16. Song, S.; Li, H. Prediction of Financial System Risk Based on Complex Network. 2024. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4330104 (accessed on 24 January 2026).
17. Guritanu, E.; Barbierato, E.; Gatti, A. Topological Machine Learning for Financial Crisis Detection: Early Warning Signals from Persistent Homology. *Computers* **2025**, *14*, 408. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.