Rolling Correlations and Granger Causality-Driven LSTM Modeling for Stock Market Prediction: A Study of NASDAQ Interdependencies

Giulia Talà, Mischa Dembech, Gloria Desideri, Edoardo Mocchi, Nicolò Dal Monte, Pietro Bottani and Simone Zani

BlackswanQuants Student Organization

ARTICLE INFO

Keywords:
Stock Prediction
Rolling Time-Window Correlation
Granger Causality
NASDAQ
Time Series Analysis
Deep Learning
Long Short-Term Memory (LSTM)
Quantitative Finance
Machine Learning

ABSTRACT

This study investigates the predictive relationship between equities in financial markets, assessing whether the future price of a target stock X can be inferred solely from the historical dynamics of another stock Y, their rolling time-window (RTW) correlation, and the assumption of directional influence from Y to X. A structured methodology is employed, combining Granger causality analysis, correlation-based feature extraction, and Long Short-Term Memory (LSTM) neural networks within a deep learning framework. The analysis is conducted using historical pricing data from the NASDAQ stock market.

Empirical results indicate that specific correlation patterns and causality signals exhibit statistically significant predictive power. These findings highlight the potential of correlation-aware and causality-informed forecasting models in quantitative finance, with direct implications for the development of systematic trading strategies.

1. Introduction

Financial markets exhibit complex interdependencies among assets, where the price dynamics of a given stock are influenced by the movements of other stocks. Understanding and quantifying these relationships is crucial for developing predictive models that can be utilized in trading strategies. This study explores whether the future price of a stock Y can be predicted solely based on the price movements of another stock X, their rolling time-window (RTW) correlation and the premise that Y exerts an influence on X. Combining statistical correlation analysis, causal inference, and deep learning methodologies, the study aims to determine whether such an approach can yield reliable predictive insights.

The implementation of this study begins with data retrieval, preprocessing and exploration, ensuring the integrity and consistency of high-frequency price data for all NASDAQ-listed stocks. A comprehensive correlation analysis is then conducted to identify the most strongly correlated stock pairs, as these relationships are fundamental to the subsequent predictive modeling. To capture the dynamic nature of these dependencies, a rolling correlation window is computed, producing a time series that reflects how the correlation structure evolves over time.

To establish whether the observed correlations reflect a true directional influence rather than mere statistical association, a Granger causality analysis is performed. This step is essential in distinguishing between coincidental co-movement and predictive causation, as it assesses whether past values of stock Y contain significant information that can improve forecasts of stock X. Establishing such causal relationships provides a foundation for incorporating these dependencies into a predictive framework.

The final stage of the analysis involves leveraging deep learning techniques to assess the predictive power of stock Y over stock X. A Long Short-Term Memory (LSTM) neural network is employed to learn patterns from stock Y's price movements, the computed RTW correlation, and the inferred causality relationship. By training the model on these features, it is possible to evaluate its ability to forecast the future price trajectory of stock X, thereby assessing the effectiveness of inter-stock dependencies in financial forecasting.

This study contributes to the broader field of financial econometrics and quantitative trading by evaluating the extent to which historical price information and correlation structures can be leveraged for stock price prediction. In doing so, it offers insights into the practical utility of causality-aware deep learning models for financial market analysis.

2. Literature Review

Inter-stock dependencies in financial markets refer to relationships where the price movement of one stock contains information about the future price of another. Understanding these lead–lag relationships is crucial for forecasting and quantitative trading, as it challenges the assumption that each stock moves independently. Researchers have studied this question through various lenses, from traditional correlation and cross-autocorrelation analyses, to formal causal inference tests and more recently using deep learning models that capture complex inter-stock linkages.

Kewei Hou (2007) [8] studied why certain stocks consistently lead others through the lens of industry information diffusion. Examining U.S. stock returns in the 1980s–1990s, Hou found the lead–lag effect is predominantly an intra-industry phenomenon: the returns of big firms lead the returns of small firms within the same industry. Once one accounts for these within-industry lead–lag links, very little predictability remains across unrelated industries. This suggests that when, say, a major company in a sector moves (perhaps due to an earnings announcement or industry news), smaller peers in that sector adjust more slowly, creating a predictive lag.

In recent years, researchers have turned to deep learning to model the complex web of inter-stock relationships for forecasting purposes. Traditional econometric or information-theoretic methods often consider pairwise relationships in isolation or impose linear structures. Deep learning, however, can learn higher-dimensional patterns and nonlinear interactions from large data.

Cheng and Li (2021) [3] introduced an approach to capture the momentum spillover effect via deep learning. In finance, momentum spillover refers to the phenomenon seen earlier: the past return of one stock can predict the future return of related stocks. Cheng and Li argued that previous machine learning models largely ignored these cross-stock interferences by assuming each stock's history is the sole driver of its future. To address this, they proposed an Attribute-Driven Graph Attention Network (AD-GAT) that learns a stock prediction model incorporating multiple firms' data.

Researchers have also tried integrating explicit knowledge of relationships into deep models. Shi et al. (2024) constructed a knowledge-incorporated graph where they hard-coded several types of relationships (such as same industry, supply chain linkage, and so on) between stocks. They then used an integrated GCN-LSTM model to predict price movements. By including multiple relationship types, their approach aims to capture various channels of influence. A cutting-edge direction has been to blend causal inference with deep learning. One example is the Implicit-Causality Graph Neural Network proposed by Li et al. (2024)[9], which first performs a neural Granger causality analysis to learn a directed graph of which stocks predict (Granger-cause) others.

Other deep learning studies have explored different ways to incorporate stock relationships. Tian et al. (2022) developed a model for inductive representation learning on dynamic stock co-movement graphs. They created time-varying graphs where edges reflected recent co-movements (correlations) among stocks, and then used a GNN to predict stock price movements.

While recent GNN-based models excel at capturing complex, time-varying inter-stock networks through multi-layer message passing and attention mechanisms, they demand extensive graph construction, hyperparameter tuning, and large training sets. By contrast, our framework distills inter-stock relationships via rolling time-window correlations and Granger causality tests, preselecting only the most causally relevant pairs and feeding these concise signals into a single LSTM.

3. Hypothesis

Hypothesis 1. *The objective of this study is to identify correlations among NASDAQ index stocks and determine a statistical edge for trading. To ensure robustness in our analysis, a set of a priori assumptions regarding the optimal intraday trading time window are imposed.*

1. ***Trading session constraints****: The analysis is restricted to the official trading hours of the New York Stock Exchange (NYSE), specifically from 15:30 to 21:30 GMT+2.*
2. ***Exclusion of the opening 15 minutes****: The initial phase of the trading session exhibits high volatility due to several factors, including the execution of pending orders from the previous day, overnight news impacting investor sentiment, and position adjustments by traders and institutional investors. These conditions introduce excessive noise, which can obscure meaningful correlations.*
3. ***Exclusion of the closing 15 minutes****: Similarly, the final phase of the trading session is characterized by elevated uncertainty. Empirical evidence suggests that approximately 20% of the daily trading volume is executed in the last minute before market closure, primarily driven by forced liquidations from intraday traders and algorithmic trading strategies. This effect is particularly pronounced on Fridays, further contributing to unpredictable price movements.*

Consequently, the analysis is conducted exclusively on NASDAQ-listed stocks within the intraday window from 15:45 to 21:15 GMT+2 (9:45 to 15:15 GMT-4, NY local time).

Furthermore, due to computational constraints, the dataset consists of minute observations spanning a three-month period. Expanding the number of training samples generally enhances predictive accuracy by capturing a more representative distribution of time series patterns, mitigating overfitting, and improving the reliability of distance metrics. Additionally, larger datasets contribute to the stability of prediction results across different initializations and subsampling procedures. However, these constraints limit the extent to which such benefits can be leveraged in this study.

Moreover, challenges in retrieving high-frequency intraday data further constrained the dataset. As a result, the analysis is conducted using 5-minute interval data over the three-month period, yielding approximately 18,000 observations. While this dataset provides a solid foundation for correlation-based studies, an extended time horizon or a higher-frequency dataset (e.g., 1-minute intervals) could enhance the robustness and statistical significance of the findings.

4. Methodology

4.1. Correlation Study

4.1.1. Rolling Time Window

The rolling time window (RTW) methodology is a statistical technique used to analyze time-series data by computing metrics over a fixed-length window that moves sequentially through the dataset. This approach is particularly useful in financial markets, where it helps capture evolving relationships between assets over time [11]. The primary objective of RTW analysis is to maintain temporal continuity while reducing the impact of short-term noise.To estimate an aggregated rolling correlation between two stocks over a period ΔT, consider two time series $S_X(t)$ and $S_Y(t)$, representing the stock prices of two assets X and Y over time, where $t = 1, 2, \ldots, T$. The sample $\{(S_X(t), S_Y(t))\}_{t=1}^{T}$ can be divided into $\frac{T}{\Delta T}$ equally sized non-overlapping sub-samples, which would allow for the calculation of the aggregated correlation over each sub-sample. However, if T is not sufficiently large, this approach may lead to unreliable statistical estimates due to the limited number of sub-samples.

A more reliable method is rolling correlation analysis, which computes the correlation between $S_X(t)$ and $S_Y(t)$ by moving the window forward one observation at a time. In this method, the total number of ΔT-rolling correlations is $T - \Delta T + 1$, significantly greater than $\frac{T}{\Delta T}$ for $\Delta T < T$.

Assuming that the time series $\{(S_X(t), S_Y(t))\}$ are stationary with means μ_X, μ_Y and autocovariance functions $\gamma_X(\cdot), \gamma_Y(\cdot)$, and cross-covariance function $\gamma_{XY}(\cdot)$, the rolling correlation $\{\widetilde{r}(t)\}$ is defined as the correlation between $S_X(t)$ and $S_Y(t)$ over each window:

$$\widetilde{r}^{(\Delta T)}(k) = \frac{\sum_{t=k}^{k+\Delta T-1}(S_X(t) - \mu_X)(S_Y(t) - \mu_Y)}{\sqrt{\sum_{t=k}^{k+\Delta T-1}(S_X(t) - \mu_X)^2 \sum_{t=k}^{k+\Delta T-1}(S_Y(t) - \mu_Y)^2}}.$$

This rolling correlation is a moving average process of order ΔT, and its mean and variance are given by:

$$\mathbb{E}(\widetilde{r}^{(\Delta T)}(k)) = \rho_{XY},$$

where ρ_{XY} is the true correlation between the two assets over the long run. The variance of the rolling correlation can be expressed as:

$$\mathrm{Var}(\widetilde{r}^{(\Delta T)}(k)) = \frac{1}{\Delta T^2}\Big[\gamma_X(0)\gamma_Y(0) + \gamma_{XY}(0) - 2\sum_{l=1}^{\Delta T-1}\gamma_X(l)\gamma_Y(l)\Big]$$

Assuming that the innovations $\{S_X(t) - \mu_X\}$ and $\{S_Y(t) - \mu_Y\}$ are i.i.d. with variances σ_X^2 and σ_Y^2, respectively, the rolling correlation process $\{\widetilde{r}(t) - \rho_{XY}\}$ is stationary, with mean and variance:

$$\mathbb{E}(\widetilde{r}^{(\Delta T)}(k)) = \rho_{XY},$$

$$\mathrm{Var}(\widetilde{r}^{(\Delta T)}(k)) = \frac{1}{\Delta T}\cdot\mathrm{Var}(r) \quad \text{where} \quad \mathrm{Var}(r) = \frac{\sigma_X^2\sigma_Y^2}{\Delta T \cdot \mathrm{corr}(X, Y)^2}.$$

Additionally, the autocovariance function of the rolling correlation $\gamma_{\widetilde{r}}(\cdot)$ decays for every lag $l > \Delta T$. The process is ergodic, meaning its limiting stationary distribution is the corresponding ΔT-stationary distribution.

Despite the advantage of preserving stationarity and providing more observations, the rolling-window approach has limitations. Specifically, extreme market observations, such as surges in asset prices, can have a lasting impact on the newly generated correlations. For example, an extreme value could affect the rolling correlation for an extended period (e.g., 12 months with a monthly rolling window). As a result, the newly generated correlation series may show a distorted influence from extreme values.

Predicting the severity of such an impact depends on the extremeness of the observation, the window width ΔT, and the available data. While autocorrelation between observations becomes less important as the number of observations increases, it still affects finite samples. Therefore, the width of the rolling window ΔT should be chosen based on the amount of data available and the need to balance the inclusion of recent data with the minimization of the impact of extreme outliers.

4.1.2. Pearson Correlation

The Pearson correlation coefficient, denoted as $\rho_{X,Y}$, measures the strength and direction of the linear relationship between two time-series variables X and Y [6]. It is defined as:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where $Cov(X, Y)$ represents the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively. The coefficient ranges from -1 to 1, where:

- $\rho_{X,Y} = 1$ indicates a perfect positive linear relationship.
- $\rho_{X,Y} = -1$ indicates a perfect negative linear relationship.
- $\rho_{X,Y} = 0$ suggests no linear correlation.

4.1.3. Rolling-Window Pearson Correlation

To capture the evolving relationships between financial time series, the rolling-window methodology to the Pearson correlation is applied. Given a rolling window of size ΔT, the Pearson correlation at time t is computed using observations from t to $t + \Delta T - 1$:

$$\rho_{X,Y}(t) = \frac{\sum_{i=t}^{t+\Delta T-1}(X_i - \bar{X}_t)(Y_i - \bar{Y}_t)}{\sqrt{\sum_{i=t}^{t+\Delta T-1}(X_i - \bar{X}_t)^2 \sum_{i=t}^{t+\Delta T-1}(Y_i - \bar{Y}_t)^2}},$$

where $\bar{X}_t$ and $\bar{Y}_t$ are the mean values of X and Y over the rolling window:

$$\bar{X}_t = \frac{1}{\Delta T} \sum_{i=t}^{t+\Delta T-1} X_i, \quad \bar{Y}_t = \frac{1}{\Delta T} \sum_{i=t}^{t+\Delta T-1} Y_i$$

By computing $\rho_{X,Y}(t)$ over a rolling window, a time series of correlations that highlights dynamic relationships between financial assets is obtained.

4.2. ADF Test and Stationarity

To ensure the reliability of the Granger causality analysis conducted in this study, it is essential to first verify the stationarity of the involved time series using the Augmented Dickey-Fuller (ADF) test. Stationarity implies that the statistical properties of a time series, such as its mean and variance, remain constant over time. This property is essential for many time series models, including Granger causality and ARIMA, as non-stationary data can yield misleading inferences due to evolving dynamics and spurious relationships [7].

The ADF test is a widely used statistical test for detecting the presence of a unit root in a time series, which would indicate non-stationarity [5]. The test is based on the following regression model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t, \tag{1}$$

where Δy_t is the first difference of the series, t is a time trend, and γ is the coefficient whose significance determines whether a unit root is present. The null hypothesis $H_0 : \gamma = 0$ corresponds to the presence of a unit root (non-stationarity), while the alternative hypothesis $H_1 : \gamma < 0$ indicates stationarity.

In our implementation, the ADF test is applied to each individual time series extracted from the NASDAQ dataset. If a series fails the test (i.e., the p-value exceeds a threshold of 0.05), differencing is applied until stationarity is achieved.

This preprocessing ensures that all time series inputs meet the stationarity assumption required for valid Granger causality testing and for training the LSTM-based predictive models.

4.3. Granger Causality Analysis

Causality in time series analysis refers to the ability to determine whether past values of one variable can provide predictive information about future values of another. Unlike simple correlation, which only measures statistical association, causality seeks to establish a directional influence between variables. In the context of economics and finance, causality analysis is critical for understanding dependencies between markets, macroeconomic indicators, and asset prices.

Traditional causality analysis is rooted in experimental designs, but in observational time series data, such experiments are often impractical. This challenge has led to the development of statistical methods such as Granger causality, which assesses predictive relationships rather than attempting to uncover true causal mechanisms.

Causality in time series can be categorized into various forms. Instantaneous causality occurs when two time series move together in real-time, typically measured through contemporaneous correlations. Lagged causality, on the other hand, is observed when past values of one time series help predict future values of another, indicating that the first variable Granger-causes the second. Structural causality involves the use of structural models grounded in economic theory, incorporating external factors and dynamic systems to explain causal relationships.

Granger causality specifically targets lagged causality, testing whether the past values of one variable improve the predictive accuracy of another.

4.3.1. Mathematical Formulation of Granger Causality

Granger causality is based on the premise that a time series X_t Granger-causes another time series Y_t if past values of X_t contain statistically significant information about future values of Y_t, beyond what is already captured by the past values of Y_t. This relationship is typically modeled using vector autoregressive (VAR) models.

4.3.2. Univariate Autoregressive Model

The behavior of a time series Y_t can be modeled using an autoregressive (AR) process. The univariate autoregressive model is given by:

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \epsilon_t,$$

where a_i represents the autoregressive coefficients, p is the number of lags, and ϵ_t is a white noise error term.

4.3.3. Granger Causality Test

To test whether X_t Granger-causes Y_t, the model is extended by including lagged values of X_t:

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{i=1}^{p} b_i X_{t-i} + \epsilon'_t.$$

Here, b_i denotes the coefficients associated with the lagged values of X_t, capturing their influence on the current value of Y_t. If these coefficients are significantly different from zero, it implies that past values of X_t contribute to explaining Y_t's behavior beyond what is explained by its own lags and if the inclusion of past values of X_t significantly improves the prediction of Y_t, then X_t is said to Granger-cause Y_t. This improvement is evaluated using an F-test, which compares the restricted model (without X_t) and the unrestricted model (with X_t):

$$F = \frac{(RSS_1 - RSS_2)/p}{RSS_2/(T - 2p - 1)},$$

where RSS_1 and RSS_2 are the residual sum of squares for the restricted and unrestricted models, respectively, and T is the number of observations. If the computed F-statistic exceeds the critical value, the null hypothesis that X_t does not Granger-cause Y_t is rejected.

4.3.4. Multivariate Extensions and Vector Autoregressive Models

Granger causality can be extended to multiple time series using vector autoregressive (VAR) models. A two-variable VAR(1) model is expressed as:

$$Y_t = c_1 + a_{11}Y_{t-1} + a_{12}X_{t-1} + \epsilon_{Y,t},$$
$$X_t = c_2 + a_{21}Y_{t-1} + a_{22}X_{t-1} + \epsilon_{X,t}.$$

Here, $a_{12} \neq 0$ implies that X_t Granger-causes Y_t, and similarly, $a_{21} \neq 0$ suggests that Y_t Granger-causes X_t. VAR models are widely used in macroeconomic forecasting and financial market analysis to examine the interdependencies between multiple time series [1].

4.3.5. Integration of Rolling-Window and Granger Causality Analysis

The proposed pipeline combines rolling-window correlation analysis and Granger causality to refine feature selection and improve forecasting accuracy. Rolling-window correlation is used to detect dynamic relationships between stock pairs over time, capturing how correlations evolve [10]. Granger causality analysis filters stock pairs by testing whether past stock price movements can predict future prices of another stock, ensuring that only relevant relationships are considered. Only the most significant stock pair, based on both correlation and Granger causality tests, is selected as inputs for the LSTM model.

This integration enhances the predictive capability of the model by incorporating both historical correlation dynamics and causal relationships in stock price movements.

4.4. RNN Architecture for Time Series Forecasting

To forecast the correlation dynamics between selected stock pairs, a Long Short-Term Memory (LSTM) recurrent neural network (RNN) (Figure 1) is employed. LSTMs are chosen for their ability to learn long-term dependencies in sequential data [2].

The implemented LSTM network consists of several key components. The input layer receives a sequence of lagged stock price values and correlation-based features. A single LSTM layer with 64 hidden units is used to capture temporal dependencies in the data. To prevent overfitting, a dropout layer with a dropout rate of 0.2 is applied. Finally, a dense output layer with a single neuron is used to predict the next correlation value.

The model is therefore formulated to ingest as inputs the observed (i.e., "caused") series X, together with the rolling-time-window correlation $\widetilde{r}^{(\Delta T)}_{X,Y}$, and to forecast the "causal" series Y. Although Granger causality establishes that past values of Y improve the prediction of X (i.e., $Y \xrightarrow{\text{Granger}} X$), the methodology deliberately inverts this forecasting direction—using X to predict Y—for two principal scientific reasons.

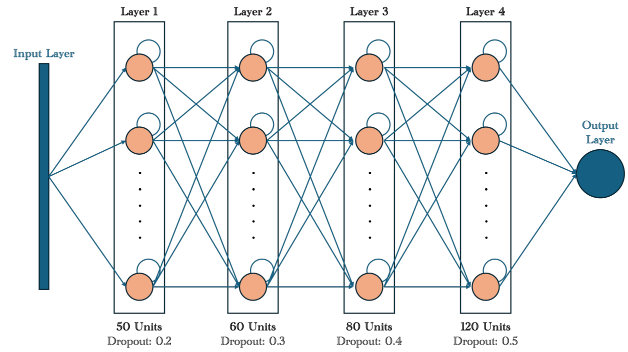

Figure 1: Graphical representation of the LSTM employed.

1. **Signal Reconstruction from Response.** In many dynamical systems, the observable response X represents a noisy and possibly filtered manifestation of the true driver Y. Training the LSTM on X enables the network to learn an inverse mapping, effectively reconstructing the latent input signal from its observable effects. This approach is conceptually analogous to deconvolution: given a system's output and knowledge of its cross-correlation with the driver, one may infer the underlying excitation.
2. **Enrichment through Rolling Correlation.** The rolling-time-window correlation $\widetilde{r}^{(\Delta T)}_{X,Y}(t)$ encodes both the strength and direction of the dynamic coupling between X and Y. When this signal is provided alongside the historical trajectory of X, the LSTM can condition its internal state on both the endogenous dynamics of X and the exogenous information contained in the correlation structure. This fusion enhances the model's capacity to infer the underlying driver Y based on the observed responses of X.

Formally, if one denotes by

$$\mathbf{h}_t = \mathrm{LSTM}\left(\left[X_{t-\Delta T+1:t},\ \widetilde{r}^{(\Delta T)}_{X,Y}(t-\Delta T+1 : t)\right]\right)$$

the hidden state that processes the past ΔT values of both the observed series and the rolling correlation, then the forecast is given by

$$\hat{Y}_{t+1} = W_o\,\mathbf{h}_t + b_o,$$

which learns the inverse representation of the Granger-causal relationship $Y \to X$. This modeling strategy allows the LSTM to recover causal signals from their empirical consequences, rather than merely replicating the original Granger direction.

The training phase of the model focuses on optimizing its parameters to minimize prediction errors. Several strategies are employed to achieve this goal. The loss function used to

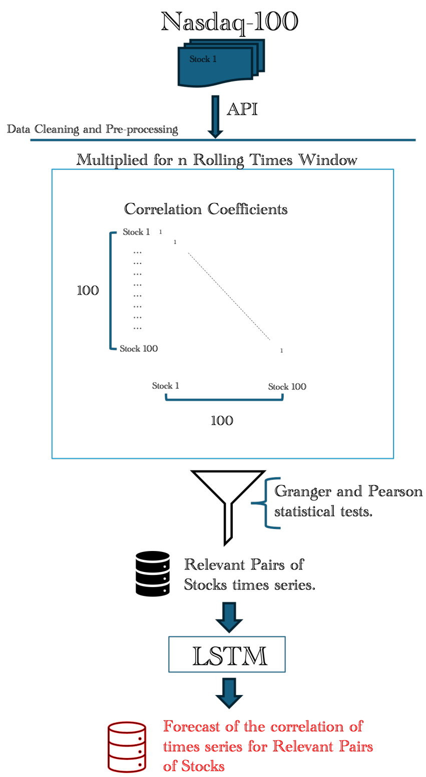

Figure 2: Model pipeline

measure prediction accuracy is Mean Squared Error (MSE). The Adam optimizer is applied for efficient gradient updates, ensuring faster convergence. Batch training is employed with a batch size of 32 to balance computational efficiency and model performance. Finally, early stopping is implemented, halting the training process when the validation loss stops decreasing to prevent overfitting [4].

4.4.1. Model Evaluation Metrics

To assess the model's performance, multiple evaluation metrics are employed:

- **Root Mean Squared Error (RMSE):** Measures prediction error magnitude.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2}$$

- **Mean Absolute Error (MAE):** Evaluates the absolute differences between actual and predicted correlation values.

- **Pearson Correlation Coefficient:** Measures the relationship between predicted and actual correlation values.

5. Model Pipeline

Source Code The authors developed all code and implementations for this study. While external libraries (e.g., SciPy, scikit-learn) are used, the core methods, algorithms, and preprocessing were fully customized for applying an statistical and deep learning methods to the data. The code is accessible at this link.

The overall workflow pipeline can be seen on **Figure** 2.

5.1. Data Collection and Preprocessing

The first step in building the model involves the collection and preprocessing of historical stock price data for NASDAQ-listed stocks. The dataset consists of time-series data, including adjusted closing prices over a specified period, which is essential for subsequent correlation and causality analyses. The preprocessing stage ensures data quality and consistency by performing several critical steps to address potential issues such as missing data, time misalignment, noise reduction, and normalization.

The dataset is initially retrieved using the Twelve Data API, which provides high-frequency intraday stock price data. A custom-built class, `IndexData_Retrieval`, was developed to automate key tasks including the extraction of NASDAQ tickers, API interaction, rate limit management, and timestamp alignment across multiple securities. The class efficiently retrieves historical price data for each stock ticker over a three-month period, with a 5-minute interval granularity. The data includes essential variables such as timestamps, open, high, low, close prices, and trading volumes, with timestamps aligned to New York trading hours.

To ensure compliance with the API rate limits, the tickers were processed in batches of up to 55 symbols per request cycle. A batching mechanism was implemented, which pauses between requests to maintain data integrity and prevent rate limit violations. Furthermore, a `Timestamping` module was incorporated to enforce data consistency within the defined trading hours of 9:45 AM to 3:15 PM Eastern Standard Time (EST), ensuring that all data points align with the hypothesis of optimal trading periods.

Upon retrieval, the data is structured into a pandas DataFrame, where each row corresponds to a timestamp and each column represents a stock ticker. The processed dataset is then stored as a pickle file to optimize storage efficiency and ensure quick access for downstream analysis.

5.1.1. Data Cleaning

To ensure the integrity of the dataset, a rigorous cleaning procedure was applied to handle missing values and preserve the continuity of the time series. Given the potential for gaps in data due to temporary unavailability or API inconsistencies, stocks with excessive missing values were removed. Using the `clean_df()` method, a threshold was set to exclude stocks with more than 30% missing data, thereby retaining only those securities with sufficiently complete time series.

For the remaining stocks, missing values were addressed using forward filling, where each missing entry was replaced by the most recent available observation. This technique is particularly suited to stock price data, as it assumes price continuity in the absence of new information. This imputation process ensures temporal consistency across the dataset, minimizing distortions that could impact the subsequent analysis.

After cleaning, the dataset is validated to confirm that all remaining stocks contain adequate historical data and no missing values persist. The cleaned dataset is then saved as a pickle file (`cleaned_nasdaq_df.pkl`) to facilitate efficient storage and rapid access during further analysis, including correlation studies, Granger causality testing, and machine learning model training.

5.1.2. Stationarity of the Time Series

For the Granger causality analysis, stationarity of the time series data is a prerequisite. The Augmented Dickey-Fuller (ADF) test was applied to each time series to assess the presence of unit roots and verify whether the data exhibits stationarity. The time series were stationary, so no further trasforming was necessary.

By applying these preprocessing techniques—data retrieval, cleaning, stationarity tests, noise reduction—the dataset is thoroughly prepared for subsequent correlation analyses, Granger causality testing, and predictive modeling with deep learning methods. This systematic approach ensures that the data is of high quality and suitable for drawing meaningful insights from the interdependencies between the stocks in the NASDAQ-100 index.

5.2. Exploratory Data Analysis

To gain initial insights into the data and prepare for the study, an exploratory data analysis on the cleaned data set was performed.

5.2.1. Descriptive Statistical Overview

To establish a foundational understanding of the NASDAQ Composite Index's behavior, a comprehensive descriptive statistical analysis of its adjusted closing prices over the observed sample period was conducted. This preliminary step is critical for characterizing the central tendency and dispersion of the distribution, thereby informing subsequent modeling and inferential procedures.

Let $\{P_t\}_{t=1}^{T}$ denote the time series of adjusted closing prices, where T is the number of observations. The sample mean $\bar{P}$, a measure of central tendency, is computed as:

$$\bar{P} = \frac{1}{T}\sum_{t=1}^{T} P_t$$

For the dataset analyzed, the mean value is $\bar{P} \approx 7434.80$, indicating moderate long-term growth of the index. However, this figure alone may obscure underlying variability, particularly in volatile financial markets.

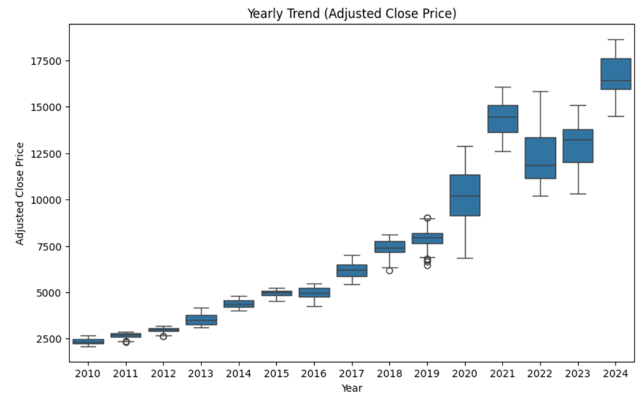

Figure 3: Adjusted close price of the index

To quantify dispersion, the sample standard deviation σ_P is calculated:

$$\sigma_P = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(P_t - \bar{P})^2}$$

A standard deviation of $\sigma_P \approx 4413.89$ reveals considerable variability in the index, underscoring the presence of substantial fluctuations likely driven by macroeconomic shocks, policy changes, and systemic events (e.g., the 2020 COVID-19 pandemic). Such volatility is characteristic of equity indices and must be considered in risk modeling and volatility forecasting.

In addition to mean and standard deviation, the sample median $\tilde{P} = 6118.27$ lies significantly below the mean, indicating a positively skewed (right-skewed) distribution. This suggests that large upward deviations—particularly in the latter years of the sample—have increased the average. Such behavior may reflect speculative bubbles or sustained bull-market phases that dominate long-term performance metrics.

5.2.2. Time Series Dynamics and Smoothing via Moving Averages

An initial visual inspection of the NASDAQ Composite Index's adjusted closing prices reveals a pronounced upward trajectory, with particularly sharp growth and heightened volatility commencing in early 2020. This inflection point aligns with globally disruptive macroeconomic events, notably the onset of the COVID-19 pandemic, which catalyzed both rapid monetary expansion and unprecedented fiscal interventions. Such conditions significantly altered market dynamics and introduced new sources of systemic risk and speculative behavior, as reflected in the price series.

To mitigate high-frequency noise and uncover underlying trends in the time series, moving average (MA) filters at varying temporal resolutions is employed. A moving average of order k is defined for a univariate time series $\{P_t\}$ as:

$$\mathrm{MA}_t^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1} P_{t-i}$$

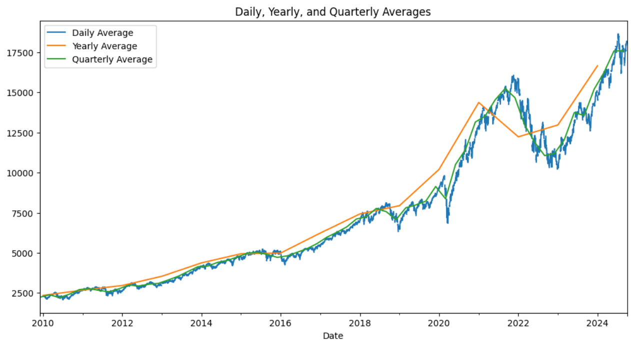

Figure 4: Average price of the index

Three different moving averages were computed for this analysis and are shown in **Figure 4**.

5.2.3. Pearson Correlation and Autocorrelation Analysis

To examine the relationship between the adjusted closing prices and their respective returns, first the percent change in adjusted closing prices is computed, defined as:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1,$$

where P_t is the adjusted closing price at time t. This allows us to analyze the returns associated with the adjusted close price over the analyzed period.

Pearson Correlation Analysis

The Pearson correlation coefficient is calculated to assess the linear relationship between the percent change in adjusted closing prices and the adjusted closing prices themselves.

In this analysis, the computed Pearson correlation coefficient between the adjusted closing price and the percent change is found to be approximately:

$$\rho_{P,r} \approx 0.015,$$

which indicates a very weak positive correlation between the adjusted closing prices and their respective returns.

Autocorrelation of Returns

To assess the temporal dependence of returns, the autocorrelation of the percent changes in adjusted closing prices is computed. The autocorrelation at lag k is defined as:

$$\rho_k = \frac{\sum_{t=k+1}^{T}(r_t - \bar{r})(r_{t-k} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2},$$

where r_t represents the return at time t, $\bar{r}$ is the mean of the returns, and T is the total number of observations. This formula quantifies the correlation between the return at time t and the return at time $t-k$, normalized by the variance of the returns.

As seen in **Figure 5** the autocorrelation at lag 0 is always 1, as any value is perfectly correlated with itself.

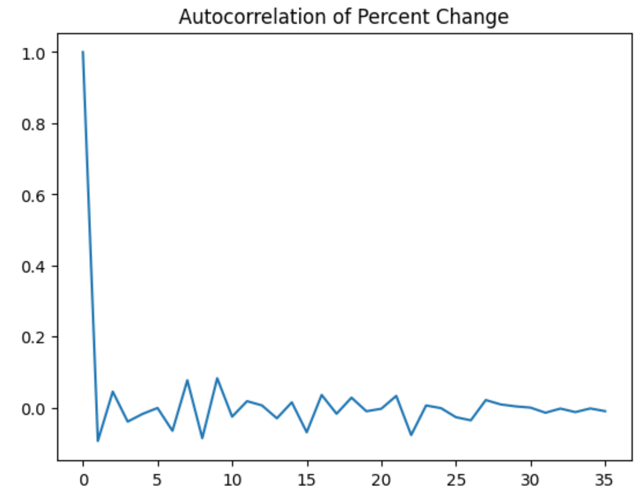

Figure 5: Autocorrelation of percent change

The empirical autocorrelation plot shows that after lag 0, the autocorrelations rapidly decrease and approach zero, with only slight fluctuations observed at higher lags. This behavior suggests that the returns in the NASDAQ adjusted close price series exhibit minimal serial dependence over time.

5.2.4. Random Walk Hypothesis using a Regression Model

To formally test the random walk hypothesis, an Ordinary Least Squares (OLS) regression model was employed. The adjusted closing price at time t, denoted P_t, was regressed on its one-period lag P_{t-1}:

$$P_t = \alpha + \beta P_{t-1} + \epsilon_t$$

Where P_t represents the adjusted closing price at time t, P_{t-1} is the adjusted closing price at time $t-1$, α is the intercept, β is the slope on the lagged price, and ϵ_t is the error term.

The key estimate is β, which here equals 1.0003 (p-value = 0.00), indicating it is statistically indistinguishable from unity—exactly the signature of a unit root process or random walk. The F-statistic of 5.34×10^6 (p-value = 0.00) confirms the overall model fit, while the intercept $\alpha \approx 2.26$ (p-value = 0.546) is not significant, suggesting no systematic drift.

The model explains 99.9% of the variance in P_t ($R^2 = 0.999$), and a Durbin–Watson statistic of 2.136 indicates no serious autocorrelation in the residuals. A high condition number (1.69×10) warns of potential numerical sensitivity. Diagnostic tests (Jarque–Bera, Omnibus) reveal non-normal, leptokurtic residuals (kurtosis = 10.608), common in financial series with fat tails.

There is no contradiction with Section 5.1.2, because:

1. Section 5.1.2 applies the ADF test to the *daily **returns*** series, which are stationary by construction, whereas here we diagnose the *price level* for a unit root.

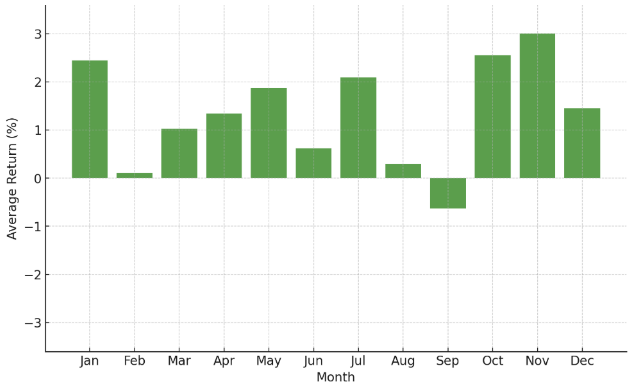

Figure 6: Average Monthly Returns for Nasdaq 100

2. The random walk regression in this subsection uses daily data on adjusted closing prices, not the already differenced or return series used for Granger causality.

Thus, while the price level behaves as a random walk (non-stationary), the transformed return series in Section 5.1.2 satisfies the stationarity requirement for subsequent Granger causality analysis.

5.2.5. *Seasonality Analysis*

As shown in **Figure 6**, the seasonality analysis highlights that August, September, and February tend to underperform. August and September coincide with the summer period, often marked by reduced trading activity and investor caution, while February also shows weaker returns, possibly due to mid-winter sentiment. In contrast, January, October, and November display stronger performance. January's returns are typically linked to the "January effect," where renewed optimism drives prices higher, while October and November may benefit from market corrections or improved clarity after weaker months.

6. Correlation & Granger Study Results

Pairwise Pearson correlation coefficients were computed over the full intraday sample (5-minute bars spanning three months) for all NASDAQ-100 tickers. After forward-filling missing observations (to accommodate market closures and data gaps) and aligning time series, the strongest correlation was found between CTSH and SIRI:

$$(\text{CTSH, SIRI}) \quad \text{with} \quad \rho_{\text{CTSH,SIRI}} = 0.96799.$$

Rolling correlation was then calculated using a window of $\Delta T = 60$ observations (approximately five trading hours). The resulting series

$$\tilde{r}^{(60)}_{\text{CTSH,SIRI}}(t)$$

exhibited a mean of roughly 0.95 and a standard deviation of about 0.03, indicating a consistently strong co-movement throughout the period.

Stationarity of both CTSH and SIRI price series was confirmed via the Augmented Dickey–Fuller test, each yielding an ADF statistic well below the 1

- SIRI $\xrightarrow{\text{Granger}}$ CTSH (F-statistic significant at $p < 0.01$)
- CTSH $\overset{\text{Granger}}{\not\longrightarrow}$ SIRI (no significance at conventional levels)

These findings reveal a unidirectional lead–lag relationship: past values of SIRI contain predictive information for future CTSH prices, whereas past CTSH values do not Granger-cause SIRI. This asymmetry motivates the inclusion of lagged SIRI (and its rolling correlation with CTSH) as an exogenous feature in subsequent forecasting models.

7. Prediction Results & Performance Evaluation

Following the correlation and Granger causality analyses, which identified a unidirectional Granger causality from SIRI to CTSH (SIRI $\xrightarrow{\text{Granger}}$ CTSH), the methodology proceeded to the predictive modeling phase using a Long Short-Term Memory (LSTM) network. As outlined in the methodology (Section 4), the forecasting direction suggested by Granger causality was inverted. The model was trained to predict the future price movements of the Granger-causing stock (SIRI) using the historical prices of the Granger-caused stock (CTSH) and their rolling time-window (RTW) correlation ($\tilde{r}^{(60)}_{\text{CTSH,SIRI}}$) as input features. This approach was designed to test the hypothesis that the dynamics of the 'caused' variable, enriched with their dynamic correlation, contain sufficient information to reconstruct or predict the 'causal' driver.

7.1. Data Preparation and Model Setup

The input data for the LSTM model consisted of sequences derived from the CTSH time series and the precomputed 60-period rolling correlation between CTSH and SIRI. The target variable was the SIRI time series. Both input features (CTSH price, correlation) and the target variable (SIRI price) were normalized using 'StandardScaler' prior to model training to ensure numerical stability and facilitate convergence. The data was transformed into suitable sequences for the LSTM using 'TimeseriesGenerator' with an input sequence length of 60 time steps (matching the RTW correlation window). The dataset was chronologically split into a training set comprising the first 70% of the observations and a testing set containing the remaining 30%. The LSTM architecture employed consisted of two sequential LSTM layers, each containing 64 hidden units and utilizing the Rectified Linear Unit (ReLU) activation function. To mitigate overfitting, a Dropout layer with a dropout rate of 0.3 was applied after each LSTM layer. The network culminated in a Dense output layer with a single neuron, tasked with producing the one-step-ahead prediction for the normalized SIRI price.

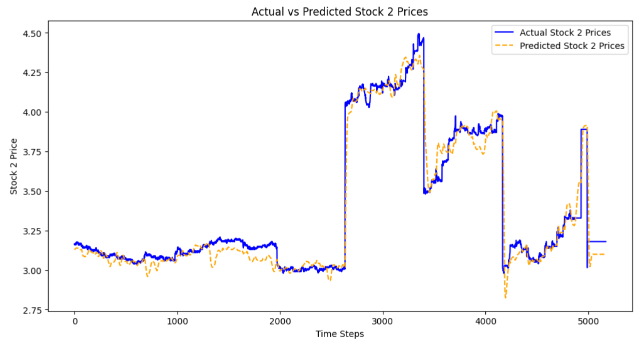

Figure 7: Prediction versus actual values of predicted Stock

7.2. Training and Evaluation

The LSTM model was compiled using the Adam optimizer with a learning rate set to 1×10^{-4}. Mean Squared Error (MSE) was selected as the loss function to quantify the difference between predicted and actual values during training. The model was trained for a maximum of 20 epochs, utilizing a batch size of 32, as implicitly defined by the 'TimeseriesGenerator' setup. An 'EarlyStopping' callback mechanism was implemented, monitoring the training loss ('loss') with a patience of 3 epochs. This ensured that training ceased if no improvement in loss was observed for three consecutive epochs, and the model weights yielding the best performance (lowest loss) were restored. The predictive performance of the trained LSTM model was rigorously evaluated on the unseen test set. Predictions were generated for the test sequences, and both the predictions and the actual target values were inverse-transformed back to their original price scale using the 'StandardScaler' previously fitted to the SIRI data. Performance was assessed using standard regression metrics, and the results can be seen on **Figure 7**:

- Mean Squared Error (MSE): 0.0089
- Mean Absolute Error (MAE): 0.0544
- R-Squared (R^2): 0.9518

7.3. Discussion of Results

The evaluation metrics demonstrate a high level of predictive accuracy for the LSTM model. The R^2 value of 0.9518 is particularly compelling, indicating that approximately 95.18% of the variance in the SIRI stock price within the test set is explained by the model using the historical price of CTSH and the dynamic rolling correlation between the two stocks. The low values for MSE (0.0089) and MAE (0.0544) further underscore the model's ability to generate predictions that closely match the actual price movements of SIRI on its original scale. Figure 7 provides a visual comparison between the actual SIRI prices and the prices predicted by the LSTM model over the test period. The plot clearly shows a strong concordance between the two time series, with the model successfully capturing the trends and fluctuations in the SIRI price. These quantitative and qualitative results strongly support the hypothesis that significant predictive information about the Granger-causing variable (SIRI) can be extracted from the Granger-caused variable (CTSH) when augmented with their dynamic correlation, thereby validating the efficacy of our inverted modeling strategy in this specific case.

8. Limitations and Practical Considerations

The strong in-sample predictive performance of the LSTM model ($R^2 \approx 0.952$) should not be conflated with immediate tradability or guaranteed profitability. In real-world settings, statistical accuracy often breaks down under the weight of trading frictions (such as transaction costs, slippage, market impact, and execution delays) all of which can significantly erode returns. Empirical studies consistently show that these frictions can reduce gross backtest profits by 15–40%, underscoring the gap between statistical fit and financial viability.

Metrics such as R^2, MSE, or MAE provide a useful snapshot of predictive skill but fall short as indicators of strategy robustness. A meaningful evaluation requires simulation under realistic trading assumptions, including bid-ask spreads, order book depth, and latency. Without such considerations, backtest results risk being overly optimistic.

Deploying LSTM models in live trading also introduces engineering challenges. Predictions must be generated with minimal delay, especially in intraday contexts, and models need frequent recalibration to account for market drift and nonstationary dynamics. LSTMs are powerful but prone to overfitting, particularly in noisy environments like financial markets, where they may latch onto transient, non-recurring patterns rather than stable economic signals.

This risk is heightened by the structural instability of financial systems. Asset relationships often change in response to macroeconomic events, regulatory shifts, or liquidity disruptions. A model trained on historical co-movements (like those between SIRI and CTSH) may falter if these dynamics shift. For instance, a sharp divergence in their rolling correlation could render previous predictive patterns obsolete.

To counteract these vulnerabilities, rigorous cross-validation is essential. Time-stratified testing across varying market regimes (bull, bear, stressed) and ensemble methods can help mitigate dependency on specific patterns or parameters. Additionally, adaptive techniques like online learning, regularized retraining, normalization layers, or even regime-switching architectures may improve generalization under dynamic conditions, though they add to the system's complexity.

High-frequency data are notoriously volatile and feature structural breaks, heavy tails, and abrupt shocks triggered by events like flash crashes or algorithmic surges. The rolling-window approach used in this study assumes a degree of continuity that may not hold under these conditions. Although LSTMs are theoretically capable of modeling long-range dependencies, their ability to generalize under distributional shift remains an open research question.

Another critical issue is the disconnect between statistical performance and economic value. A model can exhibit high accuracy on average yet fail to produce profits if it misclassifies inflection points or introduces delays that impair execution. Small prediction errors or even a few milliseconds of latency can wipe out the thin margins typical of short-horizon strategies.

Thus, future research should extend beyond backtesting and incorporate portfolio-level simulations under real trading constraints. Comparing model-driven strategies to passive benchmarks like buy-and-hold indices, with full accounting for execution costs and liquidity constraints, would yield more realistic assessments of profitability and risk-adjusted performance.

Operational concerns also abound. Deep learning models require robust infrastructure for low-latency inference, frequent retraining, and anomaly detection. Real-world financial data often contain missing values, outliers, and sudden shifts, necessitating fault-tolerant mechanisms.

Finally, the nature of the predictive signals warrants scrutiny. The model in this study relies purely on observed statistical co-movement, without incorporating causal structure. In finance, correlations can be fleeting or spurious, especially under changing regimes. Recent advances in causality-aware machine learning argue for embedding causal reasoning into model design. Techniques like structural causal models, invariant risk minimization, or graph-based learning seek to distinguish true causal drivers from confounding noise.

The limitations of Granger causality and Pearson correlation should also be acknowledged. While Granger causality testing is useful for identifying temporal precedence, it cannot establish a definitive causal relationship, especially in complex market environments where external factors or omitted variables may influence both series. Moreover, Pearson correlation, as a measure of linear dependence, may overlook non-linear relationships and be sensitive to outliers, which are common in financial data.

While beyond this paper's scope, integrating such methods could enhance both interpretability and robustness, making forecasts more resilient to shocks and structural change. Ultimately, blending deep learning with causal frameworks may offer a more stable foundation for predictive modeling in complex financial environments.

9. Conclusions

This study has investigated interdependencies within the NASDAQ stock market by developing a unified pipeline that integrates rolling Pearson correlation analysis, Granger causality testing, and Long Short-Term Memory (LSTM)-based deep learning for price forecasting. High-frequency intraday data for NASDAQ-listed equities were systematically processed, identifying CTSH and SIRI as the most strongly correlated pair ($\rho \approx 0.97$) over the three-month window. Stationarity of each price series was confirmed via Augmented Dickey–Fuller tests, and Granger causality analysis revealed a significant unidirectional influence from SIRI to CTSH (SIRI $\xrightarrow{\text{Granger}}$ CTSH).

A novel inverted modeling approach was proposed in which the historical series of the Granger-caused asset (CTSH) and its rolling correlation with the Granger-causing asset served as inputs to predict future values of SIRI. The LSTM network achieved substantial predictive accuracy, with

$$R^2 = 0.9518,$$

alongside low mean squared error (MSE) and mean absolute error (MAE), demonstrating that the "effect" series combined with dynamic interaction signals encapsulates rich information about the "cause" series.

The results suggest three main conclusions. First, strong linear correlations and Granger-causal relationships can be effectively identified and quantified even in the presence of noise typical of high-frequency financial data. Second, LSTM architectures are capable of learning complex temporal dependencies from an asset's own price dynamics, especially when augmented with its evolving correlation to a causally linked asset. Third, and most notably, the Granger-caused series, when combined with dynamic interaction terms such as rolling correlation, can encode sufficient information to allow high-fidelity prediction of the Granger-causing series itself.

This framework offers practical implications for quantitative trading and risk management by providing a data-driven mechanism for uncovering predictive cross-asset relationships. In particular, the success of the inverted LSTM approach suggests potential improvements in forecasting models through incorporation of dynamic correlation and causality features, which may capture feedback loops beyond linear causality.

Several limitations warrant consideration: the reliance on linearity and stationarity assumptions in correlation and Granger methods; sensitivity to rolling-window selection; computational scalability for larger universes; and restriction to a single asset pair over a limited interval. Future work should explore non-linear information measures (e.g., mutual information, transfer entropy), adaptive windowing strategies, advanced deep architectures (e.g., attention-based models, graph neural networks with causal priors), extension to broader asset universes and longer horizons, and comprehensive backtesting to assess economic performance under realistic trading frictions.

In summary, this research demonstrates the feasibility and value of combining rolling correlations, Granger causality, and LSTM networks for modeling and forecasting asset interdependencies. It contributes both a methodological pipeline and empirical evidence that leveraging dynamic relationships and causal signals via deep learning can enhance precision in financial time-series prediction.

10. Acknowledgements

This work would not have been possible without the support of the Quants at BlackSwan Quants, a student club that brings together enthusiasts of Quantitative Finance. The club fosters an environment where individuals can collaborate on increasingly challenging projects, applying the skills learned throughout their academic journey while building a network of ambitious people always eager to learn new things and meet new people.

References

[1] C. Amornbunchornvej, E. Zheleva, and T. Y. Berger-Wolf. *Variable-lag Granger Causality for Time Series Analysis*. https://arxiv.org/pdf/1912.10829.pdf. 2021.

[2] Dickson Chan, Michael Hsieh, and Sophie Pan. *Deep Learning for ETF Price Prediction*. International Journal of Finance. https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647694.pdf. 2020.

[3] Rui Cheng and Qing Li. "Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks". In: *Proc. Conf. AAAI Artif. Intell.* 35.1 (May 2021), pp. 55–62.

[4] H. K. Choi. *Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model*. Journal of Financial Data Science. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/. 2019.

[5] David A. Dickey and Wayne A. Fuller. "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American Statistical Association* 74.366a (1979), pp. 427–431.

[6] Ludvig Hällman. *The Rolling Window Method: Precisions of Financial Forecasting*. Degree Project, Royal Institute of Technology, School of Engineering Sciences. https://www.diva-portal.org/smash/get/diva2:1089425/FULLTEXT01.pdf. 2017.

[7] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[8] Kewei Hou. "Industry Information Diffusion and the Lead-lag Effect in Stock Returns". In: *The Review of Financial Studies* 20.4 (2007), pp. 1113–1138. ISSN: 08939454, 14657368. URL: http://www.jstor.org/stable/4494798 (visited on 05/03/2025).

[9] Ying Li et al. "Implicit-causality-exploration-enabled graph neural network for stock prediction". en. In: *Information (Basel)* 15.12 (Nov. 2024), p. 743.

[10] M. Malinda and J. Chen. "The forecasting of consumer exchange-traded funds (ETFs) via grey relational analysis (GRA) and artificial neural network (ANN)". In: *Economic Modelling* 102 (2021), pp. 61–75. DOI: 10.1007/s00181-021-02039-x.

[11] J. M. Polanco-Martínez and J. L. López-Martínez. "A non-parametric method to test the statistical significance in rolling window correlations, and applications to ecological time series". In: *Ecological Informatics* 64 (2019). https://www.sciencedirect.com/science/article/pii/S1574954121001709, p. 101379.