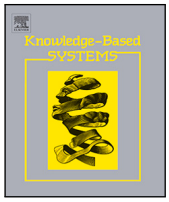
Knowledge-Based Systems 294 (2024) 111712

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Interfacing learning methods for anomaly detection in multi-country financial stress indicators

Xing Gu [a], Rogemar Mamon [a,c,*], Thibaut Duprey [b]

[a] Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada
[b] Financial Stability Department, Bank of Canada, Ottawa, Ontario, Canada
[c] Division of Physical Sciences and Mathematics, University of the Philippines Visayas, Miagao, Iloilo, Philippines

ARTICLE INFO

Keywords:
HMM filtering
Optimal parameter estimation
Prediction
Predictive analytics
Classification

ABSTRACT

This paper presents a novel ensemble supervised learning classification model designed for the early detection of financial stability anomalies. In particular, we utilise the time series of Financial Stress Indices (FSI) across multiple countries in developing an early-warning system. The innovation of this model lies in its unique integration of stochastic process modelling, hidden Markov models (HMM), random forest (RF), and XGBoost algorithms. This results to a comprehensive approach that can capture the dynamics of FSIs and forecast potential crisis episodes. The model's strength arises from the synthesis of the Ornstein–Uhlenbeck (OU) processes and HMM online recursive filters, forming a robust framework. Additionally, a feature selection module based on RF and a final classifier using XGBoost enhance the out-of-sample predictive performance. Our comparative analyses with five alternative models underscore the strong predictive power of the proposed model. A tailored feature-importance analysis highlights the substantial impact of the HMM features, emphasising their crucial role in the model's effectiveness. Furthermore, the inclusion of two projected anomaly-warning signals enhances the model's ability to predict extreme events, benefitting financial stability and public policy research.

1. Introduction

1.1. Problem definition and motivation

We shall address the central problem of detecting financial-stress episodes in advance with the view of providing bankers and regulators some tools in the mitigation of shocks attributed to the occurrence of financial stress. Since the 2007 financial crisis, there have been concerted efforts amongst researchers, market participants, and regulators to analyse and predict the potential for financial instability. As there are various interpretations of what financial stability is, we rely on the World Bank for its characterisations, namely, (i) the absence of system-wide episodes, i.e., financial crisis, in which the financial system fails to function; and (ii) the resilience of financial systems to stress. Financial stability is, therefore, a state where the financial system is immune to systemic financial crisis and able to smoothly perform its basic functions.

1.2. Related work

As alluded in [1], identifying financial instability and measuring its intensity amidst the complex interdependencies of internal and external forces on the financial system and economy are challenging tasks. The developments of a comprehensive quantitative index representing the extent of financial stress have been growing, with Duprey et al. [2] for example, introducing a country-level Financial Stress Index (FSI) covering equity markets, bond markets, and foreign exchange markets. This line of research enquiry successfully pinpointed systemic financial stress episodes using Markov-switching and threshold vector autoregressive models combined with elements quantifying stress on the business cycle.

Moreover, research on financial stability has explored machine learning algorithms for building early-warning models to predict financial crises . An early-warning model is constructed with Random

* Correspondence to: Department of Statistical and Actuarial Sciences, The University of Western Ontario, 1151 Richmond Street, London, Ontario, Canada, N6A 5B7.
E-mail address: rmamon@stats.uwo.ca (R. Mamon).

https://doi.org/10.1016/j.knosys.2024.111712
Received 5 December 2022; Received in revised form 17 March 2024; Accepted 26 March 2024
Available online 6 April 2024

Forests in [3] to predict systemic banking crisis to financial stability. Duttagupta and Cashin [4] employed a Binary Classification Tree (BCT) model to investigate banking crises, while Ward [5] demonstrated the superior out-of-sample performance of a Random Forest-based banking crisis indicator compared to logit models. Casabianca et al. [6] developed an early-warning system using Adaptive Boosting (AdaBoost), outperforming traditional logit models. In another instance, an early-warning system constructed with artificial neural networks (ANN) surpassed parametric models in predicting sovereign-debt-crisis episodes [7].

1.3. Innovation and contributions

Our innovative approach is much broader than the conventional methodologies in that it is able to accommodate the more realistic situations of partially available or limited data. In particular, we solely rely on the Financial Stress Index (FSI) and an aggregated index formed from various component indices, serving as the basis for model development. Our methodology is a fusion of algorithms culminating in a supervised-learning classification approach. It involves the processing of crucial information based on country-level FSIs reported in [2] to pinpoint instances of high financial stress.

Furthermore, our model not only captures the dynamic evolution of FSIs in parallel across multiple countries but also generates forward-looking predictions. We use discriminative signals, which are defined as the lagged differenced values of the original time series. To simultaneously model the signals from 17 countries through the Hidden Markov Model (HMM)-modulated Ornstein–Uhlenbeck (OU) processes, we employ the HMM recursive online filters to extract signal information using the change of probability measure technique. The integrated modelling algorithm, leveraging the Random Forest and XGBoost classifiers, demonstrates superior predictive-analytics performance.

A novelty of our model implementation via HMM filters is its adaptability for self-training, i.e., getting trained continuously and automatically with the arrival of new data. Consequently, this allows for the updating of model parameters reflecting the dynamic nature of input data. Employing a direct-prediction approach, our model makes six-step ahead predictions on the financial-stress status of all countries, thereby generating early-warning signals for the possibility of financial crises. Finally, we construct two anomaly-warning signal systems designed to identify distinct extreme anomalous episodes six months in advance.

Our findings affirm that our proposed modelling algorithm excels in detecting FSIs' aberrations in advance, exhibiting a higher capability with fewer false alarms. This model holds promise for economists and regulators seeking dynamic quantitative assessments of FSI status tailored to country-level financial stability management.

1.4. Main objectives

This article aims to construct an algorithmic platform integrating machine learning and statistical models that will enhance the predictive-analytics performance when detecting upcoming financial instability episodes. The focus is on capturing the evolution of multiple countries' FSIs and generating step-ahead predictions using the synthesis of Hidden Markov Models, Ornstein–Uhlenbeck processes, Random Forest, and XGBoost classifiers. We demonstrate how our methodology works in addressing one primary concern in economics and finance. Nonetheless, the generality of its principle could very well be applied to other multivariate modelling problems in the natural sciences and engineering.

1.5. Organisation of the paper

The remaining parts of this paper is structured as follows. Section 2 provides a brief overview of the underlying principles of the major components of our modelling algorithm. In Section 3, we formulate our methodology consisting of HMM filtering derivation, Random-Forests predictors' selection module, and XGBoost tuning and training. The details of a comparative examination involving the algorithm put forward in this paper vis-à-vis existing approaches are given in Section 4. The construction and diagnostics of two signal-projection systems are laid out in Section 5. Lastly, Section 6 provides a summary and implications of this study.

2. Overview of the model's major components

2.1. HMM-based regime-switching models

HMM is a doubly-embedded stochastic process comprised of an observation series and an underlying hidden process delineated by its number of states and transition probabilities or intensities. The parameters of an HMM can be estimated based on the observed data series unveiling the dynamics of the driving but unobserved Markov chain. Thus, the capacity to extract information in order to provide optimal model parameter estimates, by filtering out the noise from the data set, is the dominantly superior feature of HMM-based methods.

When dealing with real life-data, HMM could be employed as a statistical learning model to make classifications on time series according to its estimated parameters. See for instance, Li et al. [8]) who constructed an HMM-embedded device to detect anomalies in multivariate time series. In financial economics, where the manifestations of irregular or anomalous events are not apparently visible, an HMM-based model is an advantageous tool in the analysis of pertinent data. In Cao et al. [9], for example, an HMM with wavelet transformations and gradients was utilised to detect price manipulation activities in the stock markets.

Moreover, an HMM is the building block of a regime-switching-based technique in pinning down some underlying processes in finance or economics where they primarily evolve with the random shifts of their statistics (e.g., mean or variance) amongst different states. Thus, the HMM framework offers the flexibility that allows structural regimes, which are governed by the location, scale and shape parameters of a distribution, to shift stochastically over time. Such a framework is suited in capturing widely observed phenomena in financial economics, natural sciences, and engineering. The Markov-switching methodology in economics could be traced back to Hamilton's work [10], with the static estimation of the model addressed. The Markov-switching model designed mainly to describe structural changes in time series aims to (i) differentiate regimes of the economy and (ii) estimate the probability of being in an expansion or contraction.

In this paper, the HMM modulates a mean-reverting OU process to capture the dynamic behaviour of an FSI. The estimation of HMM is performed via filtering methods, which are quite common in electronics and electrical engineering. The recursive filtering, which gives rise to a self-calibrating model, is an innovation relative to the models created in the past that are heavily dependent on static-model fitting approach of maximum likelihood estimation [11]. A change of reference probability measure, which is a dominant feature in our HMM-filtering procedure and pioneered by Elliott et al. [12], facilitates the development of a self-calibrating model. An accessible introduction of this method can be found in Mamon et al. [11].

Recent research progress in this HMM filtering framework highlights various implementations in the areas of quantitative finance, insurance, economics, epidemiology, and other branches of the sciences and engineering. In finance specifically, these include short-rate modelling (e.g., Elliott et al. [13], Elliott and Mamon [14], Erlwein

and Mamon [15], Zhou and Mamon [16], and Xi and Mamon [17]); investment strategies (Tenyakov et al. [18] and Erlwein et al. [19]; commodity price forecasting (Date et al. [20]); weather derivative pricing (Xiong and Mamon [21] and Xiong and Mamon [22]); liquidity risk forecasting (Gu et al. [23]) modelling electricity spot prices (Erlwein et al. [24]).

The major motivation in this paper in the incorporation of the HMM-based model into our modelling process is to augment the capacities of learning the FSI's evolution. The turning points and pertinent evolution patterns could indicate possible financial-instability episodes. Therefore, numerical outcomes from our HMM filters could enable the detection of anomalies related to impending financial stress.

2.2. Random forests

In statistical learning, an ensemble method combines multiple predictors to obtain predictive modelling performance that is better than simply using a single-component modelling algorithm. Random forests (RFs) [25] are an ensemble machine-learning framework for classification and regression via the construction of multiple decision trees. The RF have become one of the most favoured supervised-learning models due to its considerable advantages, which are: (i) notable accuracy; (ii) robustness to outliers and noise; (iii) ease of use; and (iv) availability of internal estimates of error and variable importance [25]. In addition, the model has strong immunity from overfitting (see [25,26]), especially for classification problems [26].

The RFs are also aimed at enhancing the variance reduction of bagging (short for bootstrap aggregating) via building a large collection of de-correlated trees. This is attained by selecting variables randomly in the tree-growing procedure [26]. More specifically, all trees are grown to some bootstrap samples drawn from a training set. It can be shown that, when the sample size is large, approximately 63.2% of the instances in the training set are used to build one tree.[1] As a consequence, the remaining 36.8% of the data are used to evaluate the out-of-bag (OOB) error which is a good metric to measure the predictive power for each tree.

The OOB samples are also used to construct a measure for variable importance by evaluating the prediction power of each predictor. Once a tree is built, the OOB samples are processed by the tree, and the prediction accuracy is obtained. The algorithm randomly permutes the values for each single variable in the OOB samples to calculate the accuracy. The difference between the accuracy values before and after permutation is averaged over all trees, which is treated as a measure for the variables' importance in the RF.

Owing to its superiority as pointed out above, an RF is widely employed as a predictive model to solve various regression and classification problems in time series analysis. Tyralis and Papacharalampous [27], Karasu and Altan [28] and Hao et al. [29] found that the predictive error could be reduced by selecting features using an RF on different time-series data sets. In this research, we shall employ RFs to identify variables that contribute more to the prediction of our target variable. These selected features will be passed on to the next-stage classifier to make the final prediction.

2.3. Extreme gradient boosting (XGBoost)

Boosting is another powerful learning-ensemble algorithm that aims to convert a set of weak learners, which could only extract a small amount information, into a strong learner [30]. The weak learners are weighted in some way related to their accuracy. The weights are updated after a weak learner is added. Input samples with poor learning result lead to a higher weight, which is then emphasised more and reflected when adding the next weak learner. Eventually, the final model is established by weighting all the component learners based on their performance. The final result of the boosting model is obtained by averaging and counting votes in the solutions to regression and classification problems, respectively.

The fundamental idea of XGBoost [31] is to predict the sum of scores from multiple classification and regression trees (CARTs) considering that one tree is usually not capable of capturing sufficient information from the data. XGBoost uses a tree ensemble model that consists of a group of CARTs [32]. It is a scalable end-to-end boosting system, which is widely recognised by data scientists to achieve state-of-the-art results and outperforms many other competing models tackling various machine-learning challenges [31]. XGBoost is based on a gradient-tree boosting [33] and entails a numerical optimisation approach to minimise the loss function via the addition of trees in the gradient-descent algorithm. More specifically, new trees are built and trained on the errors of that previous model so that more importance are given to the observations in which existing learners were misclassified. A strong learner is updated by adding a new trained tree that generates a new prediction, whereby each tree's contribution is determined by minimising the overall error of the strong learner.

Following Chen and Guestrin [31], XGBoost could be described in the following way. For a given data set with n instances and m features, a tree-ensemble model uses K additive functions to predict the target variable y. That is,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{1}$$

where $f_k \in \mathcal{F}$ and $\mathcal{F} = \{f(x) = w_{q(x)} \mid q : R^m \to T, w \in R^T\}$ is a functional space that includes all possible CARTs. Note that $q(\cdot)$ is the structure of each tree that maps an instance to the corresponding leaf index. Also, T represents the number of leaves in the tree. Each f_k is related to an independent tree structure q and the leaf weight w.

Instead of growing multiple trees at once, the model is trained in an additive manner. To clarify this further, let $y_i^{(t)}$ be the prediction of the ith instance at the tth step, and f_t is added to minimise the objective function of the form given by (2). In other words, f_t is added in a greedy way that provides the optimal improvement to the model. In order to optimise the loss function efficiently, Chen and Guestrin [31] uses a Taylor approximation through the first and second-order gradients g_i and h_i in Eq. (3); in turn, more information concerning the gradients' directions are gained. Taking into account the above-mentioned descriptions, the loss function is given by

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2}$$

$$\approx \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) + \Omega(f_t), \tag{3}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$; $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$; and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$.

The function l in (2) is convex and continuously twice differentiable; it quantifies the difference between the true value y_i and the prediction $\hat{y}_i$. In order to avoid over-fitting, there is a regularisation term Ω in (2) that effectively penalises the complexity of the model. Thus, this regularisation incentivises in selecting a simple model with strong predictive power.

The major advantages and features of XGBoost are: (i) gradient-boosting-method implementation with the utility of the second-order partial derivatives, which provides more accurate approximations of the loss function; (ii) presence of regularisation that enhances model's predictive abilities; and (iii) capacity to incorporate parallel and distributed computing which makes learning faster than existing algorithms [31].

[1] Brief proof: Let n be the sample size of the training set. The probability of an instance not selected to build a tree is $1 - (1 - \frac{1}{n})^n \to 1 - e^{-1} \approx 0.632$ as n goes to infinity.

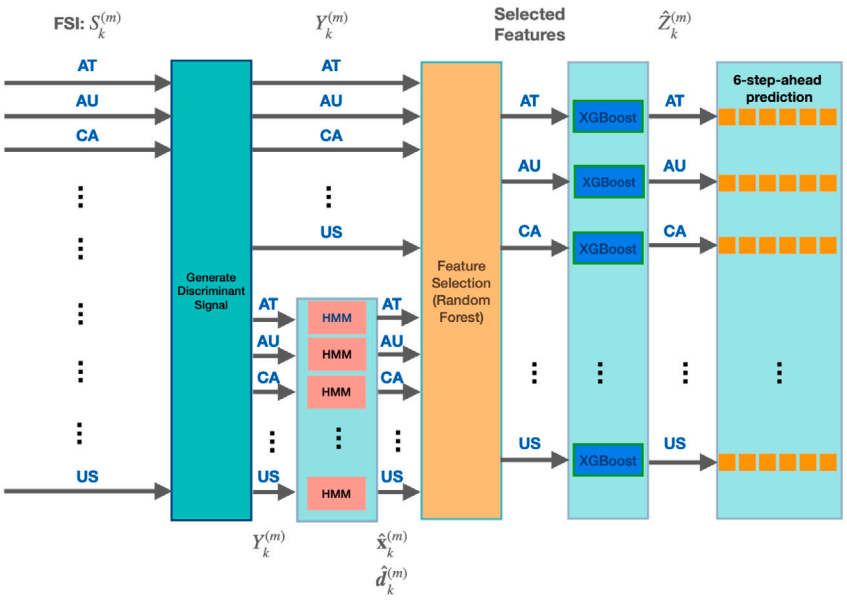


Fig. 1. Illustrating the architecture of the proposed modelling approach.

XGBoost became popular in tackling various real-life data-analysis problems. Zhang et al. [34] proposed an XGBoost-based classifier to detect indoor-human activities based on time series collected via smartphone sensors. In Basak et al. [35], it was found that models based on XGBoost and RFs are able to achieve high-accuracy predictions of the stock-price movements' direction in the medium to long term. Nobre and Neves [36] constructed an automated stock-trading system which combines Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT), XGBoost and a Multi-Objective Optimisation Genetic Algorithm (MOO-GA).

In this work, we shall employ XGBoost as a final classifier to detect anomalies related to that task of coming up with multiple-step-ahead prediction regarding the movements of all FSI time series for each country in the data set.

3. The integrated modelling algorithm construction

3.1. Model structure

This work presents an automated early-warning system designed to detect the anomaly in the FSI. This is accomplished by predicting the probability of occurrence of a rising edge in the next 6 months. A supervised learning framework consisting of HMM filters, RF and XGBoost binary classifier is proposed. Fig. 1 illustrates a schematic layout of our modelling approach, which is implemented in the R programming language. Supporting details on certain derivations and empirical aspects of this research are given in the **Supplementary Material** consisting of Appendices A (Procedure for filter's construction), B (Optimal parameter estimates) and C (Supplemental tables and figures).

3.2. Input layer

3.2.1. Data set

The data set in this work is taken from Duprey et al. [2] comprising the monthly FSIs of 17 countries for the period Jan 1968–Sep 2019 (i.e., a matrix of 621×17 data points). The descriptive statistics and the graphical evolution of the multi-dimensional time series are illustrated in Table C.1 in Appendix C and Fig. 2, respectively. Note that the range of FSI values is from 0 (no financial stress) to 1 (extremely high financial stress).

3.2.2. Discriminative signal

Let S_t denote the FSI process. We shall forecast S_t's movement which may signal financial stress in the future. More precisely, at a given time t, we need to determine whether or not the time series will have a rising edge between $t+1$ and $t+h$ inclusive. To carry this out, the original input undergoes the transformation

$$Y_t = \int_{t-\tau}^{t} \mathrm{d}S_u = S_t - S_{t-\tau}, \tag{4}$$

where τ is the length of the moving window. In Eq. (4), Y_t has two interpretations: (i) the difference in the S_t values between time t and time $t-\tau$; and (ii) the cumulative sum of S_t changes over the interval $[t-\tau, t]$.

Since FSI is a monthly data series, we set $\tau = 12$. Thus, Y_t measures the aggregate elevation of the index in the last 12 months; it could be viewed as the FSI's year-over-year net growth. Thus, Y_t's progression also traces dynamically the movement of S_t. The evolution of Y_t is exhibited (red curve) in Fig. 2.

3.2.3. Target variable formulation

Define the binary target variable Z_t as

$$Z_t = \begin{cases} 1\,(\text{Positive}), & \text{if } Y_t > 0, \\ 0\,(\text{Negative}), & \text{if } Y_t \le 0, \end{cases} \tag{5}$$

From Eq. (5), $Z_t = 1$ means that, at time t, FSI is growing relative to its value at time $t-\tau$ or, equivalently, it has positive net growth in the time interval $[t-\tau, t]$ which indicates that the rising possibility of FSI is high. The dynamics of Z_t are visualised as shaded areas (light purple) in Fig. 2. It is apparent that the major increasing edges of FSI are captured by the target variable. That is, the FSI is increasing or going to rise when $Z_t = 1$. The main purpose of this work is to propose a model to forecast the value of Z_t in the next h time periods, i.e., Z_{t+h}. Thus, for regulators, this quantity could serve as an indicator for both the direction of the FSI in the short run and a country's exposure to financial instability. The distributions of Z_t for each country are presented in Figure C.1 in Appendix C as a percent-stacked bar chart. It is worth noting that all the 17 data series are nearly balanced. That is each of them contains equal or almost equal number of samples from the positive and negative class corresponding to the value of the target variable Z_t.

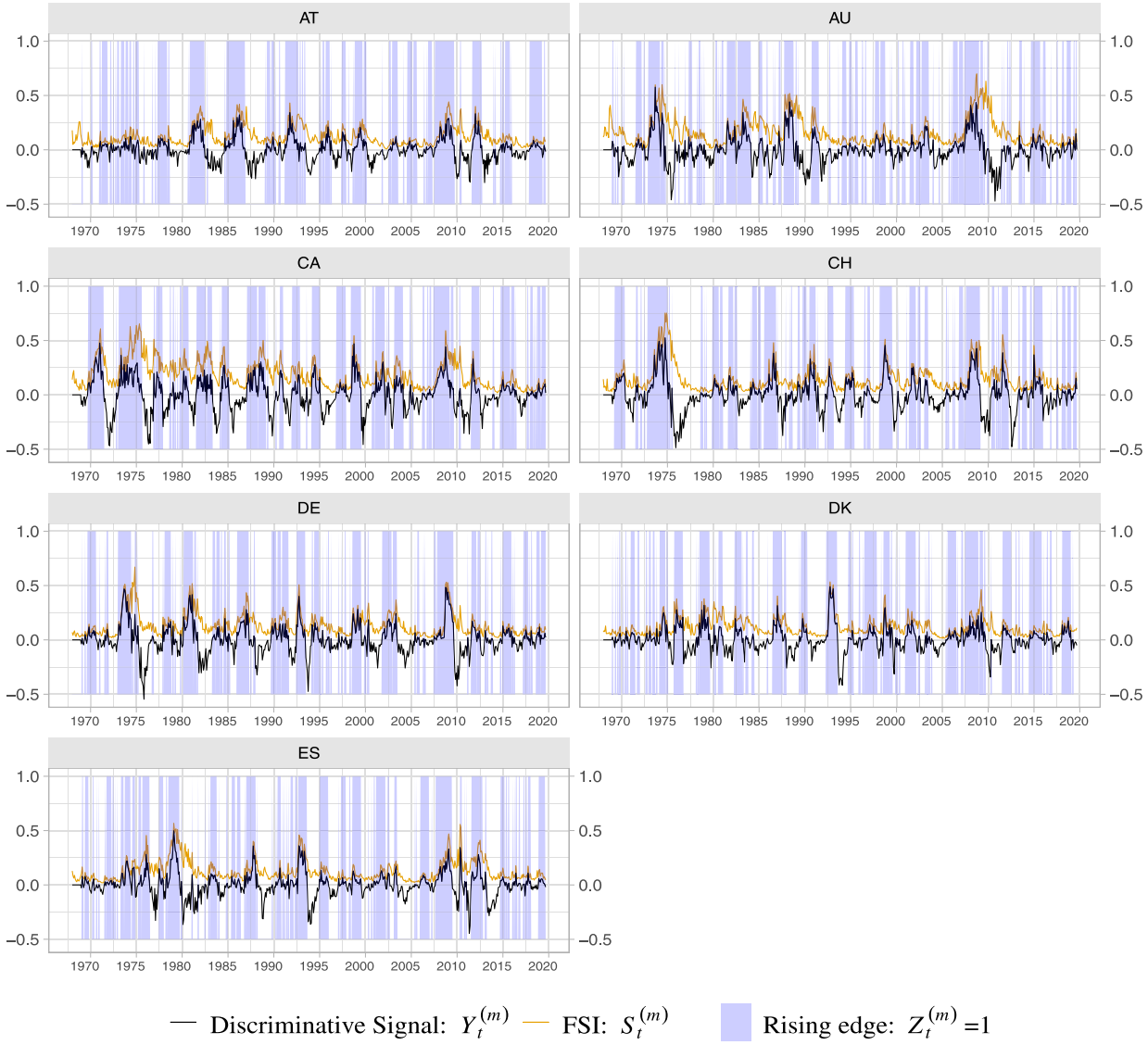


Fig. 2. Part I - Evolution of the FSIs and discriminative signals for AT (Australia), AU (Austria), CA (Canada), CH (Switzerland), DE (Germany), DK (Denmark) and ES (Spain). Part II - Evolution of the FSIs and discriminative signals for FI (Finland), FR (France), GB (United Kingdom), IT (Italy), JP (Japan), KR (Korea), NL (Netherlands), NZ (New Zealand), SE (Sweden) and US (United States).

3.3. HMM filtering module

3.3.1. Ornstein–Uhlenbeck (OU) process

Fig. 2, showing the discriminative signals Y_t derived from all the country level FSIs, portrays a mean-reverting behaviour, dynamically and randomly moving from high to low level and vice-versa. This fact provides support for an OU process as a means to capture the signals' mean-reverting attribute.

Suppose the discriminative signal Y_t in Eq. (4) Y_t follows an OU process in accordance with the stochastic differential equation

$$\mathrm{d}Y_t = \theta\left(\mu - Y_t\right)\mathrm{d}t + \sigma\ \mathrm{d}B_t, \tag{6}$$

where μ represents the mean-reverting level, θ is the speed of mean reversion, and σ is the volatility. We assume that the parameters θ, μ and σ are positive. The standard Brownian motion B_t is defined on some probability space $\left(\Omega, \mathscr{F}^{Y_t}, P\right)$, where $\mathscr{F}^{Y_t}$ is the filtration generated by Y_t.

By Itô's lemma, it may be verified that the solution to (6) is

$$Y_t = Y_0 e^{-\theta t} + (1 - e^{-\theta t})\mu + \sigma e^{-\theta t}\int_0^t e^{\theta s}\mathrm{d}B_s. \tag{7}$$

To implement Eq. (7) empirically, we require its discretisation given by

$$Y_{k+1} = Y_k e^{-\theta\cdot\Delta t} + (1 - e^{-\theta\cdot\Delta t})\mu + \sqrt{\frac{\sigma^2}{2\theta}(1 - e^{-2\theta\cdot\Delta t})}\ B_{k+1}, \tag{8}$$

where $\Delta t = (k+1) - k = 1$ and $\{B_{k+1}\}$ is a sequence of independent and identically distributed standard normal random variables. The third term in (8) is justified by the Gaussian property and the Itô's isometry.

3.3.2. HMM-governed OU process

To equip our modelling approach with the capability for time-varying or stochastic parameters, a hidden Markov-chain model (HMM) modulates the OU process in Eq. (6). In this work, we treat the states of the HMM as the country's financial-stress regime, i.e., normal or anomalous, which is an actualisation of interacting factors that cause financial instability.

To make the FSI regime-switching, the parameters θ, μ and σ are governed by a discrete-time Markov chain $\mathbf{x}_k$, for $k = 0, 1, \ldots$. The state space of $\mathbf{x}_k$ is finite and it is isomorphic to the canonical basis of $\mathbb{R}^N$, which is the set $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N\}$. The vector $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$

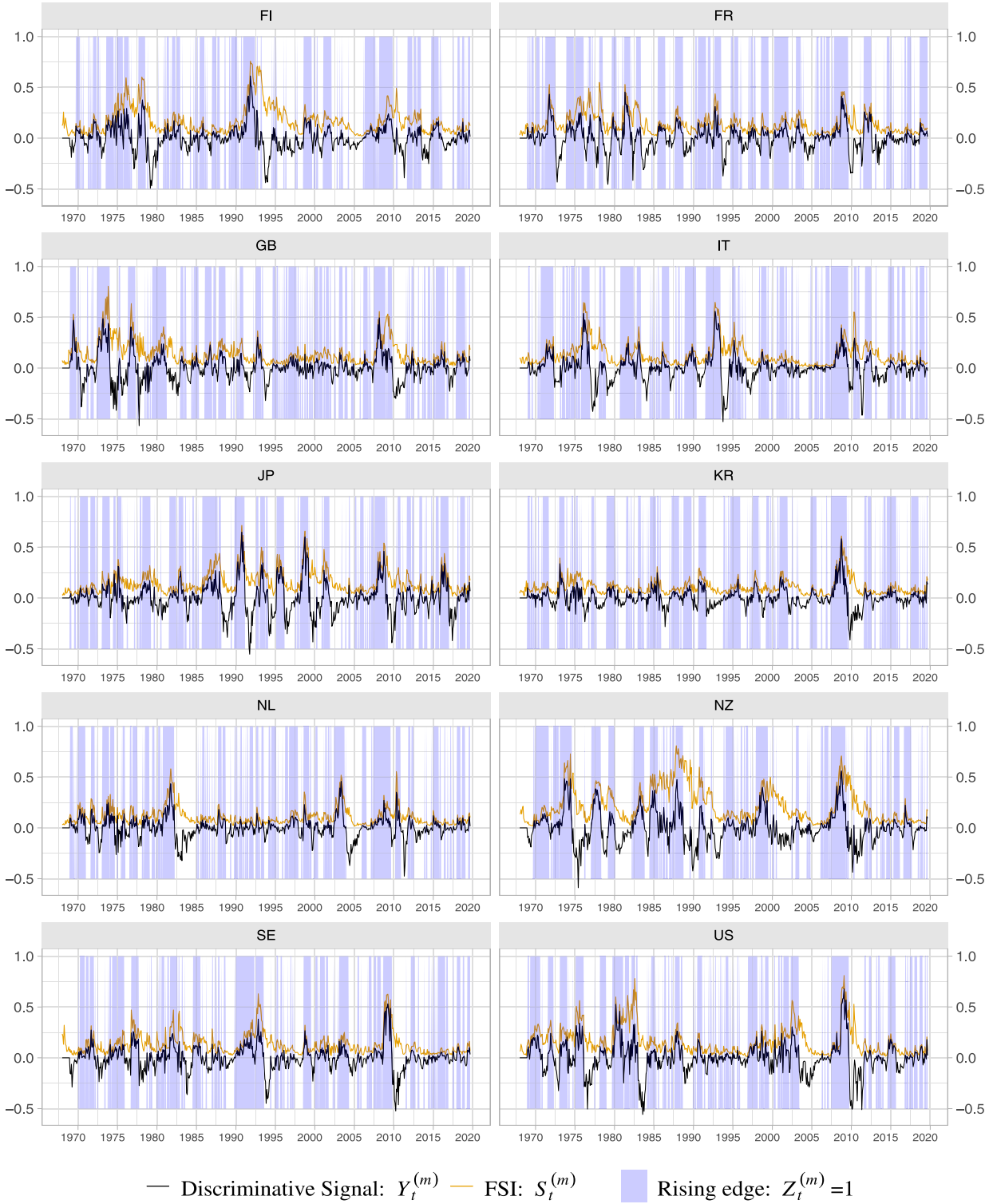


Fig. 2. (*continued*).

is a unit vector with 1 in its i^{th} component. The semi-martingale representation of $\mathbf{x}_k$ is

$$\mathbf{x}_{k+1} = \boldsymbol{\Pi}_k \mathbf{x}_k + \mathbf{v}_{k+1}, \tag{9}$$

where $\boldsymbol{\Pi}_k$ is a transition matrix; $\mathbf{v}_{k+1}$ is a martingale increment with $E[\mathbf{v}_{k+1}|\mathscr{F}^{\mathbf{x}_k}] = \mathbf{0}$; and $\mathscr{F}^{\mathbf{x}_k}$ is the filtration generated by $\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_k$. The dependence of the model parameters on the HMM should be clear from the notations $\theta(\mathbf{x}_k)$, $\mu(\mathbf{x}_k)$ and $\sigma(\mathbf{x}_k)$.

With M as the dimension of the multivariate time series, we have $M = 17$ for the data series with summary statistics in Table C.1 in Appendix C. For the m^{th} component of the M-dimensional discriminative signal, i.e., $Y_t^{(m)}$, its parameters are driven by some HMM $\mathbf{x}_k^{(m)}$ that evolves as per equation (9) so that

$$Y_{k+1}^{(m)} = Y_k^{(m)} e^{-\theta(\mathbf{x}_k^{(m)})\Delta t} + (1 - e^{-\theta(\mathbf{x}_k^{(m)})\Delta t})\boldsymbol{\mu}(\mathbf{x}_k^{(m)}) + \sqrt{\boldsymbol{\sigma}^2(\mathbf{x}_k^{(m)}) \cdot \frac{1 - e^{-2\theta(\mathbf{x}_k^{(m)})\Delta t}}{2\theta(\mathbf{x}_k^{(m)})}}\; b_{k+1}^{(m)}. \tag{10}$$

In Eq. (10), $\forall m = 1, 2, \ldots, M$,

$$\begin{cases} \mu(\mathbf{x}_k^{(m)}) = \langle \boldsymbol{\mu}_k^{(m)}, \mathbf{x}_k^{(m)} \rangle \\ \theta(\mathbf{x}_k^{(m)}) = \langle \boldsymbol{\theta}_k^{(m)}, \mathbf{x}_k^{(m)} \rangle \\ \sigma^2(\mathbf{x}_k^{(m)}) = \langle \boldsymbol{\sigma}_k^{2(m)}, \mathbf{x}_k^{(m)} \rangle, \end{cases} \tag{11}$$

with

$$\begin{cases} \boldsymbol{\mu}_k^{(m)} = (\mu_{k,1}^{(m)}, \mu_{k,2}^{(m)}, \ldots, \mu_{k,N}^{(m)})^\top \\ \boldsymbol{\theta}_k^{(m)} = (\theta_{k,1}^{(m)}, \theta_{k,2}^{(m)}, \ldots, \theta_{k,N}^{(m)})^\top \\ \boldsymbol{\sigma}_k^{2(m)} = (\sigma_{k,1}^{2(m)}, \sigma_{k,2}^{2(m)}, \ldots, \sigma_{k,N}^{2(m)})^\top \end{cases} \tag{12}$$

where $\boldsymbol{\mu}_k^{(m)}$, $\boldsymbol{\theta}_k^{(m)}$ and $\boldsymbol{\sigma}_k^{2(m)}$ are all in $\mathbb{R}^{\mathbb{N}}$; and $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^N$.

3.3.3. Parameter estimation

Eq. (10) is re-parameterised as

$$Y_{k+1}^{(m)} = \boldsymbol{\alpha}(\mathbf{x}_k^{(m)}) Y_k^{(m)} + \boldsymbol{\beta}(\mathbf{x}_k^{(m)}) + \boldsymbol{\zeta}(\mathbf{x}_k^{(m)})\; b_{k+1}^{(m)}, \tag{13}$$

where

$$\begin{cases} \boldsymbol{\alpha}(\mathbf{x}_k^{(m)}) = e^{-\theta(\mathbf{x}_k^{(m)})\Delta t} \\ \boldsymbol{\beta}(\mathbf{x}_k^{(m)}) = (1 - e^{-\theta(\mathbf{x}_k^{(m)})\Delta t})\boldsymbol{\mu}(\mathbf{x}_k^{(m)}) \\ \boldsymbol{\zeta}^2(\mathbf{x}_k^{(m)}) = \boldsymbol{\sigma}^2(\mathbf{x}_k^{(m)}) \cdot \dfrac{1 - e^{-2\theta(\mathbf{x}_k^{(m)})\cdot\Delta t}}{2\theta(\mathbf{x}_k^{(m)})}. \end{cases} \tag{14}$$

Eq. (13) implies that, given $Y_k^{(m)}$,

$$Y_{k+1}^{(m)} \sim N\left(\boldsymbol{\alpha}(\mathbf{x}_k^{(m)}) Y_k^{(m)} + \boldsymbol{\beta}(\mathbf{x}_k^{(m)}),\; \boldsymbol{\zeta}^2(\mathbf{x}_k^{(m)})\right). \tag{15}$$

Let $\hat{\mathbf{x}}_k^{(m)}$ be the conditional expectation of the state of hidden Markov chain $\mathbf{x}_k^{(m)}$ given $\mathscr{F}_k^{(m)}$ under probability measure $P^{(m)}$. Thus,

$$\hat{\mathbf{x}}_k^{(m)} := E\left[\mathbf{x}_k^{(m)} \middle| \mathscr{F}_k^{(m)}\right] = \left(\hat{x}_{k,1}^{(m)}, \hat{x}_{k,2}^{(m)}, \ldots, \hat{x}_{k,N}^{(m)}\right)^\top, \tag{16}$$

where

$$\hat{x}_{k,i}^{(m)} = P\left(\mathbf{x}_k^{(m)} = \mathbf{e}_i \middle| \mathscr{F}_k^{(m)}\right) = E\left[\langle \mathbf{x}_k^{(m)}, \mathbf{e}_i \rangle \middle| \mathscr{F}_k^{(m)}\right]. \tag{17}$$

We shall utilise the EM algorithm to find the estimate of the parameters of the HMM-driven model, which is a component of our hybridised modelling algorithm. See Elliott and Krishnamurthy [37] as well as Wu [38] for the details of the EM algorithm.

We further define the following quantities for states $j = 1, 2 \ldots, N$:

$$\begin{cases} \mathcal{G}_k^{s,j,(m)} = \sum_{n=1}^{k} \langle \mathbf{x}_{n-1}^{(m)}, \mathbf{e}_j \rangle \langle \mathbf{x}_n^{(m)}, \mathbf{e}_s \rangle \\ \mathcal{O}_k^{j,(m)} = \sum_{n=1}^{k} \langle \mathbf{x}_{n-1}^{(m)}, \mathbf{e}_j \rangle \\ \mathcal{T}_k^{j,(m)}(f) = \sum_{n=1}^{k} \langle \mathbf{x}_{n-1}^{(m)}, \mathbf{e}_j \rangle f(\cdot). \end{cases} \tag{18}$$

The quantities $\mathcal{G}_k^{s,j,(m)}$ and $\mathcal{O}_k^{j,(m)}$ represent the respective number of jumps from state j to s and the amount of time that the Markov chain $\{\mathbf{x}_n^{(m)}\}$ occupies the state j up to time k, respectively. Also, $\mathcal{T}_k^{j,(m)}$ is an auxiliary process that depends on the function $f(\cdot)$. In our empirical application, $f(\cdot)$ takes the form $Y_n^{(m)}$, $(Y_n^{(m)})^2$ and $Y_n^{(m)} Y_{n-1}^{(m)}$. Following the procedure introduced in [11,39], we obtain the optimal parameter estimates below.

$$\hat{\mathbf{x}}_k^{(m)} = \frac{\mathbf{u}_k^{(m)}}{\sum_{i=1}^{N} \langle \mathbf{u}_k^{(m)}, \mathbf{e}_i \rangle}. \tag{19}$$

$$\hat{\pi}^{j,i,(m)} = \frac{\hat{\mathcal{G}}_k^{j,i,(m)}}{\hat{\mathcal{O}}_k^{i,(m)}} \tag{20}$$

$$\hat{\alpha}_{k,i}^{(m)} = \frac{\hat{\mathcal{T}}_k^{i,(m)}\left(Y_k^{(m)} Y_{k-1}^{(m)}\right) - \hat{\beta}_{k-1,i}^{(m)} \cdot \hat{\mathcal{T}}_k^{i,(m)}\left(Y_{k-1}^{(m)}\right)}{\hat{\mathcal{T}}_k^{i,(m)}\left(Y_{k-1}^{(m)^2}\right)} \tag{21}$$

$$\hat{\beta}_{k,i}^{(m)} = \frac{\hat{\mathcal{T}}_k^{i,(m)}\left(Y_k^{(m)}\right) - \hat{\alpha}_{k-1,i}^{(m)} \cdot \hat{\mathcal{T}}_k^{i,(m)}\left(Y_{k-1}^{(m)}\right)}{\hat{\mathcal{O}}_k^{i,(m)}} \tag{22}$$

$$\begin{aligned} \widehat{\zeta_{k,i}^2}^{(m)} = {} & \frac{\hat{\mathcal{T}}_k^{i,(m)}\left(Y_k^{(m)^2}\right) + \hat{\alpha}_{k-1,i}^{(m)} \cdot \hat{\mathcal{T}}_k^{i,(m)}\left(Y_{k-1}^{(m)^2}\right)}{\hat{\mathcal{O}}_k^{i,(m)}} \\ & + \frac{\left(\hat{\beta}_{k-1,i}^{(m)}\right)^2 \hat{\mathcal{O}}_k^{i,(m)} + 2\alpha_{k-1,i}^{(m)} \hat{\beta}_{k-1,i}^{(m)} \hat{\mathcal{T}}_k^{i,(m)}\left(Y_{k-1}^{(m)}\right)}{\hat{\mathcal{O}}_k^{i,(m)}} \\ & - \frac{2\hat{\alpha}_{k-1,i}^{(m)} \hat{\mathcal{T}}_k^{i,(m)}\left(Y_k^{(m)} Y_{k-1}^{(m)}\right)}{\hat{\mathcal{O}}_k^{i,(m)}} \\ & - \frac{2\hat{\beta}_{k-1,i}^{(m)} \hat{\mathcal{T}}_k^{i,(m)}\left(Y_k^{(m)}\right)}{\hat{\mathcal{O}}_k^{i,(m)}} \end{aligned} \tag{23}$$

The details of how we develop the HMM filtering and estimation are presented in Appendices A and B. For the original model parameters specified in Eq. (10), we use Eqs. (20)–(23) to get the following optimal values:

$$\begin{cases} \hat{\mu}_{k,i}^{(m)} = \dfrac{\hat{\beta}_{k,i}^{(m)}}{1 - \hat{\alpha}_{k,i}^{(m)}} \\ \hat{\theta}_{k,i}^{(m)} = \dfrac{\ln(\hat{\alpha}_{k,i}^{(m)})^{-1}}{\Delta t} \\ \widehat{\sigma_{k,i}^2}^{(m)} = 2\hat{\theta}_{k,i}^{(m)} \cdot \widehat{\zeta_{k,i}^2}^{(m)} \left(1 - e^{-2\hat{\theta}_{k,i}^{(m)} \cdot \Delta t}\right). \end{cases} \tag{24}$$

3.3.4. Feature construction: Markov-chain state estimate

Tenyakov et al. [40] and Gu et al. [41] find the state of the hidden Markov chain filtered from the given data series carries essential information that can be treated as an early warning signal of financial instability events. In this paper, our underlying process is $Y_t^{(m)}$ which dynamically trace the trend of corresponding FSI. It is intuitive to consider that the HMM state estimate $\mathbf{x}_k^{(m)}$, which drives the OU process

$Y_t^{(m)}$, is able to provide salient guidance to forecast the movement direction of FSI.

Eq. (17) defines the estimate of hidden Markov chain $\mathbf{x}_k^{(m)}$ state at time k. The calculation formula is given in Equation A.6 in Appendix A. According to the set up in Eq. (16), we can write the ith component of $\hat{\mathbf{x}}_k^{(m)}$ as follows:

$$\hat{x}_{k,i}^{(m)} = \langle \hat{\mathbf{x}}_k^{(m)},\ e_i \rangle, \quad \forall i \in \{1, 2, \dots, N\} \tag{25}$$

Findings obtained by previous researches such as Erlwein et al. [24], Tenyakov et al. [40] and Xiong and Mamon [21] indicate that a 2-state HMM set up is sufficient to express the evolution of the underlying data series. Thus we set numbers of state $N = 2$ for all Markov chains in this work. We shall include the Markov-chain state estimates which related to anomalous state, i.e., $\hat{x}_{k,1}^{(m)}$, in the input set for next stage modelling. The evolutions of $\hat{x}_{k,1}^{(m)}$ are illustrated in Figure C.2 in Appendix C.

3.3.5. Feature construction: Deviation from the reverting mean

The evolution of $Y_t^{(m)}$ is captured by the outcomes of the HMM filters, which are used to predict its future dynamics. From Eq. (10),

$$\begin{aligned} \hat{Y}_{k+1}^{(m)} - \boldsymbol{\mu}\left(\mathbf{x}_k^{(m)}\right) &= E\left[Y_{k+1}^{(m)} \middle| \mathscr{F}_k^{(m)}\right] - \boldsymbol{\mu}\left(\mathbf{x}_k^{(m)}\right) \\ &= \left(Y_k^{(m)} - \boldsymbol{\mu}(\mathbf{x}_k^{(m)})\right) \cdot e^{-\theta(\mathbf{x}_k^{(m)})\Delta t}. \end{aligned} \tag{26}$$

Eq. (26) implies that the difference between $Y_k^{(m)}$ and its mean could be predicted using the HMM filtered parameters.

Write

$$d\left(\mathbf{x}_k^{(m)}\right) := \left(Y_{t_k}^{(m)} - \boldsymbol{\mu}\left(\mathbf{x}_k^{(m)}\right)\right) \cdot e^{-\theta(\mathbf{x}_k^{(m)})\Delta t}. \tag{27}$$

With the formulation of notation in Eq. (11),

$$d\left(\mathbf{x}_k^{(m)}\right) = \langle \boldsymbol{d}_k^{(m)},\ \mathbf{x}_k^{(m)} \rangle, \tag{28}$$

where

$$\boldsymbol{d}_k^{(m)} = (d_{k,1}^{(m)},\ d_{k,2}^{(m)},\ \dots,\ d_{k,N}^{(m)})^\top \tag{29}$$

and

$$d_{k,i}^{(m)} = \left(Y_k^{(m)} - \mu_{k,i}^{(m)}\right) \cdot e^{-\theta_{k,i}^{(m)}\Delta t}, \quad \forall i \in \{1, 2, \dots, N\}. \tag{30}$$

At time k, we obtain

$$\hat{d}_{k,i}^{(m)} = E\left[d_{k,i}^{(m)} \middle| \mathscr{F}_k^{(m)}\right] = \left(Y_k^{(m)} - \hat{\mu}_{k,i}^{(m)}\right) \cdot e^{-\hat{\theta}_{k,i}^{(m)}\Delta t}, \ \forall i \in \{1, \dots, N\}. \tag{31}$$

In addition to the discriminative signal, we incorporate $\hat{d}_{k,i}^{(m)}$ along with the Markov-chain state estimate $\hat{x}_{k,1}^{(m)}$ as the initial inputs in our proposed modelling approach. The evolutions of $\hat{d}_{k,i}^{(m)}$ are presented in Figure C.3 in Appendix C.

3.4. Random-forest-based feature-selection module

3.4.1. Model set up

The modelling-component set ups in Subsections 3.2 and 3.3 signify that the input features to the RF module can be divided into three categories: (i) discriminative signals, (ii) deviation from the reverting mean derived from the HMM filters, and (iii) Markov-chain state estimates. Each category constitutes a set of features of different countries with a range of time lags.

Our hybridised model is defined as follows:

$$Z_{k+h}^{(m)} = G_h^{(m)}\left(\mathbf{Y}_k, \mathbf{D}_k, \mathbf{X}_k | \mathscr{F}_k\right) \tag{32}$$

with

$$\begin{cases} \mathbf{Y}_k = \left\{ Y_{k-q}^{(m)} \middle| m = 1, 2, \dots, 17;\ q = 0, 1, \dots, (T-1). \right\} \\ \mathbf{D}_k = \left\{ \hat{d}_{k-q,i}^{(m)} \middle| i = 1, 2;\ m = 1, 2, \dots, 17;\ q = 0, 1, \dots, (T-1). \right\} \\ \mathbf{X}_k = \left\{ \hat{x}_{k-q,1}^{(m)} \middle| m = 1, 2, \dots, 17;\ q = 0, 1, \dots, (T-1). \right\}. \end{cases} \tag{33}$$

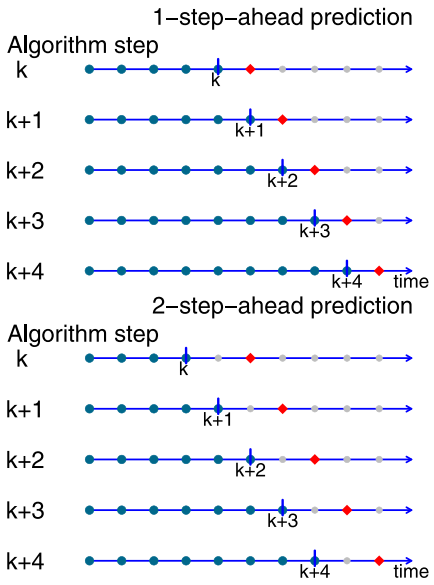


Fig. 3. The modelling scheme for training as well as 1- and 2-step-ahead predictions.

In Eq. (32), $Z_{k+h}^{(m)}$, as previously defined in equation (5), is the binary target variable corresponding to $Y_k^{(m)}$. The function $G_h^{(m)}(\cdot)$ is unknown and does not have a closed form although it could be estimated. Furthermore, q represents the memory length of the model with boundary T. Considering the monthly frequency of the data series, we set $T = 12$; that is, the proposed model has a memory length of 12 months. More specifically, for each country in the data set, we let the corresponding target variable $Z_k^{(m)}$ depends on the those three groups of predictors of itself as well as others with the maximum lag of 12 months. In this work, we use the first n ($n = 90$) rows of sample points to initialise our model.

The model for $G_h^{(m)}(\cdot)$ is essential in making multi-step predictions on the binary target variable Z_t. A recursive-forecasting strategy will be employed in our prediction. For a time series x_t, the procedure in making the h-step-ahead prediction at time k is as follows: (i) Train the model on $\{x_1, x_2, \dots, x_k\}$; (ii) Predict x_{k+1} using the previously trained model; (iii) Predict x_{k+2} with the same trained model using a data set including the previously predicted x_{k+1}. The procedure continues until x_{k+h} is predicted.

Under a recursive prediction framework, the model is only trained once based on a 1-step-ahead optimisation criterion, which could make the model fail in capturing the temporal dynamics [42]. We note that incorporating the estimated value (i.e., prediction for last time point) in the inputs inevitably introduces cumulative errors in the long-horizon forecasting [43]. This recursive prediction's limitation is further confirmed in [44] in which propagation errors are sustained when tested on real-world data set.

In this paper, we apply a direct prediction method for multi-step forecasting. Fig. 3 diagrams the training and test schemes for 1-step-ahead and 2-step-ahead predictions. At each time point, we build a specific predictive model to generate a single output for an h−step-ahead prediction. More precisely, at time k, for all m data series and all h-step-ahead ($h = 1, 2, \dots, 6$) forecasting tasks, a particular model $G_h^{(m)}(\cdot)$ is calibrated on a training set (marked by dark cyan dots in Fig. 3) with a sample size of $(n + k - h)$ to generate a prediction $\hat{Z}_{k+h}^{(m)}$ (marked by large red diamonds in Fig. 3). Such a prediction is a probabilistic value classifying the current data point as either positive (i.e., anomalous) or not.

3.4.2. Selection procedure

Tyralis and Papacharalampous [27] employed an RF to efficiently select the lag feature of a time series. In our case, we also use an RF

Table 1
Parameter setting for the XGBoost model.

Parameters	Value	Description
booster	gbtree	Booster to use
objective	binary:logistic	Objective function
eval_metric	logloss	Cross validation metric
nrounds	100	The number of boosting iteration
eta	{0.02, 0.06, 0.08, 0.1}	Learning rate
max_depth	{4, 5, 8}	Maximum depth of a tree
gamma	{0, 0.1, 0.3, 0.5}	Minimum loss reduction for a new node
alpha	{0.01, 0.02, 0.03, 0.04}	L1 regularisation term
others	Default	–

in the selection of lag variables as well as country predictors. More specifically, at time k, for every $Z_{k+h}^{(m)}$ with $m = 1, 2, \ldots, 17$, $h = 1, 2, \ldots, 6$, we first fit the model defined in Eq. (32) to the data with all the variables included in Eq. (33). After each round, the features' importance are calculated. We select the predictors with top-20 highest scores. The mean decrease of accuracy (i.e., OOB) is chosen as the measure of feature importance. The R function randomForest is utilised to implement the selection process.

We assign the number of trees to grow (ntree) to a value of 500. The number of variables selected as candidates at each split (mtry) is set to $p/2$, where p is the total number of predictors. To minimise the computation time of the feature selection, all the other parameters are set as the default value of the function. The predictors selected by RF are then passed to the XGBoost classifier to produce a final prediction.

3.5. Extreme gradient boosting (XGBoost) classification module

The purpose of the XGBoost classification module is the generation of predictions $\hat{Z}_{k+h}^{(m)}$ at time k for $h = 1, 2, \ldots, 6$. The input of the module includes the past target-variable series $Z_k^{(m)}$ and the features selected by the RF module. The output $\hat{Z}_{k+h}^{(m)}$, lying between 0 and 1, is the estimated probability that $Z_{k+h}^{(m)} = 1$, which could be viewed as the country's FSI having a rising edge at time $k + h$ as well as the country's estimated level of exposure to financial instability. The XGBoost is implemented with the R functions "xgb.train" and "xgb.cv" from "xgboost" package.

3.5.1. XGBoost model set up

Before utilising the XGBoost classifier, it is necessary to set the parameters that will direct the model's learning process. In Table 1, the model parameters' values or specified subsets of parameter space are presented; these information are based on empirical experience concerning XGBoost implementation in practice. Those parameters with potential values will be tuned via cross validation.

3.5.2. Hyper-parameter tuning

Five parameters in Table 1 are hyper-parameters: nrounds, eta, max_depth, gamma and alpha. Since these parameters could not be estimated directly from the data, the optimal values are searched in such a way that the pre-defined loss function is minimised, which consequently yield good out-of-sample prediction results.

A gridsearch is a typical method in machine learning for tuning hyper-parameters. The basic procedure consists of the following steps: (i) Generate a list that incorporates all the combinations of hyper-parameters that are candidates for tuning; (ii) Define J training and validation partitions of the original data; (iii) For each $j = 1, 2, \ldots, J$, fit the models with all hyper-parameters setting defined in (i) on the training set; (iv) Make predictions on the validation set and calculate all the error metrics for all models; (v) Select the hyper-parameters for which the model gives the best performance, i.e., having the smallest log loss.

Note that the error metric used in the grid search algorithm is typically measured by cross validation. Bergmeir and Benítez [45] found that applying the traditional cross-validation methods on time series data will lead to both theoretical and practical problems , and as a consequence, fail to provide appropriate guidance to select the optimal parameters. Hyndman and Athanasopoulos [46] introduced a cross-validation technique tailored to time series and this is known as "evaluation on a rolling forecasting origin". In the literature, we implement this cross-validation method to make the training and validation partitions in step (ii) of the grid-search procedure. The process is illustrated in Fig. 4 with the examples of 1-step-ahead and 2-step-ahead prediction cases.

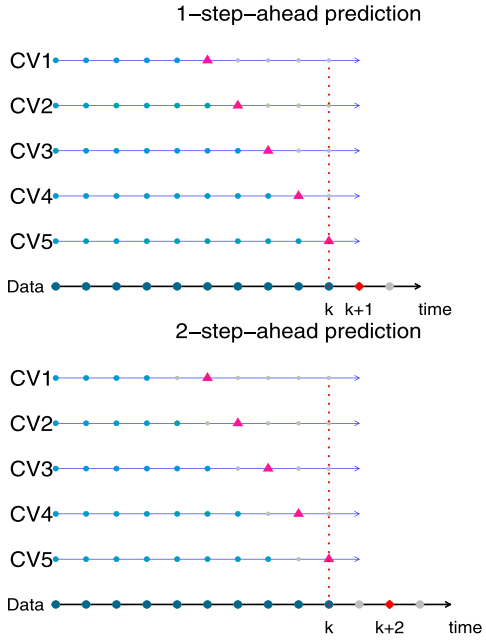


Fig. 4. A demonstration of cross-validation set ups: 1-step-ahead and 2-step-ahead predictions.

More specifically, our model makes an h-step-ahead prediction $\hat{Z}_{k+h}^{(m)}$ all data available up to time k as inputs. The dark cyan dots along the horizontal data axis in Fig. 4 represent all the data available at time k whilst the large red diamonds stand for the corresponding prediction data point. To search for the optimal hyper-parameters we need to create J training and validation partitions of the data set known at time k. We set $J = 5$ considering the balance of search efficiency and effectiveness. As exhibited in Fig. 4, each one of the short black axes CV1 to CV5 shows a training-validation partition in the cross validation. For each $j = 1, 2, \ldots, J$, (i) the data points with time index from 1 to $k - J + j - h$ are in the training set (small light black dots on the short axes) for cross validation; (ii) those with time index from $k - J + j$ to k are in the validation set (small red triangles along the black axes). Fig. 4 demonstrates the cross-validation set ups exemplifying the 1-step-ahead and 2-step-ahead predictions. The procedure could be generalised to higher values of h.

Table 2
Hardware and software specification as well as other details pertinent to numerical implementation.

Hardware/Software	Specifications/Version
Apple MacBook Pro	Retina, 15-inch, Mid 2015
Processor	2.2 GHz Quad-Core Intel Core i7
Memory	16 GB 1600 MHz DDR3
R	3.5.2
Runtime	17 h 31 mins

3.5.3. Training and prediction

The XGBoost classifier with optimal parameter are trained on all data sets with selected features. More specifically, at time k, for each $h = 1, 2, \ldots, 6$, we fit the model with the best parameter obtained using the grid-search process with cross validation on the available data set with the corresponding features selected by the aid of the RF module.

In order to fully capture the dynamics of the multivariate data series, we set the length of each algorithm step to 1 time unit (month). So, for each algorithm step k, we continuously tune the XGBoost classifier's hyper-parameters and then train the model using all available data up to time k. The main principles behind our modelling algorithmic approach is summarised in Algorithm 1.

Algorithm 1: General Algorithm for step-h prediction

Input: FSI of 17 countries
initialization;
for $k = n$ to $621 - h$ **do**
  for $m = 1$ to 17 **do**
    1. Calculate discriminative signal $Y_k^{(m)}$ and evaluate $Z_k^{(m)}$;
    2. Calculate HMM filters parameters and evaluate $\hat{x}_{k,1}^{(m)}$, $\hat{d}_{k,1}^{(m)}$, and $\hat{d}_{k,2}^{(m)}$;
    3. Generate HMM features: $\{\mathbf{D}_{k-h}, \mathbf{X}_{k-h}\}$;
    4. Random Forest: Select features from $\{\mathbf{Y}_{k-h}, \mathbf{D}_{k-h}, \mathbf{X}_{k-h}\}$;
    5. XGBoost: Tune hyperparameters via grid search cross validation on features selected in Step 4;
    6. XGBoost: Train tuned model and generate a prediciton;
    Output: $\hat{Z}_{k+h}^{(m)} = \hat{P}\left(Z_{k+h}^{(m)} = 1\right)$.
  end
end

4. Results and diagnostics

The hardware and software information used in our empirical study are reported in Table 2.

Our proposed modelling approach is tested on 17 countries' FSI data series covering the period Jan 1968–Sep 2019. We benchmark our hybridised model with 5 different alternative models. As presented in Table 3, Model_1 is our proposed model. Model_2 uses RF instead of XGBoost as the final-stage classifier. Thus, the purpose of Model_2 is to test the significance of XGBoost's contribution on system's performance. Model_3 and Model_4 are parts of Model_1 and Model_2, respectively. The motivation of introducing these two benchmark models is to verify the importance of HMM features. Model_5 is Vector Autoregressive Model with a memory length of 12. The first four models are embedded with classifiers at their final-stage component. Therefore, their outcomes are $\hat{Z}_{k+h}^{(m)}$ which could be interpreted as the model's implied probabilities of $Z_{k+h}^{(m)} = 1$.

The Naive model makes classifications based on the h-step-ahead predictions using the value at time k whilst Model_5 generates the predictions in a recursive way. The original outcomes of Model_5 and Naive method are $\hat{x}_{k+h}^{(m)}$ and the h-step-ahead predicted values of $Y_k^{(m)}$. The binary forecasting values for target variable $Z_{k+h}^{(m)}$ are obtained based on the definition of $Z_t^{(m)}$ in equation (5).

Since this research is actually solving a supervised learning problem, the model's predicted class is defined through

$$\tilde{Z}_{k+h}^{(m)} = \begin{cases} 1\,(\text{Predicted as Positive}), & \text{if } \hat{Z}_{k+h}^{(m)} \geq p_0, \\ 0\,(\text{Predicted as Negative}), & \text{if } \hat{Z}_{k+h}^{(m)} < p_0, \end{cases} \tag{34}$$

where p_0 is a discrimination threshold. Eq. (34) defines a decision rule concerning a data point at time $k + h$, whether to classify it as an anomalous point or not based on $\hat{Z}_{k+h}^{(m)}$ given all information at time k. In this work, we set $p_0 = 0.5$ for all classification models in Table 3 for simplicity and consistency of interpretation. After training the models on each individual time series $Y_t^{(m)}$, the models' forecasting power is evaluated vis-à-vis their stepwise out-of-sample prediction performance under 3 different metrics, which are discussed in the succeeding subsections.

4.1. Performance metrics derived from confusion matrix

The classification accuracy is a straightforward measure, which calculates the percentage of correct classification of a supervised learning model. The accuracy's evaluation formula is

$$\mathrm{ACC}_h^{(m)} = \frac{\sum_k \left(\mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=1,\ Z_{k+h}^{(m)}=1\right\}} + \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=0,\ Z_{k+h}^{(m)}=0\right\}} \right)}{\sum_k \left(\mathbf{1}_{\left\{Z_{k+h}^{(m)}=1\right\}} + \mathbf{1}_{\left\{Z_{k+h}^{(m)}=0\right\}} \right)}, \tag{35}$$

where $\mathbf{1}$ is an indicator function, and for $h = 1, 2, \ldots, 6$.

Although accuracy gives a direct assessment of the model, it does not tell the full story of classification performance. The metric will provide misleading outcomes when the data set is unbalanced (e.g., the distribution of observations across all categories is biased). A model's classification power in all known classes cannot definitely be assessed by the accuracy metric.

Our target is making binary classifications on multiple countries' FSI. We introduce a confusion matrix to gauge further the out-of-sample classification performance of our proposed model. More specifically, we trace the classification outcomes of all models' stepwise forecasts and compare them with the true values in the data set. For each step-ahead prediction, we aim to determine the numbers of True Positives, True Negatives, False Positives and False Negatives. For each step-ahead forecasting, the True Positive Rate (TPR) or Sensitivity, True Negative Rate (TNR) or Specificity, Positive Prediction Value (PPV) or Precision, and Negative Prediction Value (NPV).

$$\mathrm{TPR}_h = \frac{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=1,\ Z_{k+h}^{(m)}=1\right\}}}{\sum_k \mathbf{1}_{\left\{Z_{k+h}^{(m)}=1\right\}}}, \quad \mathrm{TNR}_h = \frac{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=0,\ Z_{k+h}^{(m)}=0\right\}}}{\sum_k \mathbf{1}_{\left\{Z_{k+h}^{(m)}=0\right\}}} \tag{36}$$

$$\mathrm{PPV}_h = \frac{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=1,\ Z_{k+h}^{(m)}=1\right\}}}{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=1\right\}}}, \quad \mathrm{NPV}_h = \frac{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=0,\ Z_{k+h}^{(m)}=0\right\}}}{\sum_k \mathbf{1}_{\left\{\tilde{Z}_{k+h}^{(m)}=0\right\}}}$$

The metrics in the set of Eqs. (36) are equal to 1 for perfect classification models. The numerical results of such metrics are displayed in Fig. 5. It can be observed that even though each model's forecasting power declines as the prediction horizon increases, Model_1 achieves high scores in all five metrics when applied on all countries' FSI data series.

This result indicates that the proposed model has better out-of-sample classification capability than other benchmark models. More specifically, attaining high accuracy score indicates that Model_1, in general, has a lower chance than other benchmark model to misclassify the data points. Furthermore, getting higher TPR and TNR scores implies that Model_1 has a better capacity in differentiating anomalous/normal episodes with relative lower false positive/negative errors. In addition, obtaining higher PPV and NPV scores suggests that Model_1 has a higher credibility on the correctness of its predictions. Last but not least, the above advantages of Model_1 is robust with respect to different time series within our data set.

Table 3
List of models and the benchmark included in our comparative analysis.

Model	Stage 1	Stage 2	Stage 3
Model_1	HMM feature generation	RF feature selection	XGBoost classifier
Model_2	HMM feature generation	RF feature selection	RF classifier
Model_3	RF feature selection	XGBoost classifier	–
Model_4	RF feature selection	RF classifier	–
Model_5	VAR(12)	–	–
Naive	$\hat{x}_{k+h} = Y_k,\ \forall h = 1,\ 2,\ \ldots$	–	–

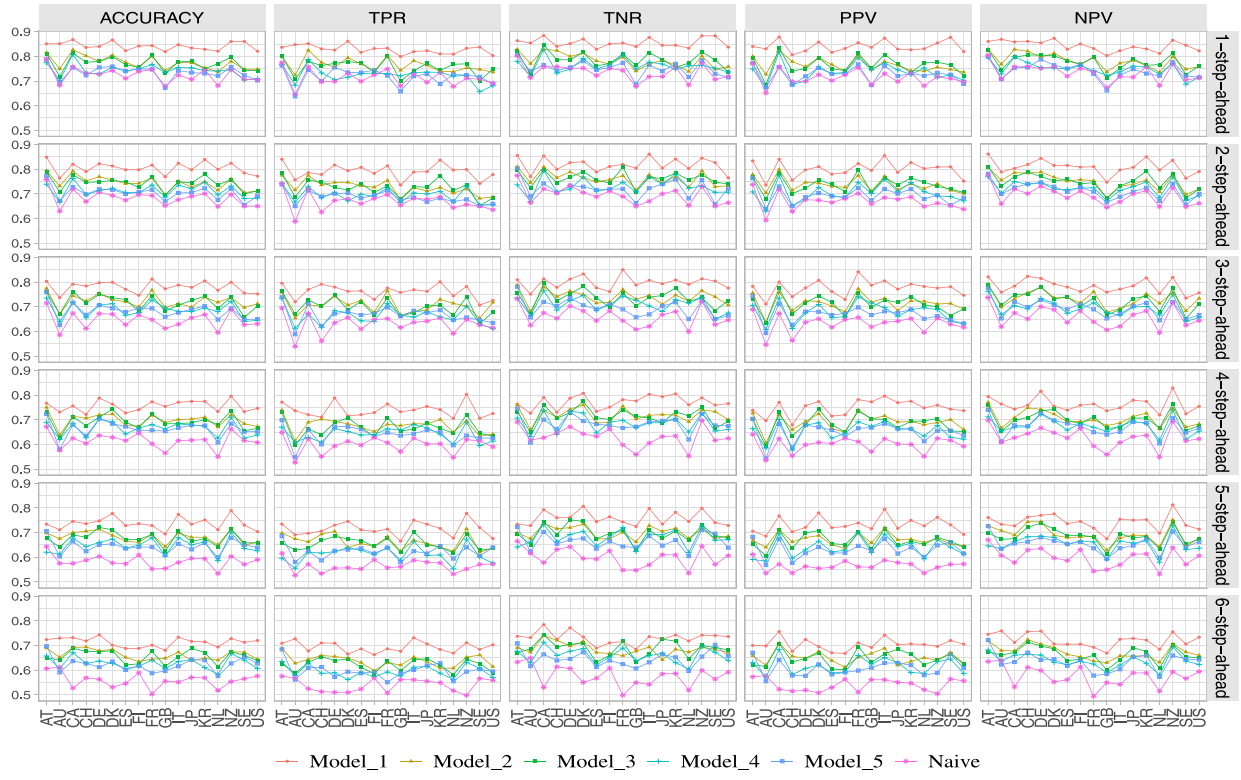


Fig. 5. Comparison of models in terms of their stepwise forecasting classification performance for various h's.

4.2. Area under the ROC curve (AUC)

The metrics derived from the confusion matrix in Section 4.1 assesses the model performance based on a fixed discrimination threshold $p_0 = 0.5$. Additional insights could be gained amongst different models with the investigation of each model's classification performance at different values of p_0.

A receiver operating characteristic (ROC) curve is a graphical tool showing the performance of a binary classification model at all discrimination thresholds. In general, the AUC considers the predicted probabilities as the criteria to judge the model's performance. The procedure to construct an ROC can be summarised as follows: (i) Order models' predicated probabilities $\hat{Z}^{(m)}_{k+h}$. (ii) Calculate $\tilde{Z}^{(m)}_{k+h}$ at multiple p_0's and obtain the corresponding TPR's and FPR's. (iii) Plot the TPRs against the FPRs.

An important summary metric is the area under the curve (AUC) representing the probability that a binary classification model ranks a randomly chosen positive example higher than a random negative instance. Note that AUC's value lies in the interval $[0, 1]$. The AUC of a worst model that makes 100% wrong predictions is 0 whilst that of a model that makes 100% right forecasting is 1. Fig. 6 gives a visualisation of the AUC values for the 4 different classification models under several step-ahead predictions. Model_1 achieves the highest AUC score in all 6-step-ahead predictions when applied on each single countries' FSI time series. Note that Model_5 and the Naive method in Table 3 are not included in the AUC comparison because the probabilistic outcomes are not available for these two models.

In order to ascertain the statistical significance of the results in Fig. 6, we conduct a pairwise one-tailed t-test facilitated by the bootstrap method. A sample size of 10,000 is chosen from each model's predicted probabilities on all countries' FSI data. A significance test of Model_1 versus Model_j ($j = 2, 3, 4$) on AUC at h-step-ahead predictions is formulated below.

$$\begin{cases} H_0 : AUC^{(m)}_{1,h} \leq AUC^{(m)}_{j,h} \\ H_a : AUC^{(m)}_{1,h} > AUC^{(m)}_{j,h}, \end{cases} \tag{37}$$

where m is the country index. The outcomes of the tests for each paired models are presented in Tables C.2 - C.4 in Appendix C. In each pairwise test, for every single FSI time series, the p-values are less than 5% on all prediction horizons. This implies that H_0's are rejected and our proposed model has more robust out-of-sample classification power than the other 3 models.

In addition to the t-test,a DeLong test [47] is performed to validate our model's significant superiority with regard to AUC. The null and alternative hypotheses are defined in a similar manner to (37). Using the R function "roc.test" from package "pROC", the outcomes are obtained and displayed in Tables C.5 - C.7 in Appendix C. All p-values are less than 5%, and so the null hypothesis is rejected. This justifies, in terms of the AUC score, that our proposed model dominates the other benchmark models.

Fig. 6. Comparison of models in terms of their stepwise forecasting classification performance under the AUC metric.

4.3. Log loss

As pointed out in Section 4.2, the AUC determines the model's capability to differentiate two classes by measuring the probability that a positive instance is ranked higher than a negative one. In this subsection, we shall compare the models' classification power based on a logistic loss, which is also known as log-loss or cross-entropy loss. This entails the evaluation of the negative log-likelihood for the predicted probabilities generated by the classification model.

In this paper, the logistic loss is calculated

$$L_h^{(m)} = -\frac{1}{N_h}\sum_k \left\{ Z_{k+h}^{(m)} \log\left(\hat{Z}_{k+h}^{(m)}\right) + \left(1 - Z_{k+h}^{(m)}\right) \log\left(\hat{Z}_{k+h}^{(m)}\right) \right\}. \tag{38}$$

From Eq. (38), the log loss could be computed using the actual class indicators and the logarithm of the predicted probabilities for each class. A classifier achieves a high log loss score if it yields high predicted probabilities for true positive example and low values for true negative instances. Theoretically, a perfect classification model's log-loss score is 0 whilst a bad one has a log-loss score is some relatively large positive value up to infinity.

In Table C.8 in Appendix C, we report the logistic loss for the proposed model and 3 other benchmark classification models. It could be seen that Model_1 has the lowest score in all 6 prediction horizons. Hence, it has higher predicted probabilities than the other models.

4.4. Kolmogorov–Smirnov test

As highlighted in Section 4.3, we assess the classification models' good of fit by analysing the logistic loss, which measures the log-likelihood of the predicted probabilities. However, a logistic loss does not capture the difference in the models' outcome for each class. Hence, we need to determine the divergence, in terms of the distribution, of a model's responses for different classes. More specifically, in our case, we shall apply the two-sample Kolmogorov–Smirnov (KS) test on the predicted probabilities' distributions covering the anomalous (True Positive) and normal (True Negative) episodes. This enables the comparison of the models' capacity to separate the actual anomalous intervals of FSI from the normal ones. The KS statistic is

$$D_h^{(m)} = \sup\left| F_{0,h}^{(m)} - F_{1,h}^{(m)} \right|, \tag{39}$$

where $F_{0,h}^{(m)}$ and $F_{1,h}^{(m)}$ are the empirical cumulative distribution functions (ECDF) of $\hat{Z}_{k+h}^{(m)}$ for the true positive and negative classes, respectively. The statistic quantifies the supremum of the set of the distances between the points on ECDFs of the two classes.

Note that a perfect model that generates two mutually exclusive categories with each one having a separate class label of observations has a KS statistic score of 100%. On the other hand, the worst model that fails to differentiate two groups has a score of 0%. In general, a high KS-statistic value s suggests that the model is good in distinguishing the two classes.

Table C.9 in Appendix C presents the KS statistics calculated from the predicted probabilities, which were produced by the proposed and benchmark models. It is noticeable that Model_1 attains the highest scores in all forecast horizons when implemented on all FSI time series. The higher the divergence between the distributions of predicted probabilities, the stronger the capacity to identify anomalous periods from regular episodes.

4.5. Analysis on features' importance

In addition to building a system with the capability to yield accurate predictions, it is also imperative to examine the model's interpretability. More specifically, we need to understand which features or input variables are most important for our system to detect the anomalous episodes in advance. Thus, in this subsection, we shall diagnose the proposed model by investigating the features' importance.

As discussed in Section 3.4, we apply the RF to select the top-20 features on the basis of their importance which are being measured by the mean decrease of the OOB accuracy. The selected features are

then used to train the XGBoost classifier for the generation of the out-of-sample (h-step-ahead) classifications. Our analysis mainly focuses on two types of quantities related to the features' importance: (i) frequencies of the features selected by the RF module, and (ii) gain score[2]) calculated by the XGBoost classifier.

At each algorithm step t, we trace the frequencies of all the features that are selected by the RF module and their importance scores calculated by the XGBoost classifier when making every h-step-ahead prediction ($h = 1, 2, \ldots, 6$). Furthermore, we calculate the marginal total frequencies of the features selected by the RF module and their marginal averaging importance scores output via the XGBoost module with respect to m (country index) and q (time lag). These quantities serve as comprehensive measurements of the contributions coming from different types of features in detecting anomalous episodes.

Let Ξ_t be the set of features selected by the RF module when performing an h-step-ahead prediction for one of the data series $Z_k^{(m)}$ at algorithm step t. We define the marginal total frequencies of these features, i.e., $\overline{D}_i(m)$, $\widetilde{D}_i(q)$ $(i = 1, 2)$, $\overline{X}(m)$, $\widetilde{X}(q)$, $\overline{Y}(m)$ and $\widetilde{Y}(q)$ as follows:

$$\begin{cases} \overline{D}_i(m) := \sum_t \sum_{q=0}^{11} \mathbf{1}_{\{\hat{d}^{(m)}_{k-q,i} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}, i = 1, 2 \\ \overline{X}(m) := \sum_t \sum_{q=0}^{11} \mathbf{1}_{\{\hat{x}^{(m)}_{k-q,1} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}, \\ \overline{Y}(m) := \sum_t \sum_{q=0}^{11} \mathbf{1}_{\{Y^{(m)}_{k-q} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}. \end{cases} \tag{40}$$

$$\begin{cases} \widetilde{D}_i(q) := \sum_t \sum_{m=1}^{17} \mathbf{1}_{\{\hat{d}^{(m)}_{k-q,i} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}, i = 1, 2, \\ \widetilde{X}(q) := \sum_t \sum_{m=1}^{17} \mathbf{1}_{\{\hat{x}^{(m)}_{k-q,1} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}, \\ \widetilde{Y}(q) := \sum_t \sum_{m=1}^{17} \mathbf{1}_{\{Y^{(m)}_{k-q} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}. \end{cases} \tag{41}$$

where $\mathbf{1}_{\{\cdot\}}$ is an indicator function. The quantities in the sets of Eqs. (40) and (41) act as the preliminary measures for evaluating the association between the target variable (i.e., Anomalous/Normal) and the different types of input features. In general, a higher marginal total frequency indicates that the related feature is treated as an important variable, with more recurrence, by the RF module in identifying the anomalies.

Let $\mathcal{I}_t(\cdot)$ be the importance score of some feature calculated by the XGBoost classifier at each algorithm step. The marginal averaging XGBoost importance scores $\overline{d}_i(m)$, $\widetilde{d}_i(q)$ $(i = 1, 2)$, $\overline{x}(m)$, $\widetilde{x}(q)$, $\overline{y}(m)$ and $\widetilde{y}(q)$ are calculated as follows:

$$\begin{cases} \overline{d}_i(m) := \dfrac{1}{\overline{D}_i(m)} \sum_t \sum_{q=0}^{11} \mathcal{I}_t\left(\hat{d}^{(m)}_{k-q,i}\right) \mathbf{1}_{\{\hat{d}^{(m)}_{k-q,i} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}, i = 1, 2 \\ \overline{x}(m) := \dfrac{1}{\overline{X}(m)} \sum_t \sum_{q=0}^{11} \mathcal{I}_t\left(\hat{x}^{(m)}_{k-q,1}\right) \mathbf{1}_{\{\hat{x}^{(m)}_{k-q,1} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}, \\ \overline{x}(m) := \dfrac{1}{\overline{Y}(m)} \sum_t \sum_{q=0}^{11} \mathcal{I}_t\left(Y^{(m)}_{k-q}\right) \mathbf{1}_{\{Y^{(m)}_{k-q} \in \Xi_t\}}, \ m \in \{1, 2, \ldots, 17\}. \end{cases} \tag{42}$$

$$\begin{cases} \widetilde{d}_i(q) := \dfrac{1}{\widetilde{D}_i(q)} \sum_t \sum_{m=1}^{17} \mathcal{I}_t\left(\hat{d}^{(m)}_{k-q,i}\right) \mathbf{1}_{\{\hat{d}^{(m)}_{k-q,i} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}, i = 1, 2, \\ \widetilde{x}(q) := \dfrac{1}{\widetilde{X}(q)} \sum_t \sum_{m=1}^{17} \mathcal{I}_t\left(\hat{x}^{(m)}_{k-q,1}\right) \mathbf{1}_{\{\hat{x}^{(m)}_{k-q,1} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}, \\ \widetilde{y}(q) := \dfrac{1}{\widetilde{Y}(q)} \sum_t \sum_{m=1}^{17} \mathcal{I}_t\left(Y^{(m)}_{k-q}\right) \mathbf{1}_{\{Y^{(m)}_{k-q} \in \Xi_t\}}, \ q \in \{0, 1, 2, \ldots, 11\}. \end{cases} \tag{43}$$

Note that the marginal averaging importance scores in the sets of Eqs. (42) and (43) serve as metrics for the dependence of the target variable on different features. In general, a higher value of the marginal averaging importance score implies that the corresponding feature is more important for the XGBoost classifier in the detection of the anomalous episodes.

The marginal total frequencies (with respect to a country index m) of the features selected through the RF module are presented as stacked bar charts in Figure C.4 in Appendix C. The plot shows that when predicting the FSI status $Z_k^{(m)}$ of a specific country, the RF module would more frequently select those features belonging to that specific country under consideration rather than the features belonging to other countries. Figure C.5 in Appendix C is a visualisation of the marginal averaging importance scores (with respect to a country index m) of those selected features. It is noticeable that the average importance scores of the HMM-related features are higher than those of the non-HMM features. This result, along with the previous error analysis, further confirms that the features generated by the HMM have stronger impacts on forecasting the upcoming anomalous episodes. It could also be observed that in the process of detecting anomalies in a specific country's FSI, the features belonging to that country in question would have the highest importance scores relative to the features' scores of other countries. It could also be seen that, when the forecasting horizon h grows, the leading position of the importance score for each country's individual feature relegates to lower ranks. In other words, the predicted anomalous/normal state of a country's FSI tends to be increasingly affected by other countries' FSI status as the prediction horizon h increases. This phenomenon could be treated as a numerical realisation of a country's idiosyncratic FSI and an aberration could propagate across countries to precipitate some systemic anomalous episodes.

Figure C.6 in Appendix C exhibits the marginal total frequencies (with respect to a time lag q) of the features selected by the RF module. The chart shows that the features with $q = 0$ value are more likely to be selected by the RF. The marginal averaging importance scores (with respect to a time lag q) are illustrated in Figure C.7 in Appendix C. It could be observed that the HMM-related features get notably higher scores than the non-HMM features do, specifically, the original discriminative signals. We could also notice that the features with a time lag $q = 0$ achieve the highest importance scores. Nevertheless, the importance scores of features with a time lag $q > 0$ climb slightly as the forecasting horizon h rises. This indicates that the projected financial stress anomalous/normal status is characterised by a strong short-memory dependence when the prediction step is small and by a weak long-memory dependence when the forecasting step increases. This serves as a numerical confirmation that a country's financial stress anomalies could accumulate to trigger a future crisis event.

5. Forecasting with anomaly-warning signals

As defined in Eq. (34), $\tilde{Z}^{(m)}_{k+h}$, which is obtained from the probabilistic outcomes generated by the XGBoost classifier, provides a set of alert signals that quantify the FSI's exposure to aberrations in the future in a point-wise manner. In this section, we shall construct two warning signals to forecast the extent of the FSI anomaly in the near future.

[2] According to the XGBoost R package document, the gain score is defined as the improvement in accuracy attained by a feature to the branches which it is on. The details of the Gain score's calculation could be found in the XGBoost Tutorials.

More specifically, given a fixed time horizon, we are going to forecast whether some extreme abnormality will happen which, in consequence, could precipitate high financial stress episodes in the short term.

We define two types of extreme abnormality episodes for all 17 countries' FSI as follows: (i) Multiple-Anomaly Episodes (MAE): $A_h^{(m)}$; and (ii) Consecutive-Anomaly Episodes (CAE): $H_h^{(m)}$. We say that the FSI enters the MAE at time t if the discriminative signal $Y_t^{(m)} \geq 0$ ($Z_t^{(m)} = 1$) for more than 50% of the time in the interval $[t,\ t+h-1]$. Thus, we define the actual and projected Multiple-Anomaly-Episodes as follows:

$$\begin{cases} A_h^{(m)} := \left\{ t \,\middle|\, \frac{1}{h}\sum_{i=0}^{h-1} Z_{t+i}^{(m)} \geq 0.5 \right\}. \\ \tilde{A}_h^{(m)} := \left\{ t \,\middle|\, \frac{1}{h}\sum_{i=0}^{h-1} \tilde{Z}_{t+i}^{(m)} \geq 0.5 \right\}. \end{cases} \tag{44}$$

The FSI is said to move into a Consecutive-Anomaly Episodes (CAE) at time t if they exists, in $[t,\ t+h-1]$, whenever at least one sub-interval with a length greater than 1 where $Y_t^{(m)} \geq 0$ ($Z_t^{(m)} = 1$) for all the time points in it. In other words, there exists at least one run of 1 for $Z_t^{(m)}$ in $[t,\ t+h-1]$. Therefore, the actual and projected CAE could be expressed as

$$\begin{cases} H_h^{(m)} := \left\{ t \,\middle|\, \exists\, j,l \in \{0,1,2,\dots,h-1\},\ j<l \text{ s.t.} \prod_{i=j}^{l} Z_{t+i}^{(m)} = 1 \right\} \\ \tilde{H}_h^{(m)} := \left\{ t \,\middle|\, \exists\, j,l \in \{0,1,2,\dots,h-1\},\ j<l \text{ s.t.} \prod_{i=j}^{l} \tilde{Z}_{t+i}^{(m)} = 1 \right\}. \end{cases} \tag{45}$$

Note that in Eqs. (44) and (45), h is a prediction-step parameter which takes integer values in $[2, 6]$.

According to Eq. (44), $A_h^{(m)}$ incorporates the time points where $Z_t^{(m)}$ is positive in no less than 50% percent of the next h time steps which indicates that the FSI has rising edges multiple times within the next $h-$step-ahead points. Eq. (45) comprises those time points, where $Z_t^{(m)}$ is positive successively within an $h-$step-ahead horizon; this implies that the FSI is likely to surge on the next h time points. Since the actual MAE and CAE, $\tilde{A}_h^{(m)}$ and $\tilde{H}_h^{(m)}$, respectively, are related to $\tilde{Z}_t^{(m)}$ which is the predicted value of $Z_t^{(m)}$, the projected MAE and CAE could serve as the respective predictions of $A_h^{(m)}$ and $H_h^{(m)}$.

In order to capture the extremely anomalous episodes, we construct two alert signals utilising the MAE and CAE: (i) Multiple-Anomaly Indicators (MAI); and (ii) Consecutive-Anomaly Indicators (CAI). The formulation for these two collections of warning signals are as follows:

$$\begin{cases} \text{MAI:} \quad U_{k,h}^{(m)} = \mathbf{1}_{\left\{k+1 \,\in\, \tilde{A}_h^{(m)}\right\}} \\ \text{CAI:} \quad V_{k,h}^{(m)} = \mathbf{1}_{\left\{k+1 \,\in\, \tilde{H}_h^{(m)}\right\}}, \end{cases} \tag{46}$$

where $\mathbf{1}_{\{\cdot\}}$ is an indicator function and $h \in \{2,3,\dots,6\}$. The MAI ($U_{k,h}^{(m)}$) takes the value 1 at time k when $\frac{1}{h}\sum_{i=0}^{h-1} \tilde{Z}_{k+1+i}^{(m)} \geq 0.5$ and 0 otherwise. Thus, it provides an aggregate value to project the mean intensity of the anomalies of the FSI in the upcoming h time points. The CAI ($V_{k,h}^{(m)}$) signals a positive at time k if there exists some integers$j$ and l with $0 \leq j < l \leq h-1$ such that $\prod_{i=j}^{l} \tilde{Z}_{k+1+i}^{(m)} = 1$. Therefore, the CAI yields an aggregate value to predict the extent of the propagation of anomalies of FSI over the $h-$step ahead time horizon.

Figures C.8 and C.9 in Appendix C depict the MAI and CAI. Note that the binary signals are plotted in a descending order of the parameter h, from top to bottom in all graphs. We see that the signal captures the FSIs' rising edges (light purple shaded areas) for all the 17 countries included in our data set. Evidently, the MAI and CAI are capable of predicting the extent of the anomalies of FSI in the upcoming $h-$step-ahead horizon since all the signals are generated in an adapted way.

The error analysis for MAI is displayed in Table 4. The medians of the TPR for 3 and 6-step-ahead predictions are 80.97% and 76.89%, respectively. The medians of the ACC are 81.29% and 77.38% associated with the 3 and 6 forecasting horizons, respectively. According to the minimums of TPR and ACC of $U_{k,6}^{(m)}$, MAI is able to detect more than 71% of the MAE in the upcoming 2–6 months amongst all countries' FSI with false discovery rate (1-PPV) less than 29%.

The performance of the CAI is reported in Table 5. The TPR medians for the 3 and 6 prediction horizons are 78.95% and 77.15%, respectively. The ACC medians are 81.47% and 77.38% for the 3 and 6-step-ahead forecasting, respectively. The minimums of TPR and ACC of $V_{k,6}^{(m)}$ indicate that the CAI is capable of identifying at least 71% of the CAE in the next 2–6 months amongst all countries' FSI with false discovery rate less than 23%.

The performance measurements provided in Tables 4 and 5 justify that the prediction power of the proposed model could be further extended via the introduction of two anomaly-warning signals: MAI and CAI. The major advantages of MAI and CAI are the (i) capacity to detect and differentiate different types of extreme anomalies of FSI in advance; and (ii) specification of the time via the parameter h to capture extreme abnormalities. The two types of indicators, along with the point-wise predicted value $Z_k^{(m)}$, certainly provide a comprehensive quantitative framework to identify extreme abnormalities of FSI in the near future for practitioners.

6. Concluding remarks

The development of a hybrid multivariate supervised learning device is this paper's key contribution. Our algorithmic approach integrated stochastic process modelling, hidden Markov model, random forest and XGBoost. The application in this research focused on tracing the movement of multiple countries' FSIs simultaneously in which signs for possible crisis episodes that will disturb financial stability are detected early.

A discriminative signal, which is evaluated as the lag-τ difference of the original data series, is introduced to characterise the anomalous behaviour of all FSI time series. The OU processes, with parameters governed by the HMM, are employed to model 17 different countries' discriminative signals altogether. Utilising the EM algorithm and change of reference probability measures, the HMM online recursive filters were constructed. This in turn, unveils the information content of the observed multiple stochastic processes. Two new features, i.e., estimate of the Markov-chain state and deviation from the mean, were created to describe the characteristics of the underlying time series under the HMM filtering framework. A feature-selection module was established that consequently improves the out-of-sample predictive performance at a lower the computing cost of modelling. Random Forest was utilised to select the HMM and Non-HMM predictors concurrently based on the OOB accuracy.

XGBoost is the final stage classifier in our modelling approach. Time series cross-validation combined with grid search was applied to tune the hyper-parameters. In order to fully capture the evolution of the input data series, the classification model is tuned and trained once new data is available. To avoid error propagation in multi-step-ahead forecasting, a direct prediction method was applied to generate a multivariate 6-step ahead predictions for the countries' financial stress status, which served as an early-warning signal for the future occurrence of financial-crisis episodes.

Benchmarking with other five models, our model's implementability and predictive performance were tested on the actual multi-dimensional time series of 17 countries' FSI. Each models' predictive power was assessed though four different model-diagnostic tools: (i) confusion matrix, (ii) area under ROC, (iii) logistic loss, and (iv) Kolmogorov–Smirnov test. The confusion-matrix analysis showed that our model had the highest TPR, TNR, PPV and NPV scores in all forecasting horizons. This means that our proposed model is able to detect positive

Table 4
An evaluation of the classification performance for the MAIs: $U_{k,h}^{(m)}$.

h		AT	AU	CA	CH	DE	DK	ES	FI	FR	GB	IT	JP	KR	NL	NZ	SE	US	Min	Med	Max
2	ACC	0.8434	0.8057	0.8377	0.8132	0.8264	0.8434	0.7981	0.8302	0.8075	0.7962	0.8472	0.7962	0.8226	0.8094	0.8434	0.8113	0.7943	0.7943	0.8132	0.8472
	TPR	0.8367	0.8188	0.8182	0.8293	0.8194	0.8459	0.8127	0.8185	0.8043	0.7953	0.8212	0.8051	0.8170	0.8000	0.8182	0.7915	0.8031	0.7915	0.8182	0.8459
	TNR	0.8517	0.7873	0.8649	0.7942	0.8355	0.8400	0.7767	0.8472	0.8125	0.7979	0.8900	0.7834	0.8310	0.8256	0.8755	0.8442	0.7810	0.7767	0.8310	0.8900
	PPV	0.8754	0.8433	0.8936	0.8264	0.8657	0.8776	0.8421	0.8862	0.8691	0.8730	0.9249	0.8428	0.8780	0.8874	0.8934	0.8942	0.8482	0.8264	0.8754	0.9249
	NPV	0.8072	0.7565	0.7742	0.7975	0.7814	0.8008	0.7389	0.7625	0.7284	0.6906	0.7511	0.7359	0.7532	0.7061	0.7907	0.7089	0.7225	0.6906	0.7532	0.8072
3	ACC	0.8412	0.7807	0.8166	0.8129	0.8355	0.8166	0.8072	0.8034	0.8412	0.7977	0.8091	0.7977	0.8393	0.8204	0.8318	0.7940	0.7788	0.7788	0.8129	0.8412
	TPR	0.8340	0.7712	0.7782	0.8233	0.8285	0.8097	0.8112	0.7903	0.8315	0.7790	0.7918	0.7967	0.8388	0.8106	0.8122	0.7395	0.7698	0.7395	0.8097	0.8388
	TNR	0.8475	0.7884	0.8555	0.8047	0.8414	0.8227	0.8036	0.8149	0.8511	0.8168	0.8269	0.7986	0.8397	0.8302	0.8467	0.8470	0.7870	0.7870	0.8269	0.8555
	PPV	0.8273	0.7459	0.8449	0.7671	0.8115	0.8000	0.7860	0.7903	0.8506	0.8125	0.8256	0.7747	0.8153	0.8263	0.8017	0.8248	0.7668	0.7459	0.8115	0.8506
	NPV	0.8536	0.8105	0.7923	0.8536	0.8561	0.8315	0.8272	0.8149	0.8321	0.7839	0.7934	0.8188	0.8607	0.8148	0.8552	0.7695	0.7899	0.7695	0.8188	0.8607
4	ACC	0.8030	0.7557	0.7898	0.7727	0.8011	0.7973	0.7936	0.7841	0.8239	0.7860	0.8011	0.7992	0.8182	0.7917	0.8163	0.7689	0.7670	0.7557	0.7936	0.8239
	TPR	0.8216	0.7821	0.7600	0.7868	0.8007	0.7840	0.7979	0.7708	0.8271	0.7778	0.7834	0.8007	0.8156	0.7799	0.8127	0.7533	0.7617	0.7533	0.7840	0.8271
	TNR	0.7838	0.7258	0.8289	0.7578	0.8016	0.8133	0.7881	0.8000	0.8197	0.7981	0.8271	0.7975	0.8211	0.8082	0.8199	0.7902	0.7739	0.7258	0.8000	0.8289
	PPV	0.7978	0.7631	0.8539	0.7754	0.8155	0.8333	0.8233	0.8222	0.8531	0.8507	0.8693	0.8237	0.8394	0.8516	0.8220	0.8297	0.8136	0.7631	0.8237	0.8693
	NPV	0.8088	0.7469	0.7241	0.7698	0.7860	0.7597	0.7592	0.7442	0.7893	0.7083	0.7224	0.7720	0.7953	0.7224	0.8106	0.7024	0.7149	0.7024	0.7592	0.8106
5	ACC	0.8197	0.7173	0.8083	0.7609	0.8197	0.8046	0.7932	0.7932	0.8216	0.7552	0.7761	0.7951	0.8216	0.7818	0.8178	0.7476	0.7552	0.7173	0.7932	0.8216
	TPR	0.8333	0.7156	0.7739	0.7585	0.8095	0.7860	0.7927	0.7875	0.7940	0.7280	0.7463	0.7886	0.8101	0.7689	0.8062	0.7043	0.7276	0.7043	0.7860	0.8333
	TNR	0.8084	0.7185	0.8421	0.7629	0.8277	0.8204	0.7936	0.7979	0.8500	0.7820	0.8069	0.8007	0.8310	0.7947	0.8267	0.7889	0.7794	0.7185	0.8007	0.8500
	PPV	0.7843	0.6545	0.8279	0.7218	0.7857	0.7893	0.7708	0.7652	0.8446	0.7661	0.8000	0.7760	0.7967	0.7899	0.7787	0.7605	0.7427	0.6545	0.7787	0.8446
	NPV	0.8529	0.7722	0.7915	0.7957	0.8478	0.8175	0.8139	0.8179	0.8007	0.7455	0.7545	0.8123	0.8427	0.7741	0.8493	0.7370	0.7657	0.7370	0.8007	0.8529
6	ACC	0.8042	0.7205	0.7909	0.7662	0.7947	0.7776	0.7719	0.7681	0.8004	0.7738	0.7548	0.7776	0.8042	0.7605	0.7795	0.7452	0.7357	0.7205	0.7738	0.8042
	TPR	0.8346	0.7366	0.7587	0.7744	0.7913	0.7656	0.7690	0.7727	0.7884	0.7641	0.7318	0.7818	0.8060	0.7492	0.7689	0.7220	0.7188	0.7188	0.7689	0.8346
	TNR	0.7744	0.7045	0.8292	0.7577	0.7978	0.7905	0.7751	0.7634	0.8155	0.7867	0.7857	0.7729	0.8023	0.7753	0.7901	0.7749	0.7563	0.7045	0.7753	0.8292
	PPV	0.7834	0.7122	0.8411	0.7658	0.7852	0.7977	0.7918	0.7669	0.8431	0.8273	0.8216	0.7904	0.8090	0.8145	0.7868	0.8038	0.7811	0.7122	0.7918	0.8431
	NPV	0.8273	0.7294	0.7425	0.7665	0.8037	0.7576	0.7510	0.7692	0.7540	0.7137	0.6848	0.7638	0.7992	0.7012	0.7724	0.6858	0.6897	0.6848	0.7540	0.8273

Table 5
An evaluation of the classification performance for the CAIs: $V_{k,h}^{(m)}$.

h		AT	AU	CA	CH	DE	DK	ES	FI	FR	GB	IT	JP	KR	NL	NZ	SE	US	Min	Med	Max
2	ACC	0.8642	0.8245	0.8547	0.8283	0.8509	0.8509	0.8377	0.8208	0.8528	0.8151	0.8358	0.8472	0.8491	0.8226	0.8491	0.8415	0.8132	0.8132	0.8415	0.8642
	TPR	0.8520	0.7955	0.8241	0.7869	0.8462	0.8209	0.8316	0.8043	0.8519	0.7500	0.7915	0.8238	0.8396	0.8223	0.8222	0.7970	0.7903	0.7500	0.8222	0.8520
	TNR	0.8713	0.8390	0.8758	0.8501	0.8534	0.8693	0.8413	0.8295	0.8535	0.8545	0.8652	0.8605	0.8542	0.8228	0.8629	0.8679	0.8256	0.8228	0.8542	0.8758
	PPV	0.7952	0.7107	0.8203	0.7347	0.7512	0.7933	0.7546	0.7150	0.8000	0.7576	0.7952	0.7718	0.7585	0.7330	0.7551	0.7811	0.7101	0.7101	0.7576	0.8203
	NPV	0.9094	0.8919	0.8786	0.8832	0.9138	0.8882	0.8949	0.8885	0.8933	0.8494	0.8625	0.8951	0.9071	0.8867	0.9042	0.8784	0.8793	0.8494	0.8885	0.9138
3	ACC	0.8526	0.7732	0.8185	0.8053	0.8318	0.8147	0.8147	0.7940	0.8336	0.7902	0.8072	0.8091	0.8393	0.8166	0.8185	0.7940	0.7713	0.7713	0.8147	0.8526
	TPR	0.8489	0.7535	0.7720	0.7917	0.8203	0.7931	0.8139	0.7798	0.8347	0.7427	0.7724	0.7895	0.8465	0.8109	0.7895	0.7393	0.7723	0.7393	0.7895	0.8489
	TNR	0.8553	0.7866	0.8602	0.8147	0.8397	0.8316	0.8154	0.8039	0.8327	0.8299	0.8375	0.8239	0.8344	0.8213	0.8375	0.8373	0.7705	0.7705	0.8316	0.8602
	PPV	0.8128	0.7074	0.8319	0.7467	0.7807	0.7863	0.7737	0.7359	0.8150	0.7851	0.8051	0.7725	0.7778	0.7878	0.7604	0.7828	0.7119	0.7074	0.7807	0.8319
	NPV	0.8844	0.8233	0.8081	0.8500	0.8704	0.8373	0.8497	0.8389	0.8509	0.7940	0.8089	0.8378	0.8881	0.8415	0.8590	0.8019	0.8217	0.7940	0.8389	0.8881
4	ACC	0.8390	0.7386	0.7784	0.7708	0.8182	0.7898	0.8030	0.7727	0.8201	0.7860	0.7973	0.7803	0.8182	0.7955	0.8049	0.7576	0.7595	0.7386	0.7898	0.8390
	TPR	0.8419	0.7362	0.7394	0.7671	0.8135	0.7643	0.8075	0.7698	0.8107	0.7553	0.7857	0.7719	0.8230	0.7849	0.8025	0.7159	0.7557	0.7159	0.7719	0.8419
	TNR	0.8364	0.7409	0.8238	0.7742	0.8225	0.8151	0.7985	0.7754	0.8306	0.8211	0.8105	0.7887	0.8140	0.8072	0.8069	0.8016	0.7632	0.7409	0.8072	0.8364
	PPV	0.8256	0.7248	0.8300	0.7520	0.8071	0.8040	0.8015	0.7578	0.8439	0.8288	0.8240	0.7838	0.7905	0.8202	0.7733	0.7918	0.7586	0.7248	0.8015	0.8439
	NPV	0.8519	0.7519	0.7309	0.7883	0.8285	0.7770	0.8046	0.7868	0.7954	0.7454	0.7701	0.7770	0.8436	0.7701	0.8327	0.7279	0.7603	0.7279	0.7770	0.8519
5	ACC	0.8046	0.7400	0.7571	0.7571	0.7970	0.7780	0.7837	0.7476	0.8065	0.7837	0.7799	0.7818	0.7913	0.7666	0.7913	0.7324	0.7533	0.7324	0.7799	0.8065
	TPR	0.8333	0.7657	0.7184	0.7706	0.7978	0.7552	0.7904	0.7545	0.8140	0.7588	0.7818	0.7751	0.7955	0.7630	0.8092	0.7209	0.7543	0.7184	0.7706	0.8333
	TNR	0.7743	0.7095	0.8119	0.7419	0.7960	0.8059	0.7754	0.7400	0.7965	0.8194	0.7773	0.7899	0.7871	0.7717	0.7736	0.7478	0.7521	0.7095	0.7754	0.8194
	PPV	0.7951	0.7578	0.8441	0.7706	0.8125	0.8264	0.8127	0.7628	0.8419	0.8582	0.8304	0.8175	0.7895	0.8246	0.7794	0.7920	0.7921	0.7578	0.8125	0.8582
	NPV	0.8156	0.7185	0.6705	0.7419	0.7804	0.7290	0.7500	0.7312	0.7627	0.7024	0.7185	0.7431	0.7931	0.6983	0.8039	0.6680	0.7097	0.6680	0.7312	0.8156
6	ACC	0.7966	0.7452	0.7433	0.7567	0.7776	0.7757	0.7776	0.7338	0.7776	0.7833	0.7738	0.7510	0.7738	0.7567	0.7757	0.7357	0.7510	0.7338	0.7738	0.7966
	TPR	0.8182	0.7683	0.7100	0.7738	0.7900	0.7692	0.7917	0.7450	0.7875	0.7715	0.7879	0.7613	0.7965	0.7522	0.7986	0.7455	0.7563	0.7100	0.7715	0.8182
	TNR	0.7708	0.7109	0.8000	0.7330	0.7611	0.7850	0.7570	0.7193	0.7621	0.8042	0.7500	0.7361	0.7469	0.7644	0.7490	0.7194	0.7427	0.7109	0.7500	0.8042
	PPV	0.8097	0.7987	0.8577	0.8000	0.8144	0.8392	0.8261	0.7762	0.8372	0.8754	0.8414	0.8055	0.7882	0.8485	0.7875	0.8173	0.8203	0.7762	0.8173	0.8754
	NPV	0.7806	0.6726	0.6190	0.7013	0.7319	0.7000	0.7137	0.6833	0.6978	0.6638	0.6774	0.6824	0.7563	0.6376	0.7615	0.6267	0.6623	0.6190	0.6833	0.7806

(i.e., anomalous) episodes more accurately with lower false positive error than the benchmark models when the discrimination threshold is set to 0.5. The comparison of AUCs showed that the proposed model scored higher than the other models at all prediction steps. The outcomes of the pairwise t-test and DeLong test supported the statistical significance of our model's AUC score. The logistic loss of the models revealed that our model has better capability in determining anomalous periods; this is based on higher predicted probabilities for positive instances than for negative ones. The result of the Kolmogorov–Smirnov test attested to the robustness of our proposed modelling approach in differentiating anomalous events from regular ones. Our modelling methodology for multivariate prediction substantially benefited from the fusion of RF, HMM and XGBoost.

A customised feature-importance analysis was conducted to measure the features' contribution to detect anomalous episodes. The results suggested that the HMM-features have significant impacts on the prediction accuracy. Our findings also indicate that, as the forecasting horizon increases, the prediction of anomalies of each country's FSI could have a long-memory effect on other countries' features.

Employing the anomalous/normal classification results, two projected anomaly-warning signals (MAI and CAI) are constructed to predict two types of extremely anomalous events (MAE and CAE). The results show that our proposed modelling approach is able to detect the specified extremely anomalous episodes in the near future (2–6 months ahead) with a good accuracy.

This research work also contributed to the integration of supervised learning, multivariate time series, and financial-stability modelling in the following ways: (i) Development of a new framework integrating HMM with the supervised learning models: RF and XGBoost. This enabled us to trace with accuracy the dynamics of multiple FSI time series in an attempt to detect future financial stability; (ii) Utility of a multiple self-tuning estimation procedure under the HMM set up to generate new features for classification; (iii) Implementation of a 2-dimensional feature selection using RF; (iv) Application of time series cross validation and grid search for hyper-parameter tuning for XGBoost; (v) Independent training framework for different prediction steps; (vi) A comprehensive model performance diagnostics using multiple statistical tools. (vii)A customised analysis of the features' importance to assess and interpret different types of input variables' impact on the anomaly detection; and (viii) Construction of two early warning alert signals MAI and CAI to detect extremely anomalous episodes in the future.

This paper laid out a new ensemble supervised learning classification model which is designed to detect future anomalous episode location for the benefit of financial-stability research and pertinent policy implications. Our proposed model provides predictive analytics for country-level FSI status which is relevant to financial-stability management. Our approach formulated a framework that allows us to recast our solution to anomalous/normal classification outcomes. Since these quantities are also given in a probabilistic form, they have natural interpretations as an estimated level of exposure to financial

instability which is relevant to central bankers and economists. In addition, we conducted a tailored features' importance analysis to quantify the relationship between the predicted FSI's anomalous/normal status and different types of input variables; this in turn provides accessible interpretation of our proposed modelling to practitioners. Lastly, we enhanced further our model by constructing two projected warning signals, MAI and CAI, to forecast some specified extremely anomalous episodes in the near future. The said enhancement gives additional quantitative insights for the policy makers in mitigating possible financial-crisis event.

Although our methodology and model have limitations, they serve as springboard for future research directions. It would be worth exploring how our approach will fair in assessing the predictive ability of crisis indicators that may change due to new risk factors or policy actions; see Duprey and Klaus [48]. This type of investigation may require designing heuristics to make the data frequency of various risk factors uniform in the creation of indicators.

The application of machine learning techniques has gained traction recently in estimating the fair market values of a large number of variable contracts (cf. [49]). However, combining this idea with the regime-switching framework driven by HMM (e.g., [50]) remains unexamined. The use of HMM filters is natural for HMM setting. Nevertheless, the literature on blending machine learning methods with other types of filters in non-linear time series models (e.g., modified sigma point filters [51]) is not well-developed yet.

In this paper, we developed a hybrid model combining the HMM with RF and XGBoost. The result demonstrated that HMMs could be utilised as an ideal feature-generating device in a machine learning framework. Implementation of the HMM in combination with a certain type of Neural Network is also anticipated to bring high predictive power in the context of big-data processing environment.

In the field of machine learning, it is deemed that both the HMM and higher-order HMM (HOHMM) are unsupervised learning models. As demonstrated in this work, the HMM could be embedded in a supervised learning model to achieve good forecasting results. In this paper, setting the number of regimes to 2 definitely simplifies the mechanics of the HMM development. As a generalisation, pinning down the data's stylised facts including the unknown number of regimes could be investigated and require a methodology that enables the HMM or HOHMM (see [52]) satisfy certain objectives, i.e., by, for example, minimising the MSE or classification error. More specifically, we could treat the number of states of the HMM and the memory length of HOHMM as hyper-parameters. The optimal values could be searched via cross validation.

Lastly, from the computing point of view, time complexity, which describes the time required to execute an algorithm, is a crucial consideration in a detection method. In our case, we indicated the time required to run our algorithm in Table 2. The size of the processed data, network load and type of machine used have a high impact on this type of complexity analysis. Since the frequency of the observations in our data sets is monthly, time complexity is not a significant issue here compared to financial or business modelling situations where updating might be conducted daily. Nonetheless, the evaluation of the asymptotic behaviour of the associated algorithms akin to the proposed model and benchmark models is an important consideration. Therefore, more advances in time-complexity analysis must be pursued in relation to the results of this research. The running times of a given algorithm can be viewed as some functions dependent on its inputs that grow in size. It must be noted that these functions could be designed to correspond to running times independent of the machine used, programming language, style of programmer, etc. So, employing these functions' limiting behaviour, as their inputs or arguments approach a certain value or infinity, could be more meaningful than simply relying on execution times in algorithm comparison.

CRediT authorship contribution statement

Xing Gu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Rogemar Mamon:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Thibaut Duprey:** Conceptualization, Formal analysis, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through R. Mamon's Discovery Grant (RGPIN-2017-04235).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2024.111712.

References

[1] B. Gadanecz, K. Jayaram, Measures of financial stability-a review, Irving Fish. Comm. Bull. 31 (1) (2008) 365–383.
[2] T. Duprey, B. Klaus, T. Peltonenc, Dating systemic financial stress episodes in the EU countries, J. Financial Stab. 32 (2017) 357–384.
[3] L. Alessi, C. Detken, Identifying excessive credit growth and leverage, J. Financial Stab. 35 (2018) 215–225.
[4] R. Duttagupta, P. Cashin, Anatomy of banking crises in developing and emerging market countries, J. Int. Money Finance 30 (2) (2011) 354–376.
[5] F. Ward, Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors, J. Appl. Econometrics 32 (2) (2017) 359–378.
[6] E. Casabianca, M. Catalano, L. Forni, E. Giarda, S. Passeri, An early warning system for banking crises: From regression-based analysis to machine learning techniques, 2019, Marco Fanno Working Papers 0235.
[7] M. Fioramanti, Predicting sovereign debt crises using artificial neural networks: A comparative approach, J. Financial Stab. 4 (2) (2008) 149–164.
[8] J. Li, W. Pedrycz, I. Jamal, Multivariate time series anomaly detection: A framework of Hidden Markov models, Appl. Soft Comput. 60 (2017) 229–240.
[9] X. Cao, X. Li, S. Coleman, A. Belatreche, T. McGinnity, Adaptive hidden Markov model with anomaly states for price manipulation detection, IEEE Trans. Neural Netw. Learn. Syst. 26 (2) (2015) 318–330.
[10] J. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica 57 (2) (1989) 357–384.
[11] R. Mamon, C. Erlwein, B. Gopaluni, Adaptive signal processing of asset price dynamics with predictability analysis, Inform. Sci. 178 (1) (2008) 203–219.
[12] R. Elliott, L. Aggoun, J. Moore, Hidden Markov Models: Estimation and Control, Springer, New York, 1995.
[13] R. Elliott, W. Hunter, B. Jamieson, Financial signal processing: A self calibrating model, Int. J. Theor. Appl. Finance 4 (4) (2001) 567–584.
[14] R. Elliott, R. Mamon, An interest rate model with a Markovian mean-reverting level, Quant. Finance 2 (6) (2002) 454–458.
[15] C. Erlwein, R. Mamon, An online estimation scheme for a Hull–White model with HMM-driven parameters, Stat. Methods Appl. 18 (1) (2009) 87–107.
[16] N. Zhou, R. Mamon, An accessible implementation of interest rate models with markov-switching, Expert Syst. Appl. 39 (5) (2012) 4679–4689.
[17] X. Xi, R. Mamon, Capturing the regime-switching and memory properties of interest rates, Comput. Econ. 44 (3) (2013) 307–337.
[18] A. Tenyakov, R. Mamon, A computing platform for pairs-trading online implementation via a blended Kalman-HMM filtering approach, J. Big Data 4 (2017) 46.

[19] C. Erlwein, R. Mamon, M. Davison, An examination of HMM-based investment strategies for asset allocation, Appl. Stoch. Models Bus. Ind. 27 (3) (2009) 204–221.
[20] P. Date, R. Mamon, A. Tenyakov, Filtering and forecasting commodity futures prices under an HMM framework, Energy Econ. 40 (2013) 1001–1013.
[21] H. Xiong, R. Mamon, A self-updating model driven by a higher-order hidden Markov chain for temperature dynamics, J. Comput. Sci. 17 (1) (2016) 47–61.
[22] H. Xiong, R. Mamon, Putting a price tag on temperature, Comput. Manag. Sci. 15 (2) (2017) 1–38.
[23] X. Gu, R. Mamon, M. Davison, H. Yu, An automated financial indices-processing scheme for classifying market liquidity regimes, Internat. J. Control 94 (3) (2021) 735–756.
[24] C. Erlwein, F. Benth, R. Mamon, HMM filtering and parameter estimation of an electricity spot price model, Energy Econ. 32 (5) (2010) 1034–1043.
[25] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[26] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, 2017.
[27] H. Tyralis, G. Papacharalampous, Variable selection in time series forecasting using random forests, Algorithms 10 (4) (2017) 114–138.
[28] S. Karasu, A. Altan, Recognition model for solar radiation time series based on random forest with feature selection approach, in: 11th International Conference on Electrical and Electronics Engineering, ELECO Bursa, Turkey, 2019.
[29] P. Hao, X. Zhan, L. Wang, Z. Niu, M. Shakir, Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA, Remote Sens. 7 (5) (2015) 5347–5369.
[30] R.E. Schapire, A brief introduction to boosting, in: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Vol. 2, 1999, pp. 1401–1406.
[31] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[32] L. Breiman, J. Friedman, C.J. Stone, R. Olshen, Classification and Regression Trees, Taylor & Francis, Boca Raton, 1984.
[33] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. 29 (5) (2001) 1189–1232.
[34] W. Zhang, X. Zhao, Z. Li, A comprehensive study of smartphone-based indoor activity recognition via XGBoost, IEEE Access 7 (2019) 80027–80042.
[35] S. Basak, S. Kar, S. Saha, L. Khaidem, S.R. Dey, Predicting the direction of stock market prices using tree-based classifiers, North Am. J. Econ. Finance 47 (2019) 552–567.
[36] J. Nobre, R. Neves, Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets, Expert Syst. Appl. 125 (2019) 181–194.
[37] R. Elliott, V. Krishnamurthy, New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models, IEEE Trans. Automat. Control 44 (5) (1999) 938–951.
[38] C. Wu, On the convergence properties of the EM algorithm, Ann. Statist. 11 (1) (1983) 95–103.
[39] A. Tenyakov, R. Mamon, M. Davison, Modelling high-frequency FX rate dynamics: A zero-delay multi-dimensional HMM-based approach, Knowl.-Based Syst. 101 (2016) 142–155.
[40] A. Tenyakov, R. Mamon, M. Davison, Filtering of a discrete-time HMM-driven multivariate ornstein-uhlenbeck model with application to forecasting market liquidity regimes, IEEE J. Sel. Top. Sign. Proces. 10 (6) (2016) 994–1005.
[41] X. Gu, R. Mamon, T. Duprey, H. Xiong, Online estimation for a predictive analytics platform with a financial-stability-analysis application, Eur. J. Control 57 (2021) 205–221.
[42] S.B. Taieb, A. Sorjamaa, G. Bontempi, Multiple-output modeling for multi-step-ahead time series forecasting, Neurocomputing 73 (2010) 1950–1957.
[43] A. Venkatraman, M. Hebert, J.A. Bagnell, Improving Multi-Step Prediction of Learned Time Series Models, Association for the Advancement of Artificial Intelligence, 2015, pp. 3024–3030.
[44] H. Cheng, P. Tan, J. Gao, J. Scripps, Multistep-ahead time series prediction, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2006, pp. 765–774.
[45] C. Bergmeir, J. Benítez, On the use of cross-validation for time series predictor evaluation, Inform. Sci. 191 (2012) 192–213.
[46] R.J. Hyndman, G. Athanasopoulos, Forecasting Principles and Practice, second ed., OTexts, Melbourne, 2018.
[47] E. DeLong, D. DeLong, D. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics 44 (3) (1988) 837–845.
[48] T. Duprey, B. Klaus, Early warning or too late? A (pseudo-)real-time identification of leading indicators of financial stress, J. Bank. Financ. 138 (2022) 106196.
[49] H. Gweon, S. Li, R. Mamon, An effective bias-corrected bagging method for the valuation of large variable annuity portfolios, Astin Bull. 50 (3) (2020) 853–871.
[50] H. Gao, R. Mamon, X. Liu, Pricing a guaranteed annuity option under correlated and regime-switching risk factors, Eur. Actuar. J. 5 (2015) 309–326.
[51] P. Date, L. Jalen, R. Mamon, A new algorithm for latent state estimation in non-linear time series models, Appl. Math. Comput. 203 (2008) 224–232.
[52] H. Xiong, R. Mamon, An enabling framework for automated extraction of signals from market information in real time, Knowl.-Based Syst. 246 (2022) 108612.