Int. J. Anal. Appl. (2025), 23:150

International Journal of Analysis and Applications

Financial Bubble Detection Using GSADF and LSTM-RNN Model: Evidence from Emerging Markets

Tran Trong Huynh[1], Bui Thanh Khoa[2,*]

[1]*Department of Mathematics, FPT University, Ha Noi, Vietnam*
[2]*Business and Management Research Group, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam*
**Corresponding author:* buithanhkhoa@iuh.edu.vn

ABSTRACT. Forecasting financial bubbles is a crucial task in financial economics due to the disruptive impact of asset price collapses on markets and economic stability. This study proposes a novel approach to bubble prediction by integrating the PSY (Phillips, Shi, and Yu) procedure for bubble detection with Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN), a machine learning technique well-suited for modeling nonlinear time-series patterns. Using weekly data from the Vietnamese stock market covering the period from 2015 to 2025, we construct a binary dependent variable indicating the presence of bubble episodes based on the GSADF test. Key macro-financial variables, including returns, volatility, and geopolitical risk, are employed as predictors. The LSTM-RNN model is trained and validated using a time-split approach (2015–2019 for training, 2020–2022 for validation, and 2023–2025 for testing), ensuring robustness and preventing overfitting. Out-of-sample results demonstrate that the LSTM-RNN achieves a high accuracy of over 81% and significantly outperforms a random walk benchmark. Our findings highlight the critical role of macroeconomic uncertainty, especially geopolitical risk, in driving bubble dynamics. This research contributes to the literature by offering an early warning framework that combines econometric detection with advanced machine learning, supporting better decision-making for investors and financial regulators in emerging markets.

1. Introduction

Asset price bubbles, periods in which asset prices deviate persistently and significantly from their fundamental values, have been a recurring feature of financial markets throughout history. From the Dutch tulip mania in the 1630s to the U.S. housing bubble of the 2000s, such episodes have often culminated in severe financial and economic crises. The collapse of bubbles

Received May 11, 2025

2020 Mathematics Subject Classification. 91G15, 62M10, 62P05.

Key words and phrases. financial bubbles; LSTM-RNN model; emerging markets; GSADF; random walk.

https://doi.org/10.28924/2291-8639-23-2025-150

ISSN: 2291-8639

can destabilize financial institutions, erode investor wealth, and trigger wide-ranging economic disruptions, as observed in the dot-com crash of 1999–2001 and the subprime mortgage crisis of 2007–2009. Beyond economic costs, the bursting of bubbles also undermines social trust and investor confidence, particularly affecting retail investors who tend to enter markets during euphoric phases and exit after steep declines. Consequently, identifying and forecasting bubble regimes has become a priority for regulators and policymakers seeking to implement timely macroprudential interventions [1, 2].

The Vietnamese stock market, inaugurated in July 2000, has exhibited several speculative episodes since its inception. Following Vietnam's accession to the WTO and the implementation of the Securities Law in 2006, the market experienced a rapid rise in capitalization, peaking during 2006–2007, before crashing in 2008. Similar dynamics were observed during 2017–2018 and again in the volatile post-pandemic period. As of mid-2023, Vietnam's market capitalization stood at approximately USD 205 billion—about 65% of GDP—reflecting its growing significance in the national economy. Despite this growth, the market remains vulnerable to price manipulation, rumor-based trading, and sentiment-driven volatility, owing in part to its nascent structure and information inefficiencies. These features make the Vietnamese market a fertile ground for asset price bubbles and highlight the need for robust monitoring frameworks [3].

Traditionally, the detection of bubbles has relied on econometric techniques such as the Sup Augmented Dickey-Fuller (SADF) and Generalized SADF (GSADF) tests developed by Phillips, Shi and Yu [4], which statistically identify explosive behavior in asset prices. These methodologies have been widely used to date-stamp bubble episodes in financial time series. However, such methods are retrospective and do not provide a predictive framework for forecasting future bubbles. To address this limitation, recent studies have begun integrating econometric insights with machine learning (ML) algorithms to enhance bubble detection and forecasting in real time [4].

Machine learning, with its ability to model complex nonlinear patterns and learn from large datasets, offers a promising alternative to traditional statistical models in forecasting applications. In the context of financial markets, ML algorithms have been applied to forecast crises [5, 6], defaults [7], and asset pricing[8]. Among these, neural networks and ensemble methods such as random forests and gradient boosting have consistently outperformed conventional models. Yet, the application of ML to forecast financial bubbles, particularly in emerging markets like Vietnam, remains limited and largely unexplored. Recent empirical evidence from Basoglu Kabran and Unlu [9] demonstrated the feasibility of using support vector machines to predict bubble episodes in developed markets [9]. However, their study did not address the dynamics of emerging financial systems nor did it incorporate sequential learning frameworks like recurrent neural networks (RNNs), which are well-suited for modeling time-

dependent phenomena. Furthermore, most existing research has focused on aggregate macroeconomic crises rather than firm-level or asset-level bubble regimes.

This study aims to fill this gap by developing a predictive framework to detect weekly stock-level bubble episodes in the Vietnamese equity market using machine learning models, with a particular focus on the Long Short-Term Memory (LSTM) architecture—a variant of RNN designed to capture long-term dependencies in sequential data. We first identify bubble periods using the GSADF test and label them as binary outcomes. We then construct lagged predictors from both technical and macro-financial domains, including volatility, skewness, return patterns, trading volume, and geopolitical risk indices. The LSTM model is trained and validated using data from 2015 to 2022 and evaluated on an out-of-sample test set spanning 2023 to 2025. For benchmarking purposes, we compare the model's performance against a naïve random walk classification rule. By integrating econometric bubble detection with machine learning prediction, our study provides a forward-looking approach to financial stability monitoring. The results offer empirical insights into the feasibility and accuracy of ML-based bubble forecasts in a frontier market, with implications for risk management, investment strategy, and regulatory oversight.

2. Literature review

Definition of Financial Bubbles

Financial bubbles, also known as asset price bubbles or speculative bubbles, describe situations where asset prices deviate significantly from their intrinsic or fundamental values. This phenomenon, while widely observed across financial history, remains theoretically and empirically contentious. Researchers often classify bubbles into two broad types: classical (irrational) and modern (rational) [1, 2].

Classical bubbles are attributed to behavioral biases and psychological forces. Shiller (2002) emphasized the role of "irrational exuberance" and media-induced feedback loops in amplifying market sentiments [10]. In contrast, modern rational bubbles suggest that asset prices can exceed fundamental value even in markets with rational agents [11]. These bubbles persist because of the belief that overpriced assets can still be sold to others at a profit. Fama (2014), a staunch proponent of the Efficient Market Hypothesis (EMH), argued that such deviations are predictable, thereby rejecting the notion of irrational bubbles altogether [12]. Recent perspectives integrate both views by proposing partially rational bubbles, where markets are influenced by a mix of rational forecasting and speculative impulses [13].

Empirical Detection of Financial Bubbles

Numerous studies have sought to empirically detect financial bubbles, especially within stock markets [14, 15]. A more robust class of methods emerged with the introduction of right-tailed unit root tests by Phillips et al. (2011), including the Supremum Augmented Dickey-Fuller

(SADF) and its extension, the Generalized SADF (GSADF) or PSY procedure [14, 16]. These techniques are particularly adept at detecting multiple episodes of explosive behavior in financial time series. Zhang, Wei, Lee and Tian [15] demonstrated the effectiveness of SADF in identifying historical bubble episodes, while the GSADF test overcame SADF's limitation by allowing both start and end points of bubbles to vary within a flexible window framework.

Applications of these methods have become standard in identifying bubbles in major markets such as the S&P 500, housing prices, and commodity indices. More recent research has also applied the GSADF framework to emerging markets, including Vietnam and China, recognizing its potential to capture regime shifts and speculative phases.

Machine Learning for Bubble Prediction

The past decade has witnessed a surge in studies exploring the use of machine learning (ML) for financial forecasting. ML models have consistently outperformed traditional econometric methods in tasks such as bankruptcy prediction [17], financial distress [18], and stock return forecasting [19].

In the specific context of bubble prediction, only a handful of pioneering studies exist: Basoglu Kabran and Unlu [9] introduced a two-step ML approach using Support Vector Machines (SVM) to forecast bubbles in the S&P 500 index. Their results indicated SVM's superiority in capturing non-linear relationships compared to logistic regression and decision trees. Tran, Le, Lieu and Nguyen [3] focused on the Vietnamese stock market from 2001 to 2021 and employed multiple algorithms, including Random Forest, ANN, and SVM. Their findings showed that Random Forest and ANN outperformed statistical benchmarks, confirming the value of ML in emerging markets.

Wang and Yampaka [2] applied logistic regression, deep learning, decision trees, and SVM to predict stock price bubbles in China from 2015 to 2023. They used four explanatory variables: inflation rate, consumer confidence index, stock yield, and P/E ratio. The logistic regression model delivered the highest accuracy and F1-score, though deep learning also showed competitive results. The authors highlight that simple models may perform better when data are limited [2].

Xiu, Kelly, Gu and Karolyi [20] emphasized that decision trees and neural networks excel due to their ability to model complex non-linear relationships, doubling the performance of linear regression in empirical asset pricing tasks. Zhou, Zhou and Long [21] confirmed similar findings using deep neural networks for forecasting equity premiums, outperforming OLS and historical averages.

Despite these successes, the literature acknowledges that ML performance can vary significantly depending on feature selection, data granularity, and tuning strategies. Moreover,

there is a notable lack of research combining bubble detection (via GSADF) with ML forecasting models—a gap that recent studies are beginning to address.

3. Method

Machine Learning for Bubble Prediction

This study employs weekly data for all stocks listed on the Ho Chi Minh Stock Exchange (HOSE) over the period from January 2015 to March 2025. The primary source of stock-level data—including prices and trading volume—is *cafef.vn*, a reliable Vietnamese financial data provider. The geopolitical risk index (GPR) is obtained from Caldara and Iacoviello (2022), available at https://www.policyuncertainty.com/gpr.html.

To construct the binary dependent variable used in our forecasting model, we follow a two-step procedure inspired by the PSY methodology [4] to identify bubble episodes in the Vietnamese stock market. Specifically, the Generalized Sup Augmented Dickey-Fuller (GSADF) test is applied to the log-transformed VN-Index price series at weekly frequency. This test recursively estimates right-tailed ADF statistics over rolling and expanding windows to detect periods of explosive behavior in asset prices, which are interpreted as financial bubbles. A bubble is deemed to occur when the ADF test statistic exceeds the simulated critical value at the 95% confidence level. Formally, the recursive ADF regression is specified as:

$$\Delta y_t = \alpha_{r_1,r_2} + \beta_{r_1,r_2} y_{t-1} + \sum_{i=1}^{p} \phi_i^{r_1,r_2} \Delta y_{t-i} + \varepsilon_t^{r_1,r_2}$$

where y_t is the stock price at time t, Δy_t is the first difference of y_t, and α, β and ϕ_i are parameters to be estimated over the sub-sample window $[r_1, r_2]$. The GSADF statistic is then computed as the supremum of SADF statistics across varying window sizes and starting points:

$$GSADF(r_0) = \sup_{r_2 \in [r_0,1]} \left(\sup_{r_1 \in [0, r_2 - r_0]} ADF_{r_1}^{r_2} \right)$$

where r_0 is the minimum window size expressed as a fraction of the full sample, and $ADF_{r_1}^{r_2}$ denotes the ADF test statistic computed over the sub-sample $[r_1,r_2]$. A week t is classified as part of a bubble regime if the GSADF test statistic exceeds the 95th percentile of the simulated critical values based on Monte Carlo simulations. Accordingly, the binary dependent variable BBt is defined as 1 if week t falls within such a bubble period, and 0 otherwise, indicating the absence of explosive price dynamics. The explanatory variables used for prediction, all lagged by one period to prevent look-ahead bias, are shown in Table 1.

These predictors are standardized using statistics computed from the training set. The entire dataset is split into training (60%), validation (20%), and test (20%) sets in chronological order to simulate a real-time forecasting scenario.

Table 1. The Variables in the LSTM-RNN Model

Variable Name	Formula	Description
Lagvol	$Lagvol_t = \log(vol_{t-1})$	Log of Volume-Weighted Average Price from the prior week
Lagvola	$Lagvola_t = \frac{1}{n}\sum_{i=1}^{n} r_i^2$	Weekly volatility as the mean squared daily return, where r is the daily return.
Laghl	$Laghl_t = High_{t-1} - Low_{t-1}$	High-low spread in the prior week
Laglo	$Laglo_t = Low_{t-1} - Open_{t-1}$	Low-open price difference in the prior week
Laggpr	$Laggpr_t = \log(gpr_{t-1})$	Log of geopolitical risk index
Lagskew	$Lagskew_t = \text{skewness}(ret_{t-1})$	Skewness of daily returns in a given week , lagged one week
LagBB	$LagBB_t = BB_{t-1}$	The lagged one week
LagVNI	$LVNI_t = \frac{VNINDEX_{t-1} - VNINDEX_{t-2}}{VNINDEX_{t-2}}$	Weekly return of VN-Index
BB		Binary label (1 if bubble detected by GSADF, 0 otherwise)

LSTM-RNN Model

To model the temporal dynamics and nonlinear characteristics inherent in financial time series, this study adopts a Long Short-Term Memory (LSTM) neural network, a variant of Recurrent Neural Networks (RNNs) specifically designed to address the vanishing gradient problem and retain long-term dependencies. The LSTM architecture used here follows the structure illustrated in Figure 1, which emphasizes the sequential flow of information through memory cells and gated operations.

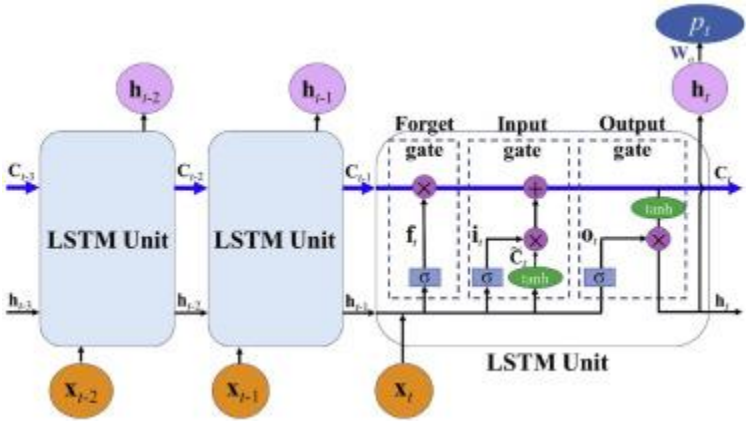

Figure. 1. LSTM architecture [22].

Each LSTM unit processes a sequence of observations spanning four weeks, where each weekly input vector $x_t \in R^7$ consists of seven lagged and standardized features: volatility, skewness, VN-Index return, the high-low price range, the low-open price difference, trading volume, and geopolitical risk. All features are lagged by one time step to avoid look-ahead bias.

Within the LSTM cell, the data flows through a series of gates that control how information is updated, forgotten, and output over time. First, the forget gate determines which components of the previous cell state C_{t-1} should be retained, computed as:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$$

Next, the input gate and candidate cell state collaborate to introduce new information:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

These components update the cell state according to the formula:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_{t-1}$$

where $\odot$ denotes element-wise multiplication. Finally, the output gate controls how much of the updated memory is transmitted forward as the hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), h_t = o_t \odot \tanh(C_{t-1})$$

The overall model architecture begins with an input tensor of shape (4,7), representing four sequential time steps and seven feature dimensions. This input is passed through an LSTM layer comprising 64 hidden units, followed by a dropout layer with a dropout rate of 0.2 to reduce overfitting. The output of the LSTM is then connected to a dense layer with 32 neurons and ReLU activation. Finally, a sigmoid-activated output neuron produces the predicted probability that the stock is experiencing a bubble regime in the given week.

The model is trained using the binary cross-entropy loss function, defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

where $y_i \in \{0,1\}$ is the observed label and $\hat{y}_i \in [0,1]$ is the predicted probability. Optimization is performed using the Adam algorithm, and early stopping is applied based on validation loss to prevent overfitting and improve generalizability. The entire flow of computation within the LSTM unit, including the gating mechanism and cell state updates, is visually summarized in Figure 1, offering a transparent view of the recurrent processing structure.

To ensure robust performance, a comprehensive hyperparameter tuning process was conducted on the training set. Several configurations were explored, varying key parameters such as the number of LSTM units (32, 64, 128), dropout rates (0.1, 0.2, 0.3), batch sizes (16, 32, 64), and learning rates for the Adam optimizer (0.001, 0.0005, 0.0001). Each combination was trained for up to 100 epochs with early stopping (patience = 10), using validation loss as the stopping criterion. The training features were standardized using the StandardScaler, with scaling parameters derived exclusively from the training set to avoid information leakage.

Model performance was evaluated on the validation set after each training run, and the optimal configuration was selected based on the validation accuracy. The best-performing model comprised an LSTM layer with 64 hidden units, a dropout rate of 0.2, a dense layer with 32 ReLU-activated neurons, and a final sigmoid output layer. It was trained using a batch size of 32 and a

learning rate of 0.001. Once the optimal configuration was identified, the model was retrained on the combined training and validation sets using the selected hyperparameters, and its predictive performance was subsequently assessed on the test set. This final evaluation offers an unbiased estimate of the model's ability to detect bubble price episodes in out-of-sample data.

Performance Metrics

To evaluate the predictive performance of the LSTM-RNN model in identifying bubble price episodes, two widely used classification metrics are employed: accuracy and F1-score. Accuracy measures the proportion of correctly classified observations over the total number of observations and provides an overall indication of model performance. However, in financial applications involving imbalanced binary outcomes, such as the detection of rare bubble events, accuracy alone may be misleading. Therefore, the F1-score, which is the harmonic mean of precision and recall, is also reported to provide a more balanced evaluation.

Accuracy is defined as: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The F1-score is given by: $F1 - score = \frac{2*Precision*Recall}{Precision+Recall}$, with $Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$

The evaluation process follows a standard procedure. After training and hyperparameter tuning, the final model is tested on the unseen test set to assess its out-of-sample generalization capability. The performance metrics are computed based on the model's predicted probabilities converted to binary classifications using a threshold of 0.5.

To provide a benchmark for comparison, a random walk classification model is used as a naive baseline. In this model, the bubble status in week t is simply assumed to be the same as in week t-1, i.e., $\hat{y}_t = y_{t-1}$. While simplistic, this approach captures the inertia commonly observed in financial regimes and serves as a useful reference point. By comparing LSTM's performance against this baseline, we assess whether the proposed model offers genuine predictive power beyond historical persistence.

The combination of these metrics enables a robust evaluation of the model's effectiveness in detecting speculative regimes, highlighting both the classification accuracy and the quality of predictions under potential class imbalance.

4. Result and Discussion

Descriptive Statistics and Correlation

Table 2 presents the summary statistics of the lagged predictors employed in the LSTM-RNN model. The variable Lagvol, representing the logarithm of trading volume from the previous week, exhibits a wide dispersion with values ranging from -100 to 616.58 and a high standard deviation of 7.98. This reflects the significant variability in trading volume across

different stocks and time periods. The weekly return volatility (Lagvola) shows a mean of 0.119 and ranges from 0 to a maximum of 6.33, which is consistent with the behavior of volatility as a non-negative and typically right-skewed variable. The high-low price spread (Laghl) and low-open spread (Laglo) also exhibit considerable variation. While Laghl has a mean of 9.36 and reaches up to 212.5, Laglo displays a more symmetric distribution around a negative mean of -4.20, reflecting the tendency of opening prices to be closer to weekly lows in many instances. The return on the VN-Index (Lagvni) is highly volatile, with extreme values ranging from -146.49 to 108.55 and a standard deviation exceeding 24.5, indicating the index's susceptibility to sharp weekly fluctuations. In terms of asymmetry in return distributions, the lagged skewness (Lagskew) centers around -0.14, with a minimum of -2.79 and a maximum of 2.39, suggesting the presence of both negatively and positively skewed return patterns. The logarithm of the geopolitical risk index (Laggpr) shows modest variability with a mean of 4.58 and a relatively narrow spread, consistent with the nature of global GPR dynamics over time.

Table 2. Descriptive Statistics of the Variables

Variable	min	Q1	Median	mean	Q3	max	Std. Dev.
Lagvol	-100	-2.5	0	0.47	2.857	616.583	7.975
Lagvola	0	0.051	0.091	0.119	0.152	6.328	0.108
Laghl	0	3.811	7.578	9.36	12.666	212.5	9.418
Laglo	-75.904	-6.775	-2.857	-4.203	0	0	5.834
Lagvni	-146.49	-8.81	3.38	1.315	14.16	108.55	24.527
Lagskew	-2.79	-1.259	-0.135	-0.138	1.376	2.394	0.749
Laggpr	4.068	4.231	4.527	4.578	4.652	5.765	0.256

The study determines a pairwise Pearson correlation matrix as shown in Table 3. Most correlation coefficients fall below ±0.3, indicating weak linear relationships between predictors. One of the stronger correlations is between Lagvol and Laglo (0.633), suggesting that trading volume is moderately associated with the weekly difference between low and open prices. Additionally, Laghl and Lagvola display a positive correlation of 0.426, reflecting the intuitive linkage between price range and return volatility. Interestingly, the correlations between each predictor and the dependent variable BB (bubble indicator) are relatively low in magnitude, ranging from -0.063 to 0.127. This highlights the nonlinear and potentially dynamic nature of bubble formation, reinforcing the need for advanced modeling approaches such as LSTM-RNN to capture such patterns. The strongest positive correlation with BB is observed for Lagskew (0.127), while the geopolitical risk index (Laggpr) shows a weaker positive association (0.093). These preliminary results suggest that linear relationships alone may not suffice in identifying bubble regimes, and more complex temporal interactions should be considered.

Table 3. Correlation Matrix

Variable	Lagvol	Lagvola	Laghl	Laglo	Lagvni	BB	Lagskew	Laggpr
Lagvol	1	0.109	0.122	0.633	0.163	-0.056	0.121	0.038
Lagvola	0.109	1	0.426	-0.187	-0.007	0.001	-0.037	0.214
Laghl	0.122	0.426	1	-0.471	-0.051	-0.063	0.081	-0.042
Laglo	0.633	-0.187	-0.471	1	0.16	0.03	0.042	0.053
Lagvni	0.163	-0.007	-0.051	0.16	1	0.023	0.185	-0.214
BB	-0.056	0.001	-0.063	0.03	0.023	1	0.127	0.093
Lagskew	0.121	-0.037	0.081	0.042	0.185	0.127	1	-0.112
Laggpr	0.038	0.214	-0.042	0.053	-0.214	0.093	-0.112	1

The LSTM-RNN and Random Walk Models

To ensure robust predictive performance, a comprehensive hyperparameter tuning procedure was carried out on the training set, which includes all weekly observations prior to 2020. Several configurations were explored, varying key parameters such as the number of LSTM units (32, 64, 128), dropout rates (0.1, 0.2, 0.3), batch sizes (16, 32, 64), and learning rates for the Adam optimizer (0.001, 0.0005, 0.0001). Each combination was trained for a maximum of 100 epochs with early stopping (patience = 10), using the validation loss as the stopping criterion. Feature standardization was applied using z-score normalization via the StandardScaler, with scaling parameters estimated solely from the training set to prevent information leakage.

Model performance was assessed on a hold-out validation set covering the period from 2020 to 2022. The optimal configuration was selected based on the highest validation accuracy and F1-score. The best-performing architecture consisted of a single LSTM layer with 64 hidden units, followed by a dropout layer with a rate of 0.2, a dense layer with 32 neurons and ReLU activation, and a final sigmoid-activated output layer. This model was trained using a batch size of 32 and a learning rate of 0.001. The selected model achieved a validation accuracy of 82.1%, and was subsequently retrained on the combined training and validation sets before being evaluated on the out-of-sample test set from 2023 onward. This final assessment provides an unbiased measure of the model's ability to identify speculative bubble regimes in unseen data.

Table 4 below presents the confusion matrices of both the LSTM-RNN and the random walk (RW) model on the test set. The LSTM-RNN model demonstrates robust predictive accuracy, correctly classifying 19,801 non-bubble weeks and 37,875 bubble weeks, with a total accuracy of 81.47% and an F1-score of 0.751. In contrast, the RW model performs only marginally better than random guessing, with an accuracy of 50.34% and F1-score of 0.445, highlighting its limited ability to capture bubble dynamics.

Table 4. Confusion matrix results of LSTM-RNN and Random Walk models on the test set

Model		Actual = 0	Actual = 1	Accuracy	F1-score
LSTM-RNN	Pred = 0	19,801	4,613	81.47%	0.751
	Pred = 1	8,503	37,875		
RW	Pred = 0	14,102	20,956	50.34%	0.445
	Pred = 1	14,202	21,532		

These results confirm the advantage of the LSTM-RNN architecture in capturing nonlinear and sequential dependencies in bubble formation. While the RW model relies solely on persistence, the LSTM-RNN learns from lagged features and temporal patterns, yielding significantly higher sensitivity, specificity, and overall predictive accuracy.

Discussion

The empirical results reveal that the LSTM-RNN model significantly outperforms the random walk benchmark in forecasting weekly asset price bubbles in the Vietnamese stock market. With an accuracy of 81.47% and an F1-score of 0.751, the LSTM-RNN demonstrates strong classification performance across both bubble and non-bubble regimes, highlighting its ability to capture nonlinear and sequential dependencies in financial data. In contrast, the random walk model, which assumes temporal persistence in bubble states, achieved an accuracy of only 50.34%, close to a coin flip, thus failing to detect meaningful patterns in the data.

Among the input features, skewness and geopolitical risk stand out with the highest positive correlations with the binary bubble indicator, at 0.127 and 0.093, respectively. Although these correlations are relatively weak, suggesting nonlinearity in the underlying relationships, the results imply that asymmetry in return distributions and global political tensions may play influential roles in bubble formation. In particular, the inclusion of GPR supports recent literature emphasizing the growing importance of non-economic shocks in financial instability [15]. The positive correlation between lagged GPR and BB, although moderate, reinforces findings from Tran, Le, Lieu and Nguyen [3], who noted that sudden increases in global risk perceptions often coincide with abrupt shifts in asset pricing behavior in frontier markets like Vietnam. This evidence supports the hypothesis that geopolitical shocks act as catalysts or amplifiers for speculative dynamics, particularly in less mature financial systems.

The ability of the LSTM-RNN to accurately forecast bubble regimes—using only lagged inputs—raises important implications for market efficiency. According to the Efficient Market Hypothesis [12], asset prices should fully reflect all available information, rendering price bubbles inherently unpredictable. However, the model's high out-of-sample predictive accuracy challenges this notion and suggests that certain patterns or structural anomalies may persist in emerging markets, allowing for the real-time detection of speculative behavior. This echoes prior

critiques of EMH in frontier markets and aligns with behavioral finance theories that stress the role of bounded rationality and sentiment-driven trading.

In terms of contribution, this research is among the first to integrate the GSADF-based bubble identification framework with deep learning techniques in a high-frequency emerging market context. While previous studies like Tran, Le, Lieu and Nguyen [3], Basoglu Kabran and Unlu [9] have applied machine learning to bubble forecasting, our study extends the literature by (i) adopting a sequential architecture (LSTM-RNN) tailored to time-series classification, (ii) focusing on stock-level weekly data for higher granularity, and (iii) incorporating geopolitical risk, volatility, and skewness as real-time predictors. These enhancements provide not only improved prediction accuracy but also practical insights for policymakers and investors seeking early warning signals in turbulent markets.

Robustness Checks

To ensure the robustness and validity of our results, several methodological safeguards were implemented. First, all independent variables were lagged by one week to avoid look-ahead bias and reduce endogeneity concerns, thereby ensuring a proper causal sequence between predictors and the binary bubble indicator. The dataset was split chronologically into training (2015–2019), validation (2020–2022), and test (2023–2025) sets, simulating real-world forecasting conditions and preventing data leakage. Feature scaling was performed exclusively on the training set to maintain the integrity of model evaluation.

Additionally, a comprehensive grid search was conducted to identify the optimal hyperparameters for the LSTM-RNN model, combined with early stopping (patience = 10) to minimize overfitting. The final model configuration—LSTM with 64 units, dropout 0.2, batch size 32, and learning rate 0.001—was selected based on its superior validation performance. Out-of-sample testing on 2023–2025 data showed that the LSTM model significantly outperformed the naïve random walk benchmark, highlighting its ability to capture nonlinear and dynamic patterns associated with bubble formation. These steps collectively confirm the model's predictive power and robustness for practical bubble forecasting applications.

5. Conclusion

This study investigates the use of machine learning, specifically the LSTM-RNN model, to forecast asset price bubbles in the Vietnamese stock market. By combining the PSY procedure for bubble detection with a supervised learning framework, we construct a weekly binary variable to indicate bubble presence and train predictive models based on macro-financial indicators. The LSTM-RNN significantly outperforms the naive random walk benchmark, achieving an accuracy of over 81% and an F1-score above 0.75 on out-of-sample data from 2023–2025. Among the predictors, geopolitical risk and volatility emerge as key drivers of bubble dynamics, aligning

with theories that highlight the role of uncertainty and investor sentiment in fueling speculative episodes.

The primary contribution of this research lies in its integration of time-varying econometric detection with modern machine learning for real-time bubble forecasting. This approach not only enhances predictive performance in emerging markets like Vietnam but also challenges the Efficient Market Hypothesis by showing that bubbles can be anticipated with high accuracy. For policymakers and investors, the findings suggest the feasibility of developing early warning systems to mitigate the adverse effects of market exuberance.

Nonetheless, this study has limitations. It relies on lagged macro variables that may not fully capture real-time sentiment shocks, and the model architecture, while effective, could benefit from further enhancement using ensemble learning or attention-based networks. Future research may extend this framework to other emerging markets, incorporate higher-frequency data, or explore causal inference methods to better isolate the drivers of bubble formation.

Conflicts of Interest: The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] T. Hirano, A.A. Toda, Bubble Economics, J. Math. Econ. 111 (2024), 102944. https://doi.org/10.1016/j.jmateco.2024.102944.

[2] Y. Wang, T. Yampaka, Predicting Stock Price Bubbles in China Using Machine Learning, Int. J. Adv. Comput. Sci. Appl. 15 (2024), 415-425. https://doi.org/10.14569/ijacsa.2024.0151139.

[3] K.L. Tran, H.A. Le, C.P. Lieu, D.T. Nguyen, Machine Learning to Forecast Financial Bubbles in Stock Markets: Evidence From Vietnam, Int. J. Financ. Stud. 11 (2023), 133. https://doi.org/10.3390/ijfs11040133.

[4] P.C.B. Phillips, S. Shi, J. Yu, Testing for Multiple Bubbles: Limit Theory of Real-time Detectors, Int. Econ. Rev. 56 (2015), 1079-1134. https://doi.org/10.1111/iere.12131.

[5] L. Alessi, C. Detken, Identifying Excessive Credit Growth and Leverage, J. Financ. Stab. 35 (2018), 215-225. https://doi.org/10.1016/j.jfs.2017.06.005.

[6] Z. Ouyang, X. Yang, Y. Lai, Systemic Financial Risk Early Warning of Financial Market in China Using Attention-lstm Model, North Am. J. Econ. Financ. 56 (2021), 101383. https://doi.org/10.1016/j.najef.2021.101383.

[7] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, A. Walther, Predictably Unequal? The Effects of Machine Learning on Credit Markets, J. Financ. 77 (2021), 5-47. https://doi.org/10.1111/jofi.13090.

[8] B.T. Khoa, T.T. Huynh, The Value Premium and Uncertainty: an Approach by Support Vector Regression Algorithm, Cogent Econ. Financ. 11 (2023), 2191459. https://doi.org/10.1080/23322039.2023.2191459.

[9] F. Başoğlu Kabran, K.D. Ünlü, A Two-step Machine Learning Approach to Predict S&P 500 Bubbles, J. Appl. Stat. 48 (2020), 2776--2794. https://doi.org/10.1080/02664763.2020.1823947.
[10] R.J. Shiller, Bubbles, Human Judgment, and Expert Opinion, Financ. Anal. J. 58 (2002), 18-26. https://doi.org/10.2469/faj.v58.n3.2535.
[11] T.E. Caravello, Z. Psaradakis, M. Sola, Rational Bubbles: Too Many to Be True?, J. Econ. Dyn. Control. 151 (2023), 104666. https://doi.org/10.1016/j.jedc.2023.104666.
[12] E.F. Fama, Two Pillars of Asset Pricing, Am. Econ. Rev. 104 (2014), 1467-1485. https://doi.org/10.1257/aer.104.6.1467.
[13] G. Cerruti, S. Lombardini, Financial Bubbles as a Recursive Process Lead by Short-term Strategies, Int. Rev. Econ. Financ. 82 (2022), 555-568. https://doi.org/10.1016/j.iref.2022.07.011.
[14] P.C.B. Phillips, S. Shi, J. Yu, Testing for Multiple Bubbles: Historical Episodes of Exuberance and Collapse in the S&p 500, Int. Econ. Rev. 56 (2015), 1043-1078. https://doi.org/10.1111/iere.12132.
[15] X. Zhang, C. Wei, C. Lee, Y. Tian, Systemic Risk of Chinese Financial Institutions and Asset Price Bubbles, North Am. J. Econ. Financ. 64 (2023), 101880. https://doi.org/10.1016/j.najef.2023.101880.
[16] P.C.B. Phillips, Y. Wu, J. Yu, Explosive Behavior in the 1990s Nasdaq: When Did Exuberance Escalate Asset Values?, Int. Econ. Rev. 52 (2011), 201-226. https://doi.org/10.1111/j.1468-2354.2010.00625.x.
[17] A. Dasilas, A. Rigani, Machine Learning Techniques in Bankruptcy Prediction: a Systematic Literature Review, Expert Syst. Appl. 255 (2024), 124761. https://doi.org/10.1016/j.eswa.2024.124761.
[18] D. Kuizinienė, T. Krilavičius, R. Damaševičius, R. Maskeliūnas, Systematic Review of Financial Distress Identification Using Artificial Intelligence Methods, Appl. Artif. Intell. 36 (2022), 2138124. https://doi.org/10.1080/08839514.2022.2138124.
[19] B.T. Khoa, T.T. Huynh, L.D. Thang, Effectiveness of Ols and Svr in Return Prediction: Fama-french Three-factor Model and Capm Framework, Ind. Eng. Manag. Syst. 22 (2023), 73-84. https://doi.org/10.7232/iems.2023.22.1.073.
[20] S. Gu, B. Kelly, D. Xiu, Empirical Asset Pricing Via Machine Learning, Rev. Financ. Stud. 33 (2020), 2223-2273. https://doi.org/10.1093/rfs/hhaa009.
[21] X. Zhou, H. Zhou, H. Long, Forecasting the Equity Premium: Do Deep Neural Network Models Work?, Mod. Financ. 1 (2023), 1-11. https://doi.org/10.61351/mf.v1i1.2.
[22] X. Wei, L. Zhang, H. Yang, L. Zhang, Y. Yao, Machine Learning for Pore-water Pressure Time-series Prediction: Application of Recurrent Neural Networks, Geosci. Front. 12 (2021), 453-467. https://doi.org/10.1016/j.gsf.2020.04.011.