



HAL
open science

Early warning system for currency crises using long short-term memory and gated recurrent unit neural networks

Sylvain Barthélémy, Virginie Gautier, Fabien Rondeau

► **To cite this version:**

Sylvain Barthélémy, Virginie Gautier, Fabien Rondeau. Early warning system for currency crises using long short-term memory and gated recurrent unit neural networks. *Journal of Forecasting*, 2024, 10.1002/for.3069 . hal-04470367

HAL Id: hal-04470367

<https://hal.science/hal-04470367>

Submitted on 21 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Early warning system for currency crises using long short-term memory and gated recurrent unit neural networks

Sylvain Barthélémy¹ | Virginie Gautier²  | Fabien Rondeau³

¹Gwenlake and CREM, University of Rennes, Rennes, France

²TAC Economics and CREM, University of Rennes, Rennes, France

³CREM, University of Rennes, Rennes, France

Correspondence

Virginie Gautier, TAC Economics and CREM, University of Rennes, Rennes, France.

Email: virginie.gautier@univ-rennes1.fr

Abstract

Currency crises, recurrent events in the economic history of developing, emerging, and developed countries, have disastrous economic consequences. This paper proposes an early warning system for currency crises using sophisticated recurrent neural networks, such as long short-term memory (LSTM) and gated recurrent unit (GRU). These models were initially used in language processing, where they performed well. Such models are increasingly being used in forecasting financial asset prices, including exchange rates, but they have not yet been applied to the prediction of currency crises. As for all recurrent neural networks, they allow for taking into account nonlinear interactions between variables and the influence of past data in a dynamic form. For a set of 68 countries including developed, emerging, and developing economies over the period of 1995–2020, LSTM and GRU outperformed our benchmark models. LSTM and GRU correctly sent continuous signals within a 2-year warning window to alert for 91% of the crises. For the LSTM, false signals represent only 14% of the emitted signals compared with 23% for logistic regression, making it an efficient early warning system for policymakers.

KEYWORDS

currency crises, early warning system, gated recurrent unit, long short-term memory, neural network

1 | INTRODUCTION

Currency or balance of payments crises have been studied since the 1970s using the founding models of Krugman (1979) and Flood and Garber (1984), the drivers of the first-generation models. With many currency crises in the 1990s,¹ the literature became considerable, and early warning systems (EWSs) were developed. There are many consequences that can affect both the financial and

real spheres (loss of central bank credibility, debt unsustainability, lack of capital, imported inflation, bankruptcy, and GDP losses). Emerging and developing economies have suffered particularly from these depreciation episodes in the past, highlighting insufficient financial development, over-regulated markets, excessive dollarization of domestic and foreign assets, and recurrent recourse to fixed exchange rate regimes due to the “fear of floating” (Calvo & Reinhart, 2000). More

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Journal of Forecasting* published by John Wiley & Sons Ltd.

recently, the global financial crisis has revealed several deficiencies in financial system regulation, including systemic risk resulting from the interdependence of financial actors, sectors and countries, and an increased need for international cooperation. While recent crises may seem less severe than past episodes, their occurrence persists (the Turkish lira crisis and the collapse of the Argentine peso in 2018) with still disastrous consequences for the affected economies.

To warn of the future occurrence of a sudden drop in currency value and limit its effects, we need to build an efficient EWS. The objective is to send a continuous signal over a given time window before the currency crisis episode begins. The EWS can be based on the monitoring of key indicators (Kaminsky et al., 1998) or on more synthetic approaches, for example, econometric models such as logit and probit regressions (Frankel & Rose, 1996; Gourinchas & Obstfeld, 2011), machine learning models (de Carvalho Filho et al., 2020; Ghosh & Ghosh, 2002), and deep learning models (Nag & Mitra, 1999; Peltonen, 2006). This system should accurately identify periods of vulnerability by sending continuous signals and managing the trade-off between sending multiple signals that also include false signals and sending reduced and less noisy signals that may lead to missing certain crises. Ideally, the warning system should not miss any crisis signals while minimizing false signals, as the cost of a missed signal is higher for policymakers. Focusing on the pre-crisis period, EWS models may have substantial value for policymakers by allowing them to detect underlying economic weaknesses and vulnerabilities, and implement preventive actions to reduce the risks of experiencing crises. In the wake of the 2008 financial crisis, proactive approaches to financial crises were strengthened in response to growing systemic risk. EWSs are part of this trend because the implementation of policies requires measuring the likelihood, magnitude, timing, and determinants of crises.

Using the criterion proposed by Frankel and Rose (1996) to identify a currency crisis, this study develops three types of models that attempt to warn of the occurrence of a currency crisis in the 2 years preceding the collapse. The data sample covers 68 countries including developed, emerging, and developing economies over the period of 1995–2020. First, a logistic regression, a standard model in the EWS literature, and a random forest (RF) were constructed. These two models were used as benchmarks for the proposed approach based on two recurrent neural networks (RNNs): long short-term memory (LSTM) and gated recurrent unit (GRU). Both networks have grown in popularity in recent years. Initially used for natural language processing, they have quickly spread to many fields, some of which are related

to health or economics, notably through the prediction of financial asset prices (Claveria et al., 2022; Dautel et al., 2020) and systemic banking crises (Tölö, 2020). However, the performance of LSTM and GRU models compared with traditional econometric models in the context of currency crises has not been explicitly analyzed in the literature. The complex relationships between signals and predictors can be captured through a data-driven non-parametric and highly nonlinear methodology, such as RF or artificial neural network (ANN). The second advantage of the proposed methodology lies its adaptation to sequential data. Time-series forecasting often requires the inclusion of lagged variables to involve dynamic interactions in an artificial manner. RNNs, such as LSTM and GRU, perfectly integrate this characteristic, owing to feedback loops and memory cells designed to identify short- and long-term dependencies between variables. The RNN preserves the temporal order of the time series so that as observations pass through the network, it can identify the accumulation of vulnerabilities and send an alert when the vulnerabilities become too numerous or when a triggering event occurs. The contribution of LSTM and GRU networks, compared with standard RNNs, lies in the construction of more complex cells, integrating a memory vector updated through a sophisticated gate mechanism. A standard RNN was constructed to measure the contribution of these new operations within the cells.

The contributions of this study can be summarized as follows. First, we used the expanding window method, which is an adaptation of cross-validation to time series, with a distinction between validation and test samples. During the training, the hyperparameters were adjusted according to the performance on the validation sample. The final performance was measured on the test sample that was not known to the model. This procedure allows to test the real generalization capacity of the models. Second, currency crises remain episodic, which can hinder the training of the models while leading to good performance based on traditional metrics that are not adapted, such as accuracy. Two techniques were tested in addition to the use of metrics adapted to the imbalanced dependent variable (F1, precision, and recall scores): the lowering of the optimal alert threshold based on the F1 score, as done by Liu et al. (2022) with the ROC curve, and the use of the SMOTEENN algorithm, allowing the creation of artificial individuals of the minority class. Additional metrics oriented on the role of the EWS were also proposed, such as the persistence and continuity of warning signals, the timing of the first and last signals, and the number of identified crises. Despite their performance, machine and deep learning models are often criticized for their lack of explicability. We

addressed this problem by using the SHAP library in Python inspired by Shapley values from game theory to identify causal links between variables and to decompose the obtained predictions for each observation. Finally, our main contribution is the novel use of RNNs involving a memory vector that can retain long-term dependencies through a gate mechanism for the design of an EWS for currency crises. Thus, we contribute to the growing literature on machine and deep learning and improve reliability for policymakers of EWS for currency crises owing to neural networks adapted to sequential data, as expected by Tölö (2020).

The remainder of this paper is organized as follows. Section 2 presents the two main criteria for identifying currency crises and reviews the literature on EWS for financial crises. In Section 3, we describe all of the models developed in this study and the structure of the selected neural networks. Section 4 elaborates on the dataset used and the detailed methodology regarding the choice of alert window, data format, imbalanced target management, and the training and testing procedure. Section 5 presents our empirical results on the test sample running from 2015 to 2020 and proposes an out-of-sample example of the contributions of the variables to the prediction. Finally, Section 6 concludes the paper.

2 | LITERATURE REVIEW

The objective of this section is to present the evolution of the literature on currency crises, both in terms of crisis definition and modeling approaches. This in-depth study made it possible to define benchmark models and an initial set of macroeconomic, financial, and banking variables.

2.1 | Crisis criteria

To train a model to declare a warning signal, we must identify episodes of strong depreciation corresponding to currency crises. Various criteria have been proposed in the literature to define and identify currency collapses. There are two types of criteria for retrospectively dating currency crises: those based solely on the loss of value of the currency and those incorporating defense mechanisms against depreciation pressures.

The criterion proposed by Frankel and Rose (1996) is based only on the extent of currency value loss against the dollar. A collapse occurs when nominal depreciation reaches 25% or greater, which is at least 10% greater than the depreciation observed in the previous year.

$$\gamma_t = \frac{S_t}{S_{t-4}} - 1; \gamma_{t-4} = \frac{S_{t-4}}{S_{t-8}} - 1; \eta_t = \frac{\gamma_t}{\gamma_{t-4}} - 1$$

Frankel and Rose criterion : currency collapse occurs if $\gamma_t \geq 25\%$ and $\eta_t \geq 10\%$.

S_t is the average nominal exchange rate in quarter t of a given country, expressed as units of local currency per unit of a foreign currency (US dollar [USD]).

By using only the currency value loss and its speed, there is a desire to consider only the attacks that have worked and thus rule out intermediate periods of tension. Frankel and Rose indicated that the thresholds used were arbitrary, although they were supported by sensitivity analysis.

These criterion can be extended by incorporating the international reserves differential and Central Bank interest rate differential with the country against which the domestic currency is quoted, as is the case with the exchange market pressure (EMP) index proposed by Eichengreen et al. (1996), derived from the foreign EMP index of Girton and Roper (1976). The literature usually adopts a simplified version of this criterion as in (Kaminsky et al., 1998; Sachs et al., 1996):

$$EMP_t = w_1 \times \frac{\Delta S_t}{S_{t-1}} - w_2 \times \frac{\Delta R_t}{R_{t-1}} + w_3 \times \Delta i_t.$$

S_t is the nominal exchange rate expressed as units of local currency per unit of a foreign currency (USD), R_t is the amount of international reserves, i_t is the policy interest rate, and w_i is the weight assigned to each component.

A currency crisis occurs when the EMP is one to three standard deviations above its mean, according to the authors: one and a half for Eichengreen et al. ((1994), (1996)); two for Fratzscher and Bussire (2002) and Peltonen (2006), excluding the variation in the interest rate for the latter; and three for Kaminsky et al. (1998) and Berg et al. (2004). Generally, the weights assigned to each component are such that the conditional variances of each component are equal. An increase in the EMP reflects increased depreciation pressure on the domestic currency.

In this first currency crisis study, the Frankel and Rose criterion was chosen for its popularity, explicability, and simplicity of implementation, thereby removing the questions of weighting and the number of standard deviations to which the EMP index is very sensitive, as explained by Pontines and Siregar (2008). The 25% and 10% thresholds limit attention to successful attacks. The binary crisis variable resulting from this dating exercise

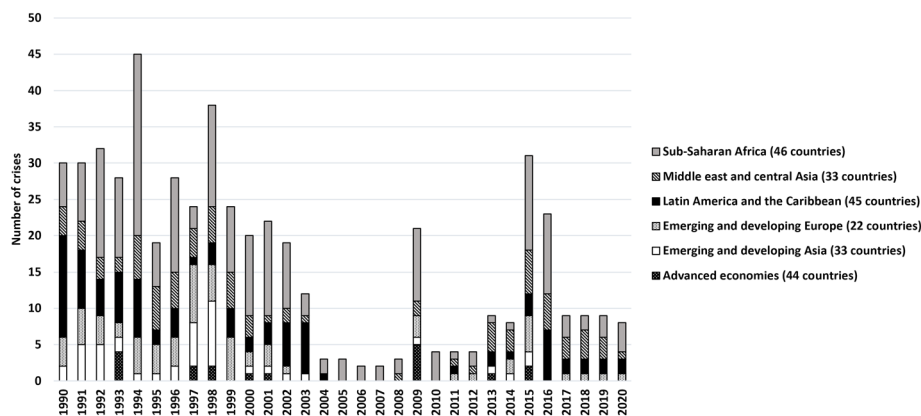


FIGURE 1 Currency crises (1990–2020) using the Frankel and Rose criterion.

made it possible to construct the dependent variable of the EWS, which is the signal before collapse.

Using a sample of 223 countries between 1990 and 2020 and the Frankel and Rose criterion, Figure 1 shows a general decline in the occurrence of crises since the early 2000s, although they have increased on average since the 2008 financial crisis. Sub-Saharan Africa remains the most affected region over the entire time horizon of the graph, which implies a geographic contagion effect or link to the level of development. Another explanation may be reduced severity over time, leading to fewer crises identified by this criterion.

2.2 | EWS for currency crises

The literature on currency crises has been extensively developed since Krugman (1979) and the severe currency tensions observed in the following decades. Initially, their occurrence was attributed to weak economic fundamentals, such as expansionary monetary and fiscal policies, which were incompatible with the maintenance of a fixed parity (Mundell's impossible trinity). However, several other theories have emerged, leading to the inclusion of expectations, banking, and financial indicators. EWSs are constructed in line with these considerations by including the indicators of these pre-crisis periods. We can distinguish two main categories of crisis warning methodology: The first is based on the individual monitoring of several key indicators, and the second is based on econometric, machine, and deep learning models to predict the probability of crisis occurrence.

The seminal paper of Kaminsky et al. (1998) [KLR] is part of the first approach, popularized as the “Signals” approach. The survey produced by these authors highlights a list of economic indicators that are considered to herald future currency collapses given their unusual behavior. In line with this methodology, KLR retained 16 indicators,² each of which was individually monitored.

A signal was sent when an indicator deviated from its “normal” level by crossing the alert threshold, defined as the value associated with a defined percentile of the indicator's distribution (between 10% and 20%). Their study confirmed that EWSs need a large and varied set of indicators, as the origins of crises can be multiple according to theory. Some variables have proven to be particularly successful, such as the real exchange rate, the banking crisis indicator, the stock index, and the ratio of M2 to international reserves. Although the authors succeeded in building a powerful set of indicators to detect the crises, their approach remains non-synthetic and does not lead to a unique and precise signal. Kaminsky (1999) considered this criticism by creating four composite indicators. One was the sum of the values taken by each of the monitored indicators (one if the variable has crossed its alert threshold and zero otherwise), weighted by the inverse of their noise-to-signal ratio.

Several studies have tested the out-of-sample performance of this procedure, with mixed results (Berg et al., 2004; Berg & Pattillo, 1999a, 1999b; Furman & Stiglitz, 1998; Mulder & Bussire, 1999). Berg and Pattillo (1999a) compared the performance of three EWS methodologies on the Asian crisis: the KLR approach, the probit model of Frankel and Rose (1996), and the cross-section country model of Sachs et al. (1996). In an attempt to replicate the KLR methodology as closely as possible, the results were inconclusive for the Asian crisis. While the predictive power remained weak, the approach allowed crises to be ranked in order of severity with some accuracy.

Gourinchas and Obstfeld (2011) constructed a fixed effects panel specification for the indicators identified in the literature. In this “event study,” each indicator was taken individually as a dependent variable of a model with four dummies as explanatory variables (dummies 1, 2, 3, and 4, respectively, for default crisis, currency crisis, systemic banking crisis, and 2008 crisis), taking the value one if a country has suffered a crisis and zero

otherwise. This specification made it possible to observe the behavior of economic and financial variables around crisis periods and the changes that herald future crises. However, in their study, this approach was not used to directly predict the probability of crises but to identify the variables to be included in a second model, which is panel logistic regression.

The Gourinchas and Obstfeld specification is therefore part of the second approach to predicting currency crises, based on the relationship between a dependent variable (taking the value one if a crisis has occurred in the following s periods and zero otherwise) and several dependent variables integrated simultaneously. Logit and probit regressions have gradually gained popularity (Berg & Pattillo, 1999a, 1999b; Chamon et al., 2012; Eichengreen et al., 1996; Frankel & Rose, 1996; Fratzscher & Bussire, 2002) due to their synthetic nature, the possibility of interpreting the results directly in terms of probability, and the study of the significance of the coefficients and their constancy over time and across countries. Their model has proven to be successful in predicting currency crises, especially for developed economies. A currency crisis was more likely to occur in the next 2 years when domestic credit rose significantly above its trend and when the currency appreciated in real terms. For emerging countries, a decline in international reserves was an important warning criterion, giving credence to the shielding policy of increasing these reserves.

However, logistic regression can only consider linear interactions between variables, which can lead to the exclusion of certain indicators that are crucial for prediction. In recent decades, the development of data mining methods has benefited currency crisis literature, and new models have been developed to bring nonlinearity to the interactions among variables. Thus, forecasting models based on decision trees have emerged.

Ghosh and Ghosh (2002) used a binary recursive tree to predict currency crises in 40 developed and emerging economies. This algorithm allowed them to consider the structural vulnerabilities of each country and accordingly adapt the critical thresholds of certain variables. They showed that fragile governance makes countries more vulnerable to corporate sector weaknesses and deteriorating macroeconomic indicators. They also identified the current account balance as the most important determinant of currency crises (top of the tree). Frankel and Wei (2004) set up a classification tree and showed that a high level of external debt does not necessarily lead to crises on its own, but it significantly increases the probability of a crisis if capital inflows are oriented toward the short term and are not used to build up reserves. However, these trees can sometimes be overly simplistic and generally suffer from overfitting. In their working paper on

forecasting periods of macroeconomic and financial stress in low-income economies, de Carvalho Filho et al. (2020) aggregated multiple decision trees into a RF to increase their predictive power.

Models based on algorithms that seek to reproduce biological mechanisms have also been tested for predicting financial crises. Nag and Mitra (1999) proposed an ANN to warn of upcoming currency tensions in Malaysia, Thailand, and Indonesia, three countries particularly impacted by the Asian crisis. For each of the three countries, a neural network including a recurrent mechanism (RNN) was built to capture dynamic features missed by traditional models. This study showed promising results; however, it was not until the late 2000s that neural networks gained popularity. In his study on 24 emerging countries, Peltonen (2006) managed to outperform logistic regression with an ANN (including lagged variables), as did Sarlin (2014) with his multilayer perceptron (MLP). To predict financial crises in BRICS countries (banking, currency, and sudden stop crises), Nik et al. (2016) also constructed a MLP, focusing on studying the importance and significance of the normalized variables by interpreting the associated weights. The private sector domestic credit variables had significant weights, as did the inflation rate, interest rate, and economic growth. In their study of currency, banking, and sovereign debt crises, Liu et al. (2022) compared logistic regression with seven machine and deep learning models, including both a RF and an ANN, for 165 countries from 1970 to 2017. Although the models appeared to perform well overall, the RF outperformed the logistic regression and neural network. The ANN outperformed logistic regression only for currency crises.

Predicting time series requires considering dynamic relationships between variables, but this feature is omitted in traditional currency crisis EWSs or artificially recreated through lagged variables. RNNs such as those of Nag and Mitra are specifically designed to accommodate sequential series by incorporating a memory mechanism through feedback loops. However, in its simplest version, the RNN is only able to retain the near past (see the vanishing gradient problem described in Appendix A.0.1) so that more sophisticated neural networks have been developed, such as LSTM and GRU. Two EWSs for the detection of systemic banking crises were built on the basis of a LSTM and a GRU for 17 developed countries by Tölö (2020) using logistic regression as a benchmark. The results showed that the predictions can be significantly improved with this type of recurrent network and became more robust. To the best of our knowledge, no other EWS dedicated to financial crises and, in particular, currency crises, is based on this type of model. If these models have been successful in forecasting financial asset

prices (Claveria et al., 2022; Dautel et al., 2020; Ranjit et al., 2018) and other sequence problems, such as video, text, and speech recognition (Wu et al., 2016), their contribution to currency crisis EWS remains to be demonstrated.

3 | MODEL DEVELOPMENT

The objective of this section is to present the different models tested and the selected structures for the neural networks.

3.1 | Logistic regression

Logistic regression is a model adapted to two-modality classification problems and is the predilection model of currency crisis EWS. It is one of the simplest models to set up and interpret because the interactions between the dependent and explanatory variables are linear. It predicts the probability of an event by optimizing the value of the coefficient assigned to each explanatory variable by maximizing the log-likelihood function of the sample. Thus, it is assumed that the probability of belonging to each of the two classes is a linear function of these input variables. During modeling, the bias and variance of the model should be reduced to improve the predictive quality, although they generally move in opposite directions. Adding a penalty parameter to the logistic regression can lead to better management of this trade-off, as is the case in ridge, lasso, and elastic net regressions. In this study, all of these regressions were modeled using the Sklearn library in Python with hyperparameter optimization based on the grid search algorithm. Logistic regression appeared to be more efficient than the other types of regression mentioned previously and was chosen as the benchmark because of its significant use in the literature.

3.2 | RF

RFs are among the most powerful machine learning models and can be used for both regression and classification problems. These forests are composed of multiple decision trees to reduce the prediction variance that result from a single tree. The decision tree is an iterative algorithm that, at each iteration, splits the sample observations into k groups according to a defined threshold value for a particular model variable. The first division is obtained by choosing the most informative explanatory variable with respect to the target. This split results in subpopulations corresponding to the first node of the

tree. The splitting process is then repeated several times for each subpopulation until the final nodes are reached. Each tree composing the forest is trained on a sub-sample of observations and variables from a bootstrap selection to reduce the overfitting risk. The forecasts of each tree are then aggregated using the bagging method, favoring performance generalization. In this study, the modeling was performed using the Sklearn library in Python.

3.3 | RNNs

ANNs are models originally inspired by the biological neural system by integrating statistical and optimization methods, allowing the machine to learn and make predictions. Two types of networks can be distinguished: feed-forward network and RNN. In this study, we are interested in RNNs whose functioning is adapted to the analysis of sequential data, including time series. Appendix A.0.1 presents the functioning of standard recurrent networks and the vanishing gradient problem which led us to resort to two types of more sophisticated recurrent networks: the LSTM (Hochreiter & Schmidhuber, 1997) and the GRU (Cho et al., 2014). Standard recurrent networks have evolved to include an updated memory vector owing to a gate mechanism that enables the retention of information for an extended time. The LSTM includes three gates (input gate, forget gate, and output gate) and the GRU includes two gates (update gate and reset gate), as it is a simplified version of the LSTM. The functioning of the LSTM and GRU neurons is detailed in the appendix mentioned above.

LSTM and GRU neural networks, which are close in structure, are particularly efficient in time series and textual analysis. In this study, these two types of networks were built using the Keras and TensorFlow libraries in Python. Six networks were built: two standard RNNs, the

TABLE 1 Hyperparameter space.

Hyperparameters	Values
Neurons	5–45, step = 5
Layers	1, 2
Epochs	50–1000, step = 10
Dropout	0–0.5, step = 0.05
Recurrent dropout	0–0.5, step = 0.05
Activation function	Tanh, Softsign
Recurrent activation function	Sigmoid, Softmax
Optimization objective	F1 score, precision, and recall
Early stopping	With (200), without
L2 regularization	0, 0.001, 0.01

first with one hidden layer and the second with two hidden layers, and two LSTMs and GRUs, again with one and two hidden layers respectively. The structures are presented in Figures A4–A6. Each neural network built has many-to-one form: The network receives a sequence of inputs and generates a unique output in the form of a probability transformed into an integer value according to a defined threshold. The hyperparameters were defined using the Bayesian optimization algorithm, testing combinations of hyperparameters available in Table 1 so that the objective function was optimized.

4 | DATA COLLECTION AND PREPROCESSING

4.1 | Signal window

In this study, we aim to warn of the possible occurrence of a crisis in a given window and not to identify its precise timing. To do so, we defined a warning window of eight quarters before the collapse of a currency, during which a continuous signal must be sent by our models in the form of a probability transformed into a binary variable according to a defined threshold (0.5 in the standard case). Thus, each quarter corresponds to a tranquil period (dependent variable equal to zero) or to an alert period corresponding to the pre-crisis period (dependent variable equal to one). If the model sends a signal, a crisis is expected to occur in the next one to eight quarters. The 2-year warning window is a standard in the literature and is found in KLR and Berg and Pattillo (1999b).

No treatment of post-crisis bias has been performed, such as suppressing crises and subsequent observations over several years after the currency collapse (Demirci-Kunt & Detragiache, 1998; Gourinchas & Obstfeld, 2011) or assigning a particular value to these observations and transforming the target into a three-modality variable (Fratzschler & Bussire, 2002). When no treatment is assigned to these post-crisis observations, they are treated as observations of tranquil periods during which economic indicators are healthy and sustainable, despite that they are in a period of adjustment toward a “normal” level. As a result, these periods can be wrongly considered indicators of future crises. We chose not to delete any of the observations to avoid breaking the order of the sequences and altering the memory mechanism of the tested neural networks. In this study, the notions of crisis episodes and currency crises (generally comprising several episodes) were confounded, but the switch to a target with more than two modalities could constitute a source of improvement for the continuation of this work.

4.2 | Data collection

The study is based on quarterly data from Q1 1995 to Q4 2020 for 68 currencies quoted against the USD as the base currency, including currencies of developed, emerging and developing economies (complete list available in Table B1). The dataset was derived from Datastream, IMF International Financial Statistics, and the Global Economy.

In line with the previously presented literature, particularly the KLR paper, current account, capital account, international, monetary, real, fiscal, and development variables were incorporated into our models. A first attempt to materialize the contagion phenomenon was also tested through six dummies for the six major world regions according to the IMF World Economic Outlook,³ taking a value of one if a crisis occurred in one of these regions during the last eight quarters and zero otherwise. The full list of variables tested is presented in Table B2 and descriptive statistics for non-categorical variables in Table B3.

KLR identified indicators for which the particular dynamics in the pre-crisis period herald a collapse in the value of a currency. The early nature of these indicators makes it possible to partially remove the question of endogeneity, assuming that their characteristic dynamics are sufficiently early for them to be considered determinants of crises and not consequences. However, the criteria for identifying crises are such that it is sometimes necessary to wait several months before the associated extreme thresholds are reached, even though the consequences are already materializing.

Depending on the model, the final set of indicators differed in line with the results of the variable selection algorithms (recursive feature elimination, mean decrease impurity, and permutation feature importance). For the regressions, the full set of variables was used; for the RF and neural network models, an identical subset of variables, including nominal and real exchange rate growth rates, real exchange rate deviation from its trend, country fixed effect, inflation rate, inflation differential with the United States, CPI volatility, M2 growth rate, M2 to international reserves, and public debt to GDP, performed better. From the variables included in the dataset, we expected a characteristic dynamic in the pre-crisis period, the time pattern of which may differ according to the indicators. Therefore, it was necessary to include several lags for each variable. However, this can lead to collinearity problems in logistic regression models, contrary to RNNs, for which these multiple lags are the basis of the performance. Thus, a time window of seven quarters (current and past six quarters),⁴ selected on the basis of F1 score performance, was used for the 11 variables

included in the RNNs. Figure B1 presents the empirical regularities in the pre- and post-crisis periods of the non-categorical variables retained in the neural networks and RFs.

4.3 | Data preprocessing

4.3.1 | Feature scaling

Prior to modeling, the dataset was standardized (values are centered around the mean with a unit standard deviation) so that all variables were on the same scale. Standardization generally increases the performance of models, especially when they are highly sensitive to differences in scale across variables. For neural networks, standardization makes it possible to bring the data into ranges more comparable with those of the activation functions ($[0:1]$, $[-1:1]$), facilitating the achievement of a zero value by the gradient during the training phase. To maintain a real independence between the sample on which each model was trained and the one on which the final performance was measured, the standardization was built on the training data only and then applied to the test sample.

4.3.2 | Imbalanced classification target management

Owing to the scarcity of currency crises, the dependent variable in this study contained few observations taking the value of one. The imbalanced nature of the target to be predicted may imply learning difficulties, especially for machine learning models that tend to neglect the minority class. Predicting warning signals rather than crises directly made it possible to partially compensate for this drawback: Crises represent 2.7% of the global sample, while warning signals correspond to 9.6%. However, this was still insufficient; therefore, two techniques have been considered, supplemented by an adapted performance measurement criterion (F1 score). On the one hand, the SMOTEENN algorithm allows to act upstream of the modeling by creating artificial individuals of the minority class and removing the least characteristic individuals of each class using the K -nearest neighbors algorithm. This method can only be used with models that do not explicitly include the sequential nature of the series as a parameter because the artificial individuals created cannot be replaced in the order of the sequence. Thus, this method has only been used in regressions and RFs. On the other hand, the lowering of the critical threshold for transforming probabilities into signals (by default

fixed at 0.5) allows to artificially increase the proportion of predicted signals. A lower threshold indicates greater attention to crises by policymakers.

4.3.3 | Training and validation procedure

The expanding window method, which is an adaptation of cross-validation to time series, was used in this study. The sample was divided into a training sample starting at the first available observation up to a certain date and a validation sample comprising the following observations up to a given date. This prevents data leakage, that is, the use of future data to predict the past. As with cross-validation, the operation was repeated several times, with the difference that the number of observations contained in the training window increased with each new training, incorporating more recent data. The validation window maintained a fixed size but slid toward the most recent observations as the training sample grown. Thus, in the last training iteration, the training sample included all observations except the most recent ones. To guarantee real independence between model learning and performance evaluation and limit overfitting, the validation and test samples correspond to different observations, as recommended by Sarlin (2014). The training phase comprising the training and validation samples was based on data from Q1 1995 to Q4 2014, respecting the split shown in Figure 2 for each of the four training iterations. The hyperparameters and structure of the models were optimized according to the average performance obtained from the validation samples. The test sample, which runs from Q1 2015 to Q4 2020, allowed the final performance of each model to be evaluated without it having been used during training.

5 | EXPERIMENTAL EVALUATION

5.1 | Performance criteria

To compare the performance of each model on this classification problem, performance metrics like precision, recall, F1 score, and area under the ROC curve (AUC) were used. While the accuracy is traditionally a decisive criterion, in the presence of an imbalanced target it can lead to high performance even though few good warning signals are predicted by the models. The three criteria of precision, recall, and F1 score have the advantage of considering this particularity by basing their construction on the confusion matrix (Table 2). The F1 score offers the advantage of combining the precision and recall scores,

FIGURE 2 Training and testing procedure.

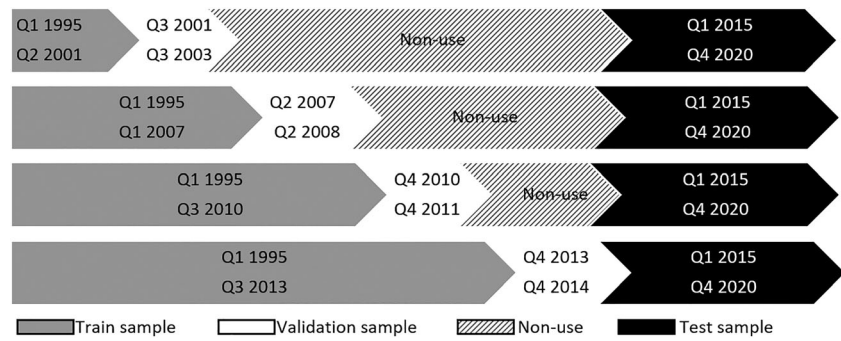


TABLE 2 Confusion matrix.

		True class	
		Crisis (within eight quarters)	No crisis (within eight quarters)
Predicted class	Signal was sent	True positive (TP)	False positive (FP)
	No signal was sent	False negative (FN)	True negative (TN)

Note: TP: number of crisis signals rightly sent. TN: number of quarters correctly identified as tranquil. FN: number of missed crisis signals. FP: number of crisis signals sent during tranquil periods.

so it is the metric chosen for comparing the performance of models, increasing with prediction quality.

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN};$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}.$$

5.2 | Results

5.2.1 | Global performances

For each model (logistic regression, RF, standard RNN, LSTM, and GRU), three specifications were tested on the panel dataset. For logistic regression and RF, the standard specification is compared with two specifications that consider the imbalanced character of the target: one by lowering the probability threshold of assignment to class one and the other using the SMOTEENN algorithm. For neural networks, a specification with one hidden layer is compared with an identical specification with a lower probability threshold and a specification with two hidden layers.

Metrics

Information on the performance of each model on the test sample from Q1 2015 to Q4 2020 is presented in

Table 3. The performance of all models is acceptable, with F1 scores above 0.70, except for the RF (1). Sophisticated neural networks⁵ appear to perform significantly better, with F1 scores above 0.80 for the single-hidden layer (LSTM (1) and GRU (1)). The best version of the standard model in the literature, namely, logistic regression (1), still appears to be a good candidate for the implementation of an EWS for currency crises with a F1 score of 0.73. However, of all of the signals sent, 77% were true (precision), and only 70% of true signals were identified (recall). For the one-hidden layer GRU (1), 83% of the signals sent were true, and 79% of the expected signals were identified, which represents a considerable margin of improvement. Recall is the second most decisive performance criterion if we consider that an undetected crisis represents a higher cost for policymakers than does a false signal. Based on the F1 score alone, sophisticated neural networks were systematically more efficient than logistic regressions, RFs, and standard RNNs, regardless of the number of hidden layers or the technique used for managing the imbalanced target. It should also be noted that neural networks and RFs performed on a more limited number of variables (11 variables) than logistic regressions (33 variables), which could allow the study to be more easily extended to other countries and over time.

The standard RNN shows acceptable performance, with a F1 score of 0.76 in its two-hidden layer specification (model (3)), which is higher than the scores of logistic regressions and RFs but still below LSTM and GRU,

TABLE 3 Metrics for test sample.

	F1	Precision	Recall	AUC
Logistic regression				
(1) Standard	0.73 (0.09)	0.77 (0.10)	0.70 (0.11)	0.84 (0.05)
(2) Probability threshold 45%	0.72 (0.09)	0.74 (0.10)	0.71 (0.11)	0.84 (0.05)
(3) SMOTEENN 20%	0.71 (0.09)	0.66 (0.10)	0.76 (0.10)	0.86 (0.05)
Random forest				
(1) Standard	0.59 (0.11)	0.86 (0.11)	0.45 (0.11)	0.72 (0.06)
(2) Probability threshold 30%	0.75 (0.08)	0.79 (0.10)	0.72 (0.11)	0.85 (0.05)
(3) SMOTEENN 40%	0.71 (0.09)	0.70 (0.11)	0.71 (0.10)	0.84 (0.05)
Standard RNN				
(1) 1 hidden layer	0.73 (0.09)	0.72 (0.10)	0.74 (0.10)	0.85 (0.05)
(2) 1 hidden layer + probability threshold 40%	0.73 (0.08)	0.69 (0.10)	0.78 (0.09)	0.87 (0.05)
(3) 2 hidden layers	0.76 (0.08)	0.77 (0.10)	0.75 (0.10)	0.86 (0.05)
LSTM				
(1) 1 hidden layer	0.81 (0.07)	0.86 (0.09)	0.76 (0.10)	0.87 (0.05)
(2) 1 hidden layer + probability threshold 40%	0.80 (0.08)	0.82 (0.09)	0.78 (0.10)	0.88 (0.05)
(3) 2 hidden layers	0.78 (0.08)	0.76 (0.10)	0.80 (0.09)	0.89 (0.05)
GRU				
(1) 1 hidden layer	0.81 (0.08)	0.83 (0.09)	0.79 (0.09)	0.89 (0.05)
(2) 1 hidden layer + probability threshold 35%	0.78 (0.08)	0.72 (0.09)	0.86 (0.07)	0.91 (0.04)
(3) 2 hidden layers	0.77 (0.07)	0.70 (0.10)	0.86 (0.08)	0.91 (0.04)

Note: (): estimation of the statistic's standard deviation using the bootstrap technique (1000 random draws of 200 observations).

regardless of the specifications chosen. The LSTM and GRU networks represent an improvement over standard RNNs in that they have both reduced noise and increased the number of good signals emitted, demonstrating the contribution of the gate mechanism in the network cells. It should also be noted that for LSTM and GRU, the best performance was obtained owing to a specification with only one hidden layer against two for the best standard RNN. The more complex mechanism governing the LSTM and GRU cells allowed for the use of fewer deep networks, facilitating learning.

With respect to the imbalanced target to be predicted, only the performance of the RF was affected. In its standard specification (1), the RF failed to integrate this characteristic, leading to poor performance (F1 score = 0.59). The use of the SMOTEENN algorithm with a ratio of the number of observations in the minority class to the number of observations in the majority class reaching 40% after resampling (model (3)) significantly improved the predictive power of the model, but lower the probability threshold to 0.30 proved to be more effective (model (2)). Logistic regression appeared insensitive to this target characteristic, showing comparable results regardless of the model. Single-hidden layer neural networks were also

insensitive to this particularity because they had already taken it into account during the optimization phase of the parameters, aiming at maximizing the F1 score or the precision for a defined recall score.

Identified crises

The final objective was to identify the maximum number of crises, and the model that has predicted the greatest number of signals was not necessarily the one that has identified the greatest number of crises. For the best specification of each of the models on the basis of the F1 score, Table 4 shows the number of crises identified by each of them, that is, for which at least one signal was sent during the alert window. The results are presented with respect to the previously used country classification. The test sample contained 22 crises, with at least one in each of the six regions. The models performed well, missing a maximum of only four crises for the standard RNN. We see that the RF managed to identify 20 crises, similar to the single-hidden layer LSTM. The logistic regression identified 19 crises, similar to the single-hidden layer GRU. Sophisticated neural networks stood out in this study not by the number of identified crises but by the precision of the signals sent; the signals sent were more

TABLE 4 Identified crises in the test sample (Q1 2015–Q4 2020).

	Crises	Logit	RF	RNN	LSTM	GRU
Advanced Economies	2	2	2	1	2	2
Emerging and Developing Europe	5	5	5	5	5	5
Middle East and Central Asia	4	3	3	2	3	2
Emerging and Developing Asia	1	1	1	1	1	1
Latin America and the Caribbean	5	5	4	5	5	5
Sub-Saharan Africa	5	3	5	4	4	4
Total	22	19	20	18	20	19
Missed crisis		KGZ	KGZ	KGZ	KGZ	KGZ
		NAM	MEX	MYS	NAM	MYS
		ZAF		NAM		NAM
				SWE		

Note: The results are presented for the best specification of each type of model, based on the F1 score (standard logistic regression, random forest with a 30% probability threshold, standard RNN with two hidden layers, standard LSTM with one hidden layer, and standard GRU with one hidden layer).

reliable because they were less noisy with false signals and were more numerous during the alert window. The standard RNN, which had a F1 score of 0.76 for its two-hidden layer specification, is the poorest performing model in terms of identified crises. It has sent signals that were most often continuous over the entire alert window, but it seemed to adapt less well to the particularities of each currency.

Regardless of the model used, the collapse of the Som (Kyrgyzstan's currency) in Q4 2015 was not anticipated. Several explanations can be proposed for this phenomenon. On one hand, the year-on-year depreciation of the Som in Q4 2015, the first component of the Frankel and Rose criterion, was just one percentage point above the 25% threshold of the criterion. The second component was above the 10% threshold, although it remained relatively low compared with the other crises in the sample. On the other hand, the Q4 2015 crisis was the first observed for the Som in the available sample. We can assume that the unprecedented nature of the event for the currency studied and the fact that it was not very representative compared with the other crises in the sample may have misled the models. Using the SHAP library in Python (described in Section 5.2.2), it was possible to measure the contribution of the explanatory variables to the predicted probability by the neural networks. The country effect variable showed a negative contribution (decrease the predicted probability) in relation to the infrequency of crises in the country. M2 growth rate and the ratio of M2 to international reserves, considered sustainable by the networks, largely compensated for the deterioration of the other monetary variables (inflation rate, inflation differential with the US and CPI volatility) and a slight overvaluation of the real exchange rate. If we

compare the neural networks, the missed crises are globally the same because of their similar functioning (KGZ and NAM are common to the three networks, and MYS is common to the GRU and standard RNN). Only the standard RNN was unable to alert to the drop in the Swedish krona in Q1 2015. The main reason for this was the granting of a large weight to the country effect variable, limiting the probability of a crisis occurring because the country had never experienced one in the past.

In their study, KLR used different sample sizes depending on the availability of data for each of the leading indicators. While, on average, their sample contained 61 crises as opposed to 22 in our testing sample, it should be noted that their study was not conducted out of sample as was the case here. The best performing models in terms of detected crises, the RF with a probability threshold set at 30%, and the standard LSTM with one hidden layer managed to identify 91% of the crises in the sample, while KLR identified an average of 70%.

Robustness checks

During the training phase, each model was tested on four different validation samples to identify the combination of hyperparameters that performed best on average and promote the performance generalization on unseen data. Neural networks offered more stable performance, displaying reduced box plots compared with the other models in Figure C1, indicating a superior generalization capacity regardless of the period used.

Table C1 compares our results with those of Berg and Pattillo (1999b) and Peltonen (2006). We compared the performance of our benchmark model (logistic regression) and our best LSTM and GRU specifications to the

KLR signal model reconstructed by Berg and Pattillo as well as their probit model. As RNNs are absent from the related literature, we compared our results with those of Peltonen's ANN. The differences in results should be interpreted with caution because the estimation sample (including temporality and countries), the criterion for identifying crises, the duration of the alert window, the input variables, and the probability threshold may differ across studies. Our out-of-sample results appear stronger than those in the existing literature, both for logistic regression and neural networks, outperforming the KLR model.

5.2.2 | LSTM and GRU: Detailed performances

EWS metrics

In this study, neural networks appear to be the best candidates for the implementation of an EWS for currency crises. However, the performance criteria used above are not sufficient on their own to identify the best EWS. Additional measures specific to EWSs, such as the number of identified crises, signal continuity, and timing of the first and last signals, may prove decisive.

Information on the performance of the single and two-hidden layer networks on the test sample is presented in Table 5. The objective of an EWS is to send as many good warning signals as possible. In this sense, the GRU with two hidden layers appears to be the best

performing, with 118 good signals identified out of the 138 expected. However, its F1 score was penalized by the 51 false signals sent, which increases the uncertainty of policymakers concerning the reliability of the model predictions. LSTM with one hidden layer is more reliable because only 17 false signals were sent, 10 of which were attributed to post-crisis bias. All models in this study were affected by this bias, which could be improved in future studies by switching to a three-modality target after an in-depth study on the duration of this bias.

The ultimate objective was to succeed in identifying each crisis, that is, sending at least one signal before each of them. Each model managed to warn of 20 crises of the 22 crises in the test sample, excepting the GRU with one hidden layer, which missed three crises. The latter appeared to be the best performing model based on the precision, recall and F1 score. For all models, the collapse of the Som (Kyrgyzstan) against the dollar in Q4 2015 was not anticipated. More complexity (i.e., adding an additional hidden layer) allowed us to anticipate the fall of the Namibian dollar against the USD in Q4 2015.

Policymakers can take preventive action if they have confidence in the EWS, and signal continuity over the entire warning window can strengthen it. In this study, persistence or continuity is the average number of good signals emitted compared with the number of expected signals for each alert window. Thus, a continuity indicator close to 100% tells policymakers that long signals should be watched carefully (under the condition that the noise metric is low) and that the first signal issued

TABLE 5 EWS metrics on the test sample.

	LSTM		GRU	
	1 hidden layer	2 hidden layers	1 hidden layer	2 hidden layers
True signals to be identified	138	138	138	138
True signals sent by the model	105	110	109	118
Missed true signals by the model	33	28	29	20
False signals sent by the model	17	34	22	51
<i>of which related to post-crisis bias</i>	10	22	7	27
Number of crises to be identified	22	22	22	22
Number of crises identified by the model	20	20	19	20
Missed crises (no signal sent)	KGZ, NAM	KGZ, SWE	KGZ, NAM, MYS	KGZ, SWE
Average % true signals sent/expected signals	69%	72%	73%	78%
Average % continuous true signals sent/expected signals	68%	71%	71%	78%
Average number of quarters of delay for the first signal sent	1	1	0	0
Average number of quarters in advance for the last signal sent	1	0	0	0

precisely suggests a crisis coming in eight quarters, completing the timing indicator. For all crises detected by the two-hidden layer GRU, the signals sent represented, on average, 78% of the alert window and were continuous over 78% of the alert window. GRUs also appear to be more accurate than LSTMs because the first signal for each alert window was, on average, sent in the first quarter of the window (no delay) and the last one at the end of the window.⁶ However, the two-hidden layer GRU sent signals beyond the window because it was very sensitive to post-crisis bias.

The two-hidden layer GRU appears to perform better here, with a higher proportion of good-predicted signals, and showed some signal continuity. However, the signals sent were also noisier and altered by post-crisis bias. LSTMs were less sensitive to noise but slightly less accurate over the warning window. Finally, the single-hidden layer GRU is the least efficient, missing three crises, while the others missed two.

SHAP values

As is the case for traditional models, it was possible to assess the overall importance of the variables (Figures D1 and D2), the direction of the correlations (Figures D3 and D4), and identify, for a specific signal, the variables that have contributed to the predicted probability by measuring the impact of past observations (Tables D1 and D2). This has been done using SHapley Additive exPlanation (Lundberg & Lee, 2017), also known as SHAP value, inspired by game theory (Shapley, 1953). Deep learning models, although powerful, are at first sight more difficult to explain because of the large number of parameters involved as well as the stacking of several hidden layers. The use of SHAP values allowed us to compensate for this weakness by precisely identifying the causes of crises, which could allow policymakers to implement adapted policies. The base value corresponds to the average payoff deduced from the set of possible combinations of players in the framework of comparative game theory. Similarly, for econometric, machine, or deep learning models, the base value corresponds to the average predicted probability by the model for the entire sample, based on the average contributions of each explanatory variable. Therefore, the SHAP value associated with a variable for a given observation is the difference between the contribution of this variable to the prediction associated with this observation and its average contribution to all predictions. The predicted probability by the model for a given observation is the sum of the base value and SHAP values of all explanatory variables for that observation.

Figures D1 and D2 show the importance of the variables, and Figures D3 and D4 show the direction of the correlations with the target for the LSTM and GRU with

one hidden layer. The blue dots correspond to indicators with low values, and the red dots correspond to indicators with high values. The *x*-axis indicates the sign and magnitude of the contribution to the model-predicted probability. Negative contributions decrease the probability of a crisis. The variables are ranked in order of importance. The top three most important variables were identical for both models, including the deviation of the real exchange rate from its trend, real exchange rate growth rate, and the country fixed effect. The deviation of the real exchange rate often appears to be a good indicator of future crises, as shown by Kaminsky et al. (1998) and Gourinchas and Obstfeld (2011). The overvaluation of a currency relative to its trend with the USD can increase the vulnerability of an economy by reducing the competitiveness of its exports and maintaining a current account deficit, which can be decisive in emerging and developing economies. The real exchange rate growth rate is positively correlated with the warning signals, meaning that the currencies under study were gradually losing value before their collapses, probably because of a sharp rise in inflation and persistent speculative attacks. Finally, the appearance of the country effect in the top three implies structural fragilities, leading to history repeating itself.

Tables D1 and D2 present the contributions of present and past variables (up to six quarters back) to the sending of a warning signal in Q3 2017, 1 year before the collapse of the Turkish lira in Q3 2018. For the LSTM, the base value was 0.074, and the predicted probability for the third quarter of 2017 was 0.839. For the GRU, the base value was 0.072, and the predicted probability was 0.745. The sum of the contributions of the variables over the seven quarters and the base value corresponds to the model prediction. The variables in red increased the predicted probability, whereas those in blue decreased it. For both LSTM and GRU, all indicators seemed to deteriorate, increasing the probability of a signal above the alert threshold. The sending of a warning signal was particularly influenced by the real exchange rate growth rate, public debt to GDP ratio, and country effect. Monetary variables such as CPI volatility, M2 growth rate, inflation rate, and inflation differential with the United States reinforced the sending of a warning signal with greater weight in more recent quarters. The Turkish lira had already suffered currency crises in the past (2001 and 2015) in addition to other episodes of extreme depreciation that did not fully satisfy the Frankel and Rose criterion (the 2008 financial crisis and post-crisis period). Therefore, the evolution of the USD/TRY was historically destabilizing for the country, making the lira a vulnerable currency due to a lack of confidence. The real exchange rate depreciation fluctuated during the alert window but

remained positive, reaching ranges comparable with past crises and tensions. The ratio of public debt to GDP has continued to decrease since the Global Financial Crisis, reaching about 27% in 2015, compared with over 45% in 2008. At the beginning of the warning window, public debt was thus relatively low, but the trend has reversed, and it started to increase because of the impact of depreciation on external debt (representing almost 40% of public debt). Finally, in an open economy such as Turkey's, depreciation can represent gains in competitiveness but can also lead to an increase in prices through imported inflation (pass-through effect), which occurred in early 2017. Therefore, the crisis of 2018 was the result of the country's history and mistrust in the Turkish lira leading to indebtedness in foreign currency, which proved to be very risky in the face of an already depreciated currency reinforcing price increases.

6 | CONCLUSIONS AND FUTURE WORK

This study has investigated the implementation of an EWS for currency crises over a 2-year warning window for 68 countries, including developed, emerging and developing economies over the period of 1995–2020 and using the Frankel and Rose criterion. The objective was to evaluate the performance of sophisticated RNNs, representing the state of the art for financial asset price prediction, compared with traditional models, such as logistic regressions and RFs. The advantage of these models lies in a gate mechanism that allows the memory vector to retain information over an extended period and to distinguish between short and long-term dependencies. Thus, the LSTM and GRU models have been tested for the first time for the implementation of an EWS for currency crises and compared with the traditional models of the associated literature, which are still widely used.

For all of the models tested, the performances were acceptable, meaning that the calibration on past data is a good way to forecast the future in the case of currency crises. LSTM and GRU outperformed traditional models and standard RNNs, in line with the growing literature on the use of machine and deep learning for predicting financial crises. It should be noted that the imbalanced specificity was integrated by neural networks during the parameter optimization phase and, therefore, did not reduce the predictive power of the models. The best LSTM and GRU built offered similar performances and allowed for the warning of 20 of the 22 crises of the test sample. The traditional models built also performed well in terms of the number of detected crises and showed

performances superior to those in the past literature. This can be explained by the availability of data over a longer history and for a larger number of countries, improving the training of the models. However, neural networks outperformed these traditional models in terms of precision and continuity of signals throughout the alert window and were characterized by less noisy signals (representing 14%, 17%, and 23% of the emitted signals for LSTM, GRU, and logistic regression, respectively), making them more reliable EWSs for policymakers.

All models were sensitive to post-crisis bias, concentrating false signals after the collapse of a currency due to the continued deterioration of economic and financial indicators. The treatment of this bias could be improved in future studies by conducting an appropriate study on the duration of this bias and switching to a three-modality target.

As with traditional models, it was possible to explain precisely how neural networks led to a prediction and to identify the contributions of past variables to the current prediction owing to SHAP decomposition. Each of the constructed networks was based on a limited number of variables from the initial set, including only variables related to the real and nominal exchange rates and monetary, fiscal, and development variables. The deviation of the real exchange rate from its trend appeared here as the most decisive variable, which is in line with the literature presented in this paper. The nonlinear nature of the operations within the neural network cells also allowed us to obtain warning signals motivated occasionally by variables that had less importance for the whole sample, as in the example of Turkey. This made it possible to detect crises with various origins. The accurate results of our two neural networks, in addition to being superior to those of traditional models, were also interpretable, which made it possible to identify the causes and vulnerabilities leading to crises.

Finally, one may wonder whether the valuable results obtained can be generalized to all types of financial crises. This could be the subject of comparative studies and the implementation of new types of neural networks, as the associated literature is constantly growing.

ACKNOWLEDGMENTS

We would like to thank the participants of the Forecasting Financial Markets (2022, Rimini Center) and the IMAC workshop (2022, Rennes) for their helpful remarks.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from [Datastream](#), [IMF International Financial Statistics](#), and [the Global Economy](#). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with permission from Datastream and the Global Economy.

ORCID

Virginie Gautier  <https://orcid.org/0000-0002-7077-789X>

ENDNOTES

- ¹ The European Monetary System's collapse in 1992–1993, the Tequila crisis following the collapse of the Mexican peso in 1994, and the Asian crisis in 1997.
- ² International reserves (USD), imports (USD), exports (USD), the terms of trade, deviations of the real exchange rate from trend (%), the differential between US and domestic real interest rates on deposits, “excess” real M1 balances, the money multiplier of M2, the ratio of domestic credit to GDP, the real interest rate on deposits, the ratio of nominal lending to deposit interest rates, the stock of commercial bank deposits (nominal terms), the ratio of M2 to gross international reserves, output, the index of equity prices (USD), and banking crises (dummy).
- ³ (1) Advanced Economies, (2) Emerging and Developing Europe, (3) Middle East and Central Asia, (4) Emerging and Developing Asia, (5) Latin America and the Caribbean, and (6) Sub-Saharan Africa.
- ⁴ We use seven quarters for the lagged variables included as explanatory variables in the neural networks and eight quarters for the alert window during which a continuous signal must be emitted by the model, as in the literature.
- ⁵ LSTM and GRU.
- ⁶ It does not guarantee that the signal was emitted continuously between the first and last emitted signals.

REFERENCES

- Berg, A., Borensztein, E., & Pattillo, C. (2004). Assessing early warning systems: How have they worked in practice?. (*IMF Working Papers No. 52*): International Monetary Fund. <https://www.imf.org/external/pubs/ft/wp/2004/wp0452.pdf>
- Berg, A., & Pattillo, C. (1999a). Are currency crises predictable? A test. *IMF Staff Papers*, 46(2), 107–138. <https://ideas.repec.org/a/pal/imfstp/v46y1999i2p1.html>
- Berg, A., & Pattillo, C. (1999b). Predicting currency crises: The indicators approach and an alternative. *Journal of International Money and Finance*, 18(4), 561–586. <https://ideas.repec.org/a/eee/jimfin/v18y1999i4p561-586.html>
- Calvo, G., & Reinhart, C. (2000). Fear of floating. (*NBER Working Papers No. 7993*): National Bureau of Economic Research, Inc. <https://ideas.repec.org/p/nbr/nberwo/7993.html>
- Chamon, M., Ghosh, A., & Kim, J. I. (2012). Are all emerging market crises alike? In Obstfeld, M., Cho, D., & Mason, A. (Eds.), *Global economic crisis* (pp. 228–249), Chapters: Edward Elgar Publishing. https://ideas.repec.org/h/elg/eechap/14951_10.html
- Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111), Association for Computational Linguistics. <https://aclanthology.org/W14-4012>
- Claveria, O., Monte, E., Soric, P., & Torra, S. (2022). An application of deep learning for exchange rate forecasting. (*AQR Working Papers No. 202201*): University of Barcelona, Regional Quantitative Analysis Group. <https://ideas.repec.org/p/aqr/wpaper/202201.html>
- Dautel, A. J., Hrdle, W. K., Lessmann, S., & Seow, H.-V. (2020). Forex exchange rate forecasting using deep recurrent neural networks. *Digital Finance*, 2(1), 69–96. https://ideas.repec.org/a/spr/digfin/v2y2020i1d10.1007_s42521-020-00019-x.html
- de Carvalho Filho, I. E., Weisfeld, H., Liu, F., Comelli, F., Presbitero, A. F., Meyer-Cirkel, A., Lizarazo Ruiz, S. V., Hellwig, K.-P., Giri, R., & Huang, C. (2020). Predicting macroeconomic and macrofinancial stress in low-income countries. (*IMF Working Papers No. 289*): International Monetary Fund. <https://ideas.repec.org/p/imf/imfwpa/2020-289.html>
- Demirci-Kunt, A., & Detragiache, E. (1998). The determinants of banking crises in developing and developed countries. *IMF Staff Papers*, 45(1), 81–109. <https://ideas.repec.org/a/pal/imfstp/v45y1998i1p81-109.html>
- Eichengreen, B., Rose, A. K., & Wyplosz, C. (1994). Speculative attacks on pegged exchange rates: An empirical exploration with special reference to the European Monetary System. (*NBER Working Papers No. 4898*): National Bureau of Economic Research, Inc. <https://ideas.repec.org/p/nbr/nberwo/4898.html>
- Eichengreen, B., Rose, A. K., & Wyplosz, C. (1996). Contagious currency crises. (*NBER Working Papers No. 5681*): National Bureau of Economic Research, Inc. <https://ideas.repec.org/p/nbr/nberwo/5681.html>
- Flood, R. P., & Garber, P. M. (1984). Collapsing exchange-rate regimes: Some linear examples. *Journal of International Economics*, 17(1–2), 1–13. <https://ideas.repec.org/a/eee/inecon/v17y1984i1-2p1-13.html>
- Frankel, J. A., & Rose, A. K. (1996). Currency crashes in emerging markets: an empirical treatment. (*International Finance Discussion Papers No. 534*): Board of Governors of the Federal Reserve System (USA). <https://ideas.repec.org/p/fip/fedgif/534.html>
- Frankel, J. A., & Wei, S.-J. (2004). Managing macroeconomic crises. (*NBER Working Papers No. 10907*): National Bureau of Economic Research, Inc. <https://ideas.repec.org/p/nbr/nberwo/10907.html>
- Fratzscher, M., & Bussire, M. (2002). Towards a new early warning system of financial crises. (*Working Paper Series No. 145*): European Central Bank. <https://ideas.repec.org/p/ecb/ecbwps/2002145.html>
- Furman, J., & Stiglitz, J. E. (1998). Economic crises: Evidence and insights from East Asia. *Brookings Papers on Economic Activity*, 29(2), 1–136. <https://ideas.repec.org/a/bin/bpeajo/v29y1998i1998-2p1-136.html>
- Ghosh, A. R., & Ghosh, S. R. (2002). Structural vulnerabilities and currency crises. (*IMF Working Papers No. 9*): International

- Monetary Fund. <https://ideas.repec.org/p/imf/imfwpa/2002-009.html>
- Girton, L., & Roper, D. E. (1976). A monetary model of exchange market pressure applied to the post-war Canadian experience. (*International Finance Discussion Papers No. 92*): Board of Governors of the Federal Reserve System (USA). <https://ideas.repec.org/p/fip/fedgif/92.html>
- Gourinchas, P.-O., & Obstfeld, M. (2011). Stories of the twentieth century for the twenty-first. (*NBER Working Papers No. 17252*): National Bureau of Economic Research, Inc. <https://ideas.repec.org/p/nbr/nberwo/17252.html>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kaminsky, G. (1999). Currency and banking crises: The early warnings of distress. (*IMF Working Papers No. 178*): International Monetary Fund. <https://ideas.repec.org/p/imf/imfwpa/1999-178.html>
- Kaminsky, G., Lizondo, S., & Reinhart, C. (1998). Leading indicators of currency crises. (*MPRA Paper No. 6981*): University Library of Munich, Germany. <https://ideas.repec.org/p/pramprapa/6981.html>
- Krugman, P. (1979). A model of balance-of-payments crises. *Journal of Money, Credit and Banking*, 11(3), 311–325. <https://ideas.repec.org/a/mcb/jmoncb/v11y1979i3p311-25.html>
- Liu, L., Chen, C., & Wang, B. (2022). Predicting financial crises with machine learning methods. *Journal of Forecasting*, 41(5), 871–910. <https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2840>
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Thirty-first conference on neural information processing systems*, Neural Information Processing Systems.
- Mulder, C. B., & Bussire, M. (1999). Political instability and economic vulnerability. (*IMF Working Papers No. 46*): International Monetary Fund. <https://ideas.repec.org/p/imf/imfwpa/1999-046.html>
- Nag, A., & Mitra, A. (1999). Neural networks and early warning indicators of currency crisis. *Reserve Bank of India Occasional Papers*, 20(2), 183–222. https://www.researchgate.net/profile/Ashok-Nag/publication/284480973_Neural_networks_and_early_warning_indicators_of_currency_crisis/links/59e048ada6fdcca98423759d/Neural-networks-and-early-warning-indicators-of-currency-crisis.pdf
- Nik, P. A., Jusoh, M., Shaari, A. H., & Sarndi, T. (2016). Predicting the probability of financial crisis in emerging countries using an early warning system: Artificial neural network. *Journal of Economic Cooperation and Development*, 37(1), 25–40. <https://jeed.sesric.org/pdf.php?file=ART15012802-2.pdf>
- Peltonen, T. A. (2006). Are emerging market currency crises predictable? A test. (*Working Paper Series No. 571*): European Central Bank. <https://ideas.repec.org/p/ecb/ecbwps/2006571.html>
- Pontines, V., & Siregar, R. (2008). Fundamental pitfalls of exchange market pressure-based approaches to identification of currency crises. *International Review of Economics & Finance*, 17(3), 345–365. <https://ideas.repec.org/a/eee/reveco/v17y2008i3p345-365.html>
- Ranjit, S., Shrestha, S., Subedi, S., & Shakya, S. (2018). Comparison of algorithms in foreign exchange rate prediction. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)* (pp. 9–13), IEEE.

- Sachs, J. D., Tornell, A., & Velasco, A. (1996). Financial crises in emerging markets: The lessons from 1995. *Brookings Papers on Economic Activity*, 27(1), 147–216. <https://ideas.repec.org/a/bin/bpeajo/v27y1996i1996-1p147-216.html>
- Sarlin, P. (2014). On biologically inspired predictions of the global financial crisis. *Neural Computing and Applications*, 24(3–4), 663–673. https://www.scs-europe.net/conf/ecms2012/ecms2012%20accepted%20papers/fes_ECMS_0065.pdf
- Shapley, L. (1953). A value for n-person games. In Kuhn, H. W., & Tucker, A. W. (Eds.), *Contributions to the theory of games II* (pp. 307–317), Princeton University Press.
- Tölö, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, 49(C), 100746. <https://ideas.repec.org/a/eee/finsta/v49y2020ics1572308920300243.html>
- Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. <https://arxiv.org/abs/1609.08144>

AUTHOR BIOGRAPHIES

Sylvain Barthélémy is the founder and CEO of Gwenlake, an associate professor at the University of Rennes, and a lecturer in HEC Paris. His research interests include machine learning, artificial intelligence, and statistical analysis in the fields of finance, economics, and life science.

Virginie Gautier is a PhD student at the University of Rennes and a member of the Center of Research in Economics and Administration (CREM). She is also a junior economist and data scientist at TAC Economics. Her research interests include macroeconomics, international economics, and machine learning.

Fabien Rondeau is an associate professor at the University of Rennes in the economics department, a member of the Center of Research in Economics and Administration (CREM), and a fellow of the Euro Area Business Cycle Network. His research interests include macroeconomics, international economics, trade and integration, and European Union.

How to cite this article: Barthélémy, S., Gautier, V., & Rondeau, F. (2024). Early warning system for currency crises using long short-term memory and gated recurrent unit neural networks. *Journal of Forecasting*, 1–28. <https://doi.org/10.1002/for.3069>

APPENDIX A: NEURAL NETWORKS

A.1 | Functioning

Neural networks consist of at least two layers of neurons (also called nodes): the input layer (explanatory variables) and output layer (prediction). Additional layers, called hidden layers, can also be added if the interactions between variables are more complex. Each node is connected to one or more nodes in the previous and next layers with associated weights and thresholds. The input variables are multiplied by their respective weights (a set of weights specific to each neuron) and then summed before passing through an activation function, which determines the output of the neuron and allows nonlinearity to be incorporated into the process. If this value is higher than the predefined threshold, the node is activated, and the information is transmitted to the nodes of the next layer. The mechanism is repeated as many times as there are layers, resulting in a prediction.

Two types of networks can be distinguished: feed-forward and recurrent neural networks. Feed-forward networks, as shown in Figure A1, are acyclic unlike RNNs. The first type is the simplest, as the information only circulates in one direction: from the input layer to the output layer, passing through the hidden layers if they exist.

RNNs, as shown in Figure A2, are designed to adapt to sequential data, owing to their feedback loops. The input variables and data transformed during the previous iterations in the network (also called the hidden state) allow us to obtain a prediction that consider the sequence order and related dependencies. In other words, they can retain the state from one iteration to the next and use their own output as one of the input for the next iteration. Therefore, the points are not treated independently but are seen as sequences by the model.

With each new passage through the network, the predicted value is compared with the expected value to calculate the prediction error. This error is minimized by modifying the network weights according to the importance of the contribution of the associated nodes (update proportional to the partial derivative of the error function with respect to the current weight). This procedure, called the backpropagation algorithm, is repeated until the loss function minimum is reached. To do this, an optimization algorithm such as the descending gradient is used, which is capable of identifying the minimum of any convex function by converging to it at a speed depending on its parameterization (learning rate).

However, standard RNNs, also called “vanilla,” have a limitation during the learning phase: the vanishing gradient problem. When the weights are extremely small (less than one), their successive multiplication at each time step causes the gradient to decrease exponentially until it reaches zero (or causes the gradient to explode if the weights are very large), causing the network to stop learning early. Indeed, the descent of the gradient allows the RNN parameters to be updated; and the more the gradient decreases, the more insignificant the updates become, contributing to retaining only the near past. For this reason, recurrent networks have evolved to include an updated memory vector owing to a gate mechanism that enables the retention of information for an extended time, as with the LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014). These networks reduce the risk of explosion or vanishing gradients by using an additive gradient combined with a gate mechanism.

All recurrent networks are built as a chain repeating an identical cell, the structure of which is quite simple for standard recurrent networks. In a standard RNN cell, two inputs are necessary to produce the output, as shown in Figure A3: the input variables corresponding to the

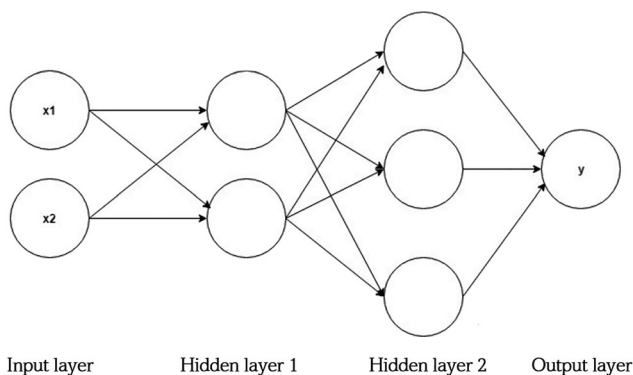


FIGURE A1 Fully connected feed-forward neural network.

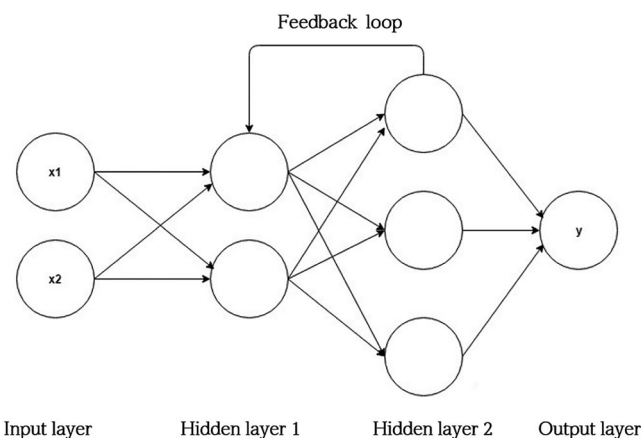


FIGURE A2 Recurrent neural network.

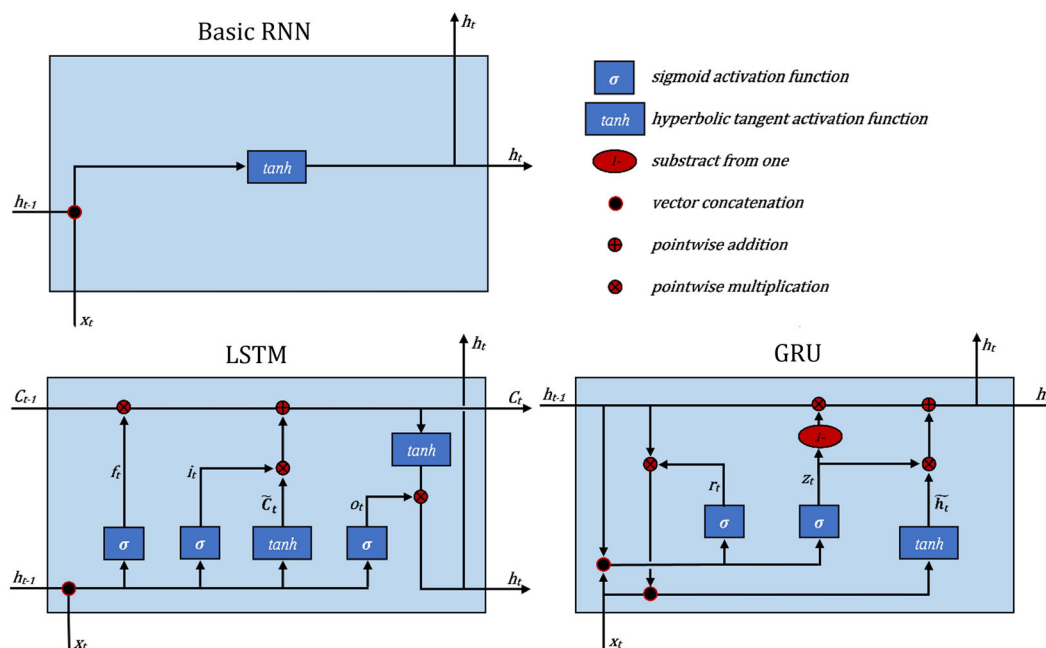


FIGURE A3 Standard RNN, LSTM, and GRU cells.

explanatory variables (x_t) and the previous hidden state (h_{t-1}) corresponding to the last output of the cell. These components are weighted and summed before passing through the activation function (\tanh). The process is slightly more complex in the LSTM and GRU cells. For the LSTM, three inputs are needed to perform the prediction: again the observed explanatory variables (x_t) and the output of the cell from the previous time step (h_{t-1}), the latter corresponding to the short-term memory, and the long-term memory (C_t), which contains the long-term dependencies between the model variables. This long-term memory is updated by the gate mechanism. This mechanism is also found in GRU, which is a simplified version of LSTM.

The LSTM and GRU cells attempt to keep track of all past information that has passed through the network while forgetting irrelevant information, thanks to a memory vector called the cell state. The contribution of LSTMs and other sophisticated recurrent networks lies in this cell state, which passes through all cells of the network while undergoing only minor transformations (horizontal line running through the top of the LSTM and GRU diagrams in Figure A3). Thus, the information relative to the more distant time step can easily be stored in memory in an unchanged form. Information can be added or deleted from this memory vector owing to gate regulation. LSTM has three gates: forget gate, input gate, and output gate. The forget gate decides which information should be deleted from the long-term memory used in the previous time step (C_{t-1}) based on the new

information available (x_t and h_{t-1}). The input gate identifies which new relevant information from the explanatory variables (x_t) and short-term memory (h_{t-1}) should be stored in the long-term memory (C_t). Through these two gates, the cell state (C_{t-1}) is updated (C_t). Finally, the output gate filters the information available (x_t, h_{t-1}, C_t) to create the output of the cell (h_t), also used as short-term memory in the next time step. The GRU is a compromise between a standard RNN and a LSTM because it limits the risk of exponential gradient decay owing to a gate mechanism but contains only one memory vector. The GRU includes two gates: the reset gate and the update gate. First, the reset gate determines how much of the memory produced in the previous time step (h_{t-1}) should be combined with new variables (x_t) to provide a new hidden state (h'_t). The update gate retains information relevant to the current and future predictions from the previous hidden state (h_{t-1}) and the new state proposed by the reset gate (h'_t), creating the final hidden state (h_t). Thus, long-term dependencies can be retained by the model if the update gate decides to retain a significant proportion of the previous hidden state (h_{t-1}). The performance of these networks is based on the efficiency of the filters applied to the data and associated weights. During the learning phase, the optimal weights are defined using backpropagation to improve prediction. The deletion and addition of information in the cell state and the cell output are the result of successive filtering processes performed by the activation function(s) of each gate.

A.2 | Architectures

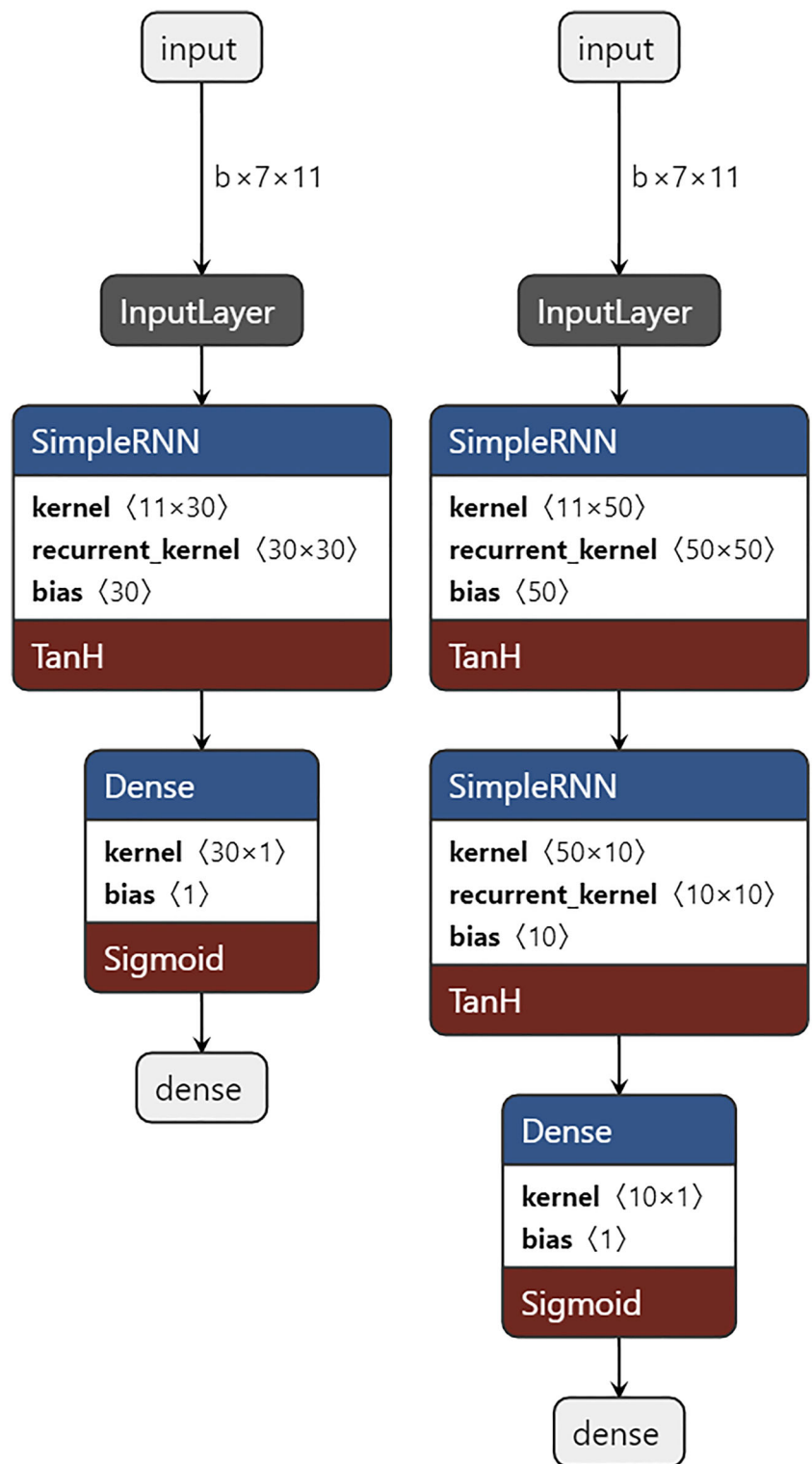


FIGURE A4 Standard RNN proposed structures.

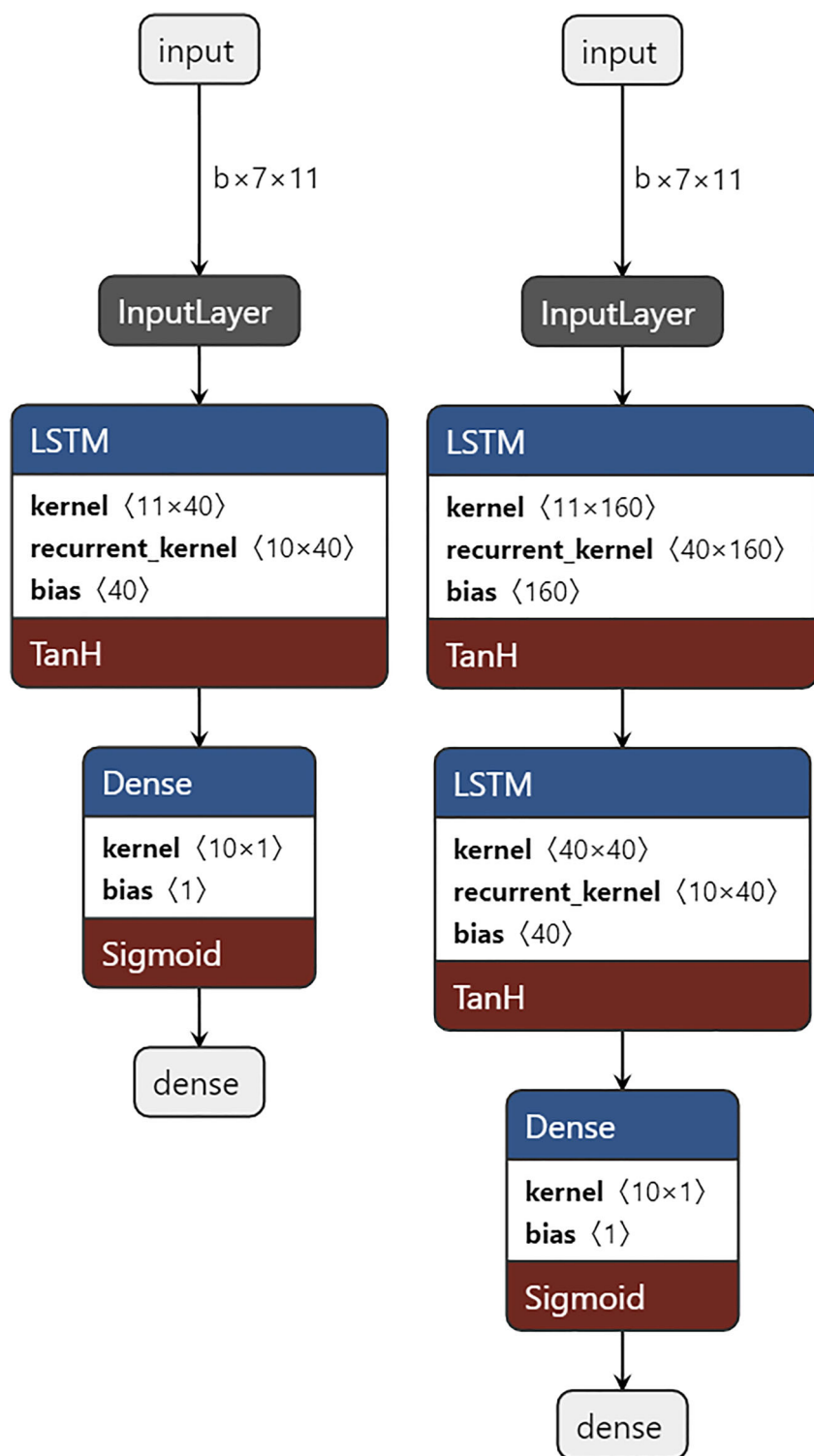
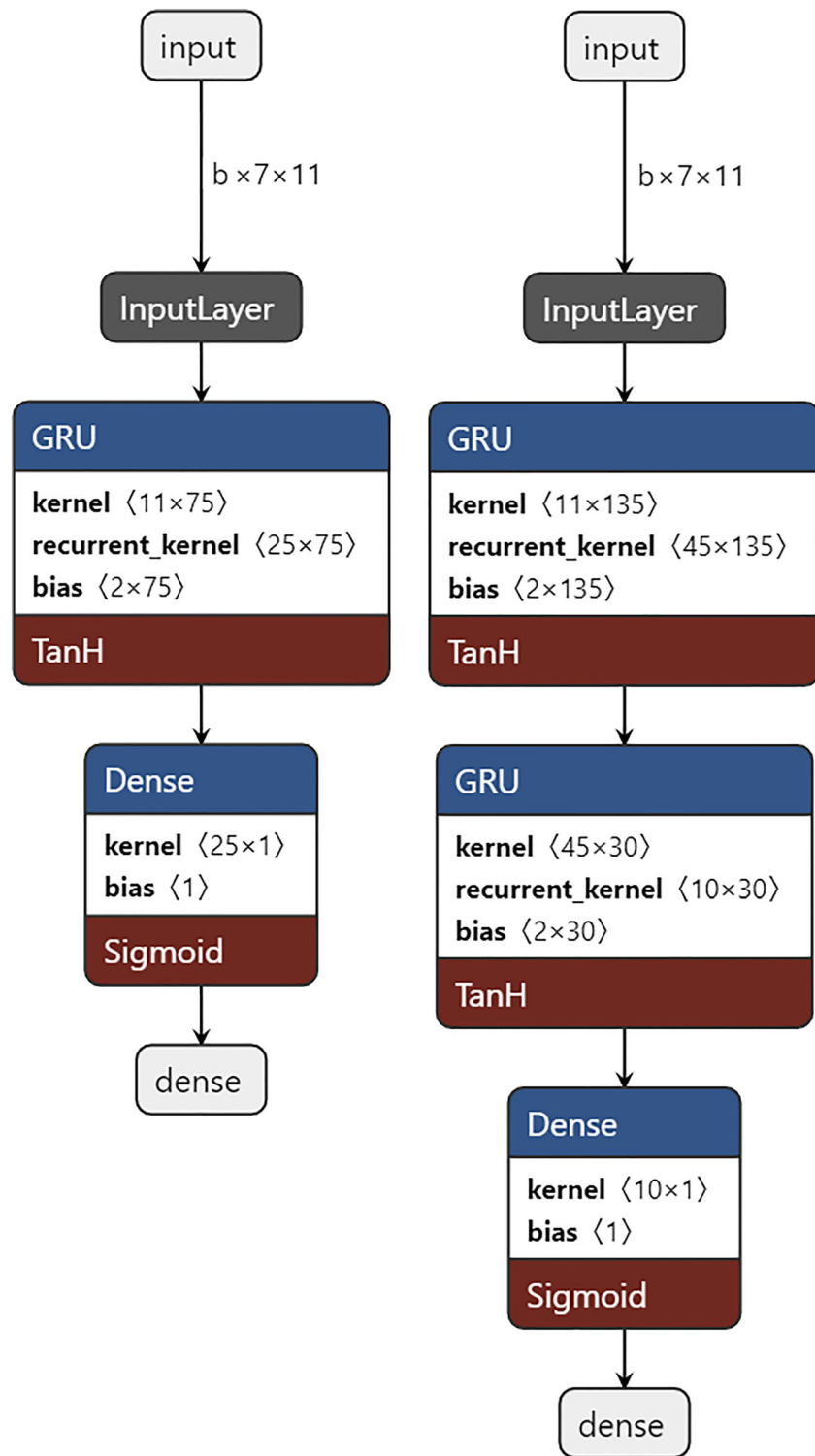


FIGURE A5 LSTM proposed structures.

FIGURE A6 GRU proposed structures.



APPENDIX B: DATABASE

B.1 | Composition

TABLE B1 IMF WEO country classification.

IMF World Economic Outlook regions	Countries
Advanced Economies	Australia, Canada, Switzerland, Czech Republic, Germany, Denmark, Eurozone,* Finland, France, United Kingdom, Iceland, Japan, South Korea, Netherlands, New Zealand, Singapore, Sweden, United States, Norway
Emerging and Developing Europe	Albania, Bulgaria, Belarus, Hungary, Moldova, Northern Macedonia, Poland, Romania, Russia, Serbia, Turkey, Ukraine
Middle East and Central Asia	Armenia, Egypt, Georgia, Jordan, Kazakhstan, Kyrgyzstan, Kuwait, Morocco, Tunisia, Israel
Emerging and Developing Asia	China, Hong Kong, Indonesia, India, Malaysia, Philippines, Thailand
Latin America and the Caribbean	Argentina, Brazil, Chile, Colombia, Dominican Republic, Guatemala, Jamaica, Mexico, Peru, Paraguay
Sub-Saharan Africa	Botswana, Cape Verde, Ghana, Kenya, Mauritius, Namibia, Nigeria, Rwanda, Uganda, South Africa

*Includes Germany, Austria, Belgium, Spain, Finland, France, Greece, Ireland, Italy, Portugal, Netherlands, Cyprus, Estonia, Latvia, Lithuania, Malta, Luxembourg, Slovenia, and Slovak Republic since 1999.

TABLE B2 Initial set of variables.

	Variables
Current account	Exports to GDP, imports to GDP, terms of trade, current account to GDP, nominal and real exchange rate level and growth, exchange rate volatility, real exchange rate deviation from its trend, international reserves growth
Capital account	Interest rate, interest rate growth
International	Differential between domestic and US real GDP growth, inflation and interest rate
Monetary	Inflation, CPI volatility, domestic credit to the private sector to GDP, M2 growth, M2 to international reserves
Real	Real GDP growth, output gap
Fiscal	Public debt to GDP
Level of development	World Economic Outlook classification in six regions, country fixed effect (encoded variable of the country names)
Contagion	Six dummies equal to one if a crisis has occurred in one of the six WEO regions within the last eight quarters (one hot encoding)

B.2 | Stylized facts

TABLE B3 Descriptive statistics.

	Count	Mean	Std	Min	25%	50%	75%	Max
Exports to GDP (%)	4483	40.0	29.8	5.4	24.3	33.4	46.5	239.7
Imports to GDP (%)	4483	40.9	26.5	7.5	25.8	34.3	49.8	218.3
Terms of trade (%)	4483	98.7	26.4	29.1	83.9	99.3	112.1	284.4
Current account to GDP (%)	4483	-0.9	11.1	-65.1	-5.5	-0.1	4.4	54.5
Nominal exchange rate volatility	4483	28.4	140.6	0.0	0.1	0.6	4.2	2967.7
Nominal exchange rate growth (q/q %)	4483	0.8	5.1	-16.3	-1.6	0.2	2.3	62.2
Nominal exchange rate growth (y/y %)	4483	3.4	13.8	-38.9	-4.0	1.2	7.6	173.8
Real exchange rate growth (y/y %)	4483	0.9	10.6	-42.5	-5.6	0.1	5.6	76.4
Undervaluation of real exchange rate (% of trend)	4483	-0.2	6.5	-25.5	-4.0	-0.6	3.1	42.9
International reserves growth (y/y %)	4483	10.9	26.6	-66.6	-1.6	6.6	18.7	409.5
Real central bank policy rate (%)	4483	1.3	3.7	-69.8	-0.3	1.1	3.0	30.7
Central bank rate differential with the United States (pp)	4483	1.7	3.5	-67.2	0.0	1.6	3.3	31.2
CPI inflation (y/y %)	4483	4.4	5.9	-4.9	1.5	3.0	5.6	107.8
CPI inflation differential with the United States (pp)	4483	2.4	5.9	-6.0	-0.6	0.9	3.5	105.0
CPI volatility	4483	5.92.15 ^_frp_secowid=85.65 0	12.2	0.2	1.6	2.8	5.0	281.4
Private credit to GDP (%)	4483	64.2	41.8	8.4	31.7	52.7	90.2	304.6
M2 growth (q/q %)	4483	2.7	3.2	-25.1	1.1	2.2	3.9	53.5
M2 growth (y/y %)	4483	11.5	10.3	-32.8	5.6	9.2	14.9	94.1
M2 to international reserves (%)	4483	6.1	9.4	0.4	1.9	3.1	5.0	70.3
Real GDP growth (y/y %)	4483	3.7	5.8	-31.7	1.0	3.6	6.3	69.6
Output gap (% trend)	4483	0.0	7.0	-42.2	-3.0	0.0	3.1	55.7
GDP growth differential with the United States (pp)	4483	2.0	5.5	-36.3	-0.9	1.5	4.3	68.5
Public debt to GDP (%)	4483	50.6	32.3	3.1	29.8	43.3	65.6	256.4

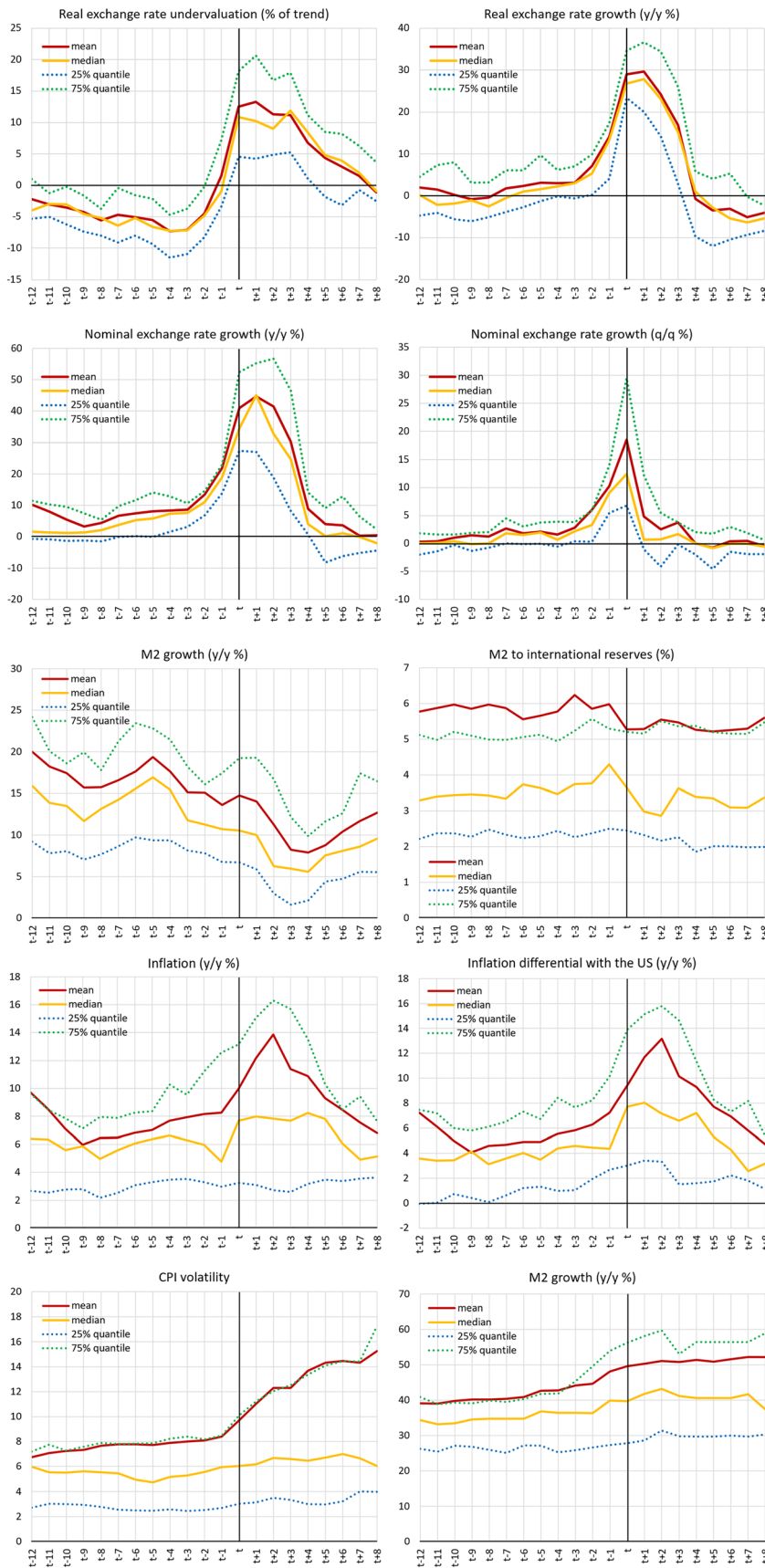
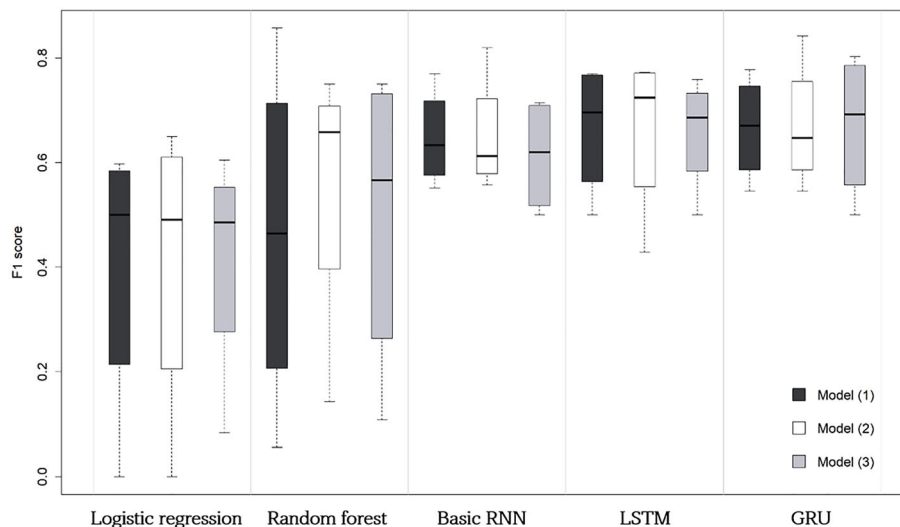


FIGURE B1 Empirical pre and post-crisis regularities. *Note:* Authors' calculations. Mean, median, and 25% and 75% quartiles of each of the numerical variables selected for the neural networks, over the 12 quarters prior to a crisis and the following eight quarters.

APPENDIX C: ROBUSTNESS CHECKS

C.1 | Capacity for generalization

FIGURE C1 Box plot of F1 scores during cross-validation (expanding window). *Note:* Model (1) refers to the standard specification (with one hidden layer for neural networks). Model (2) refers to a specification with a lower probability threshold of assignment to Class 1. Model (3) refers to a specification using the SMOTEENN algorithm for logistic regression and random forest and a standard specification with two hidden layers for neural networks.



C.0.2 | Comparison with the literature

Table C1 compares our results with those of Berg and Pattillo (1999b) and Peltonen (2006). The globally performing model of Berg and Pattillo and that of Peltonen seem to be affected by the imbalanced character of the target to be predicted, leading to a high accuracy but a low recall (reaching 25% for the KLR model and 3.6% for the ANN). Our models were trained over a longer period

with more countries, so we were able to correct this defect, allowing the identification of more than 70% of warning signals, regardless of our model. Our models are also less sensitive to noise, with only 14% of the signals sent as false for the LSTM, as opposed to 61% for Berg and Patillo's probit model and 72% for Peltonen's ANN. Our models appear to be more efficient overall (identification of alert and tranquil periods), more precise during alert periods, and less noisy.

TABLE C1 Comparison of out-of-sample performances.

	Logit ¹	LSTM ¹	GRU ¹	BP ² -Signal	BP ² -Probit	P ³ -Probit	P ³ -ANN
Percentage of observations correctly called ^a	95.0	96.5	96.4	69.0	76.0	89.2	89.0
Percentage of pre-crisis periods correctly called ^b	70.0	76.1	79.0	25.0	16.0	0.0	3.6
Percentage of tranquil periods correctly called ^c	97.6	98.7	98.2	85.0	93.0	99.9	99.5
False alarms as percentage of total alarms ^d	23.4	13.9	16.8	63.0	61.0	100.0	71.4

¹The models selected are the best performing ones based on F1 score, that is, the standard logistic regression and the standard single-hidden layer LSTM and GRU. Crises were identified using the Frankel and Rose criterion for an eight-quarter alert window. The out-of-sample period runs from Q1 2015 to Q4 2020 for 68 countries.

²BP corresponds to Berg and Pattillo (1999b). The signal model is a reproduction of the KLR model (weighted sum of indicators), whereas the probit model is a regression using a set of variables identical to that in KLR. The specifications retained are those of the 25% threshold for a 24-month alert window using the EMP criterion. The out-of-sample period runs from May 1995 to December 1997 for 23 countries.

³P corresponds to Peltonen (2006). The author compares two specifications, a probit model and an artificial neural network. The specifications chosen are those of the 25% threshold for a 3-month alert window using the EMP criterion. The out-of-sample period runs from December 1997 to December 2001 for 24 countries.

^aNumber of identified true signals and true tranquil periods/number of samples (accuracy).

^bNumber of identified true signals/number of expected signals (recall or sensitivity).

^cNumber of identified true tranquil periods/number of tranquil periods (specificity).

^dFalse signals sent/number of signals sent.

APPENDIX D: ANALYSIS OF SHAP VALUES

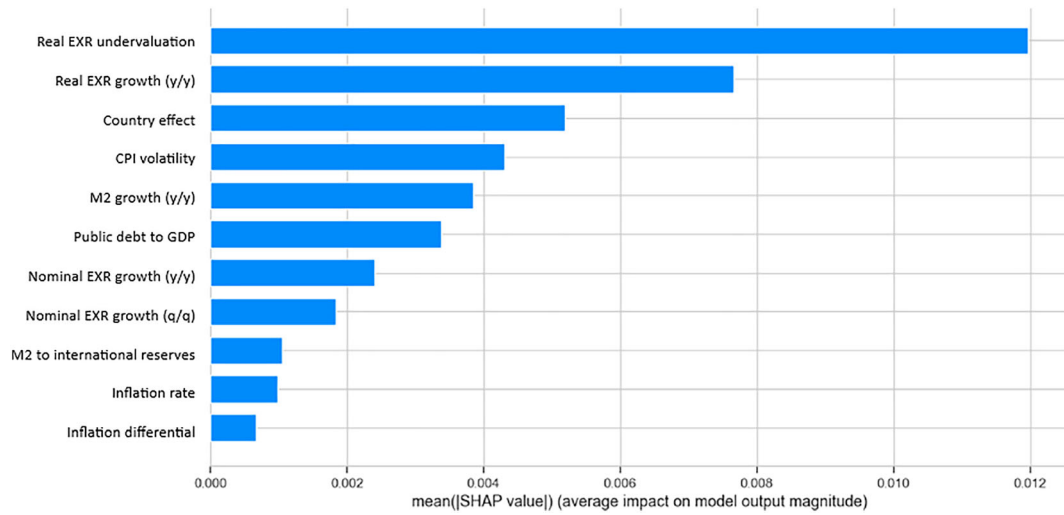


FIGURE D1 One-hidden layer LSTM, feature importance.

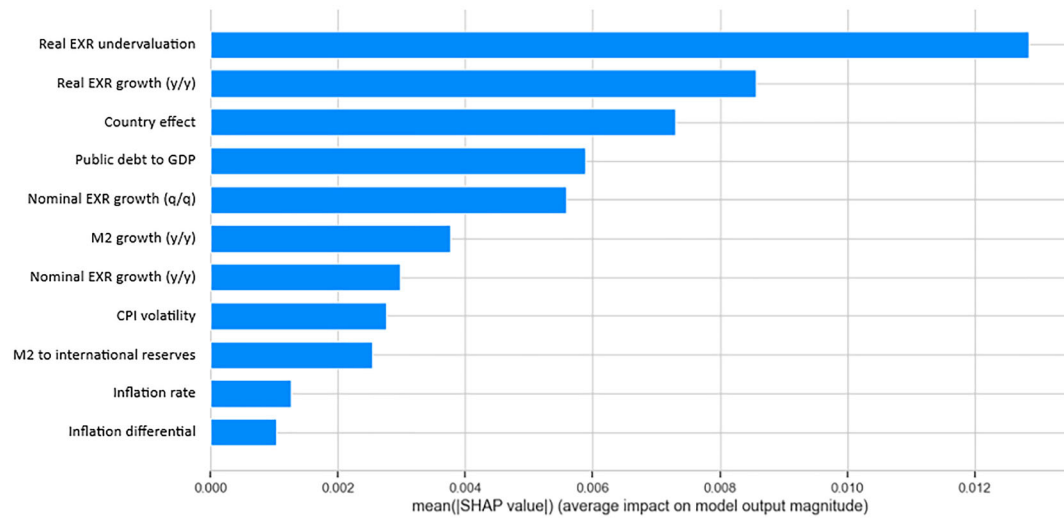


FIGURE D2 One-hidden layer GRU, feature importance.

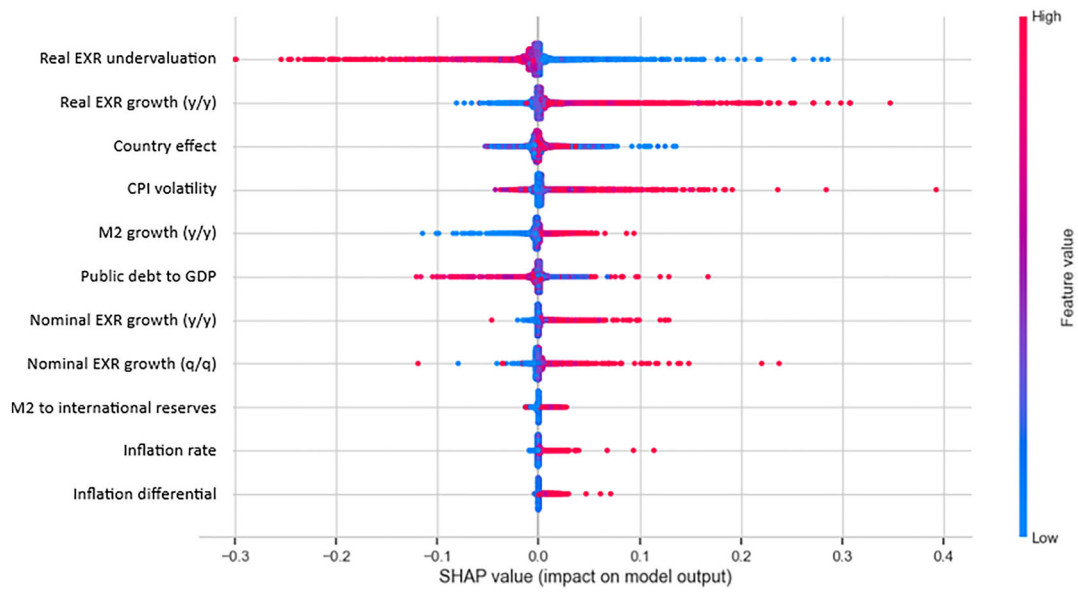


FIGURE D3 One-hidden layer LSTM, feature importance, additional.

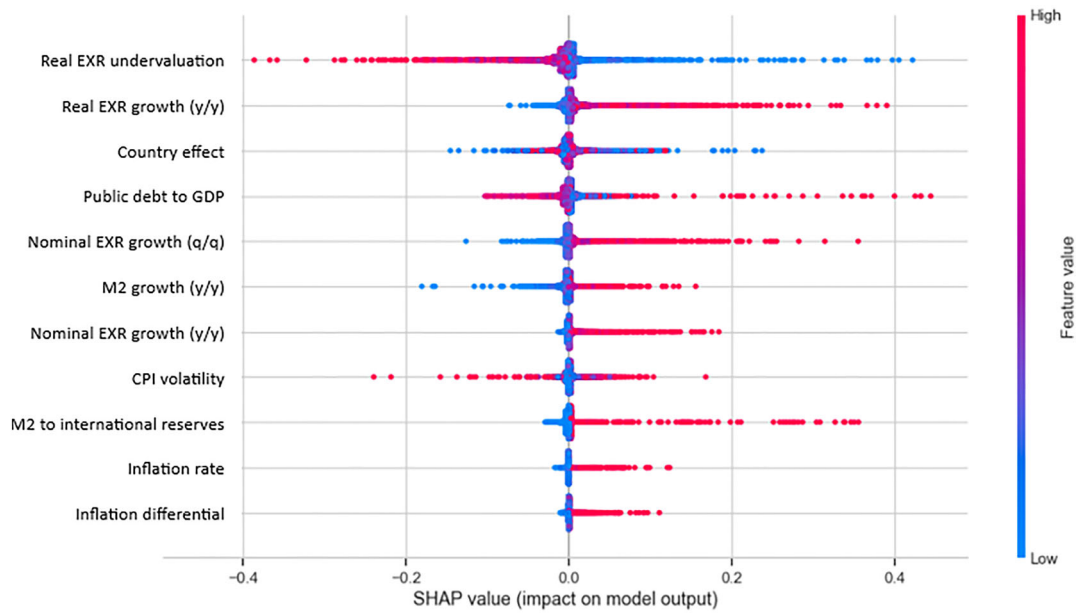


FIGURE D4 One-hidden layer GRU, feature importance, additional.

TABLE D1 Contribution of the variables from the current and past time steps to the sending of a signal in Q3 2017 by the one-hidden layer LSTM 1 year before the collapse of the Turkish lira (Q3 2018).

	Q1 2016	Q2 2016	Q3 2016	Q4 2016	Q1 2017	Q2 2017	Q3 2017	Tot. feature
Nom. EXR growth (y/y)	1.00	0.80	0.30	0.70	1.30	1.10	2.30	7.50
Real EXR growth (y/y)	2.70	2.10	-0.10	2.60	5.60	3.40	3.50	19.80
EXR undervaluation	-0.20	1.60	2.60	-1.60	-7.00	-1.00	8.00	2.40
Nom. EXR growth (q/q)	0.20	0.20	0.30	0.40	0.90	-0.10	-1.00	0.90
Country effect	0.70	3.90	5.90	3.50	2.10	2.20	1.00	19.30
Inflation (y/y)	0.20	0.10	0.10	0.10	0.20	0.40	1.20	2.30
Inflation differential	0.00	0.10	0.10	0.10	0.10	0.20	0.80	1.40
CPI volatility	-1.30	-0.20	0.50	1.30	1.20	1.50	4.30	7.30
M2 growth (y/y)	0.70	0.20	-0.20	0.60	0.50	1.10	2.70	5.60
M2 to int. reserves	0.20	0.10	0.00	-0.10	-0.20	0.30	0.30	0.60
Pub. debt to GDP	1.70	2.70	2.10	1.00	0.40	0.60	0.90	9.40
Total quarter	5.90	11.60	11.60	8.60	5.10	9.70	24.00	76.50
							Base value	7.40
							Predicted probability	83.90

Note: The LSTM and GRU take as input data matrices covering several time steps. We use seven quarters for the lagged variables included as explanatory variables in the neural networks: the current quarter and six past quarters. SHAP values are calculated on all explanatory variables, with a distinction by time step, as recurrent neural networks such as LSTM and GRU are able to differentiate between them. By breaking down SHAP values by variable and time step, we can see whether or not the same variable is deteriorating over the seven quarters and whether this deterioration is likely to affect signal probability.

TABLE D2 Contribution of the variables from the current and past time steps to the sending of a signal in Q3 2017 by the one-hidden layer GRU 1 year before the collapse of the Turkish lira (Q3 2018).

	Q1 2016	Q2 2016	Q3 2016	Q4 2016	Q1 2017	Q2 2017	Q3 2017	Tot. feature
Nom. EXR growth (y/y)	0.69	0.48	0.29	0.63	1.29	2.10	2.31	7.80
Real EXR growth (y/y)	2.33	1.34	0.16	1.90	4.80	7.50	5.86	23.88
EXR undervaluation	-0.17	0.39	0.17	0.36	-3.69	-2.18	7.47	2.35
Nom. EXR growth (q/q)	0.24	-0.24	0.38	1.83	5.07	-2.78	-2.45	2.06
Country effect	0.41	1.76	2.14	0.46	-1.54	-1.99	7.26	8.51
Inflation (y/y)	0.06	0.05	0.07	0.01	-0.01	0.42	1.45	2.04
Inflation differential	0.06	0.04	0.04	-0.04	-0.15	0.14	1.37	1.47
CPI volatility	0.07	0.08	0.06	0.27	0.61	2.23	2.91	6.23
M2 growth (y/y)	0.05	-0.02	-0.09	-0.02	0.10	1.28	2.99	4.30
M2 to int. reserves	0.04	0.03	0.03	-0.01	-0.10	-0.30	-0.54	-0.85
Pub. debt to GDP	0.18	0.62	0.77	1.57	2.45	3.28	0.62	9.50
Total quarter	3.96	4.55	4.01	6.97	8.83	9.70	29.25	67.28
							Base value	7.23
							Predicted probability	74.51

Note: The LSTM and GRU take as input data matrices covering several time steps. We use seven quarters for the lagged variables included as explanatory variables in the neural networks: the current quarter and six past quarters. SHAP values are calculated on all explanatory variables, with a distinction by time step, as recurrent neural networks such as LSTM and GRU are able to differentiate between them. By breaking down SHAP values by variable and time step, we can see whether or not the same variable is deteriorating over the seven quarters and whether this deterioration is likely to affect signal probability.