mathematics

MDPI

Article

Major Issues in High-Frequency Financial Data Analysis: A Survey of Solutions

Lu Zhang and Lei Hua *

Department of Statistics & Actuarial Science, Northern Illinois University, DeKalb, IL 60115, USA; lzhang3@niu.edu
* Correspondence: lhua@niu.edu

Abstract: We review recent articles that focus on the main issues identified in high-frequency financial data analysis. The issues to be addressed include nonstationarity, low signal-to-noise ratios, asynchronous data, imbalanced data, and intraday seasonality. We focus on the research articles and survey papers published since 2020 on recent developments and new ideas that address the issues, while commonly used approaches in the literature are also reviewed. The methods for addressing the issues are mainly classified into two groups: data preprocessing methods and quantitative methods. The latter include various statistical, econometric, and machine learning methods. We also provide easy-to-read charts and tables to summarize all the surveyed methods and articles.

Keywords: nonstationarity; signal-to-noise ratios; asynchronous data; imbalanced data; intraday seasonality; deep learning

MSC: 91G15; 68T07

1. Introduction

With the emergence of great computing power with deep neural networks and easy access to high-quality ultra-high-frequency financial data, the analysis of high-frequency financial data has become more popular for many researchers and practitioners. Incorporating high-frequency financial data analysis into traditional statistical and financial econometric analysis can be very helpful to better understand the financial market. In the literature, many techniques have been developed to analyze financial data that are of lower frequency, such as daily and monthly data. However, high-frequency financial data have unique characteristics, and traditional statistical, econometric, and shallow machine learning methods may not adequately suit the needs of analyzing high-frequency financial data.

In recent years, numerous studies have been conducted to address the challenges posed by high-frequency financial data. A notable trend in this area is the application of deep learning techniques to tackle critical issues such as nonstationarity and low signal-to-noise ratios. For instance, studies such as [1–3] have successfully leveraged deep learning approaches to overcome significant obstacles in forecasting high-frequency returns. Nevertheless, deep learning methods are often criticized for their inherent "black-box" nature, which complicates detailed data preprocessing and interpretability. Consequently, alternative approaches, including signal decomposition, feature engineering, sequence reconstruction, and various mathematical and statistical techniques, merit a closer examination of the processing of high-frequency financial data. Although the existing body of literature on mathematical and machine learning applications in financial data is extensive

check for updates

Academic Editor: Antonella Basso

Received: 9 December 2024
Revised: 2 January 2025
Accepted: 13 January 2025
Published: 22 January 2025

Citation: Zhang, L.; Hua, L. Major Issues in High-Frequency Financial Data Analysis: A Survey of Solutions. *Mathematics* **2025**, *13*, 347. https://doi.org/10.3390/math13030347

(see Appendix A) , a critical gap remains: no comprehensive review has systematically synthesized all available methodologies to address the core challenges associated with high-frequency financial data.

This study aims to fill this gap by organizing and synthesizing the classical and advanced methodologies documented in the literature. We provide a valuable resource for academics, industry professionals, and market participants interested in high-frequency financial data modeling and time series forecasting. By presenting a holistic perspective, this work not only highlights existing approaches but also identifies potential extensions to broader applications in time series analysis.

In our research, we used Google Scholar as the primary database due to its authoritative and comprehensive coverage. Keywords used for the search included "high-frequency financial data", "intraday financial data", and"financial data modeling", among others. To ensure the reliability and relevance of the search results, we focused on articles published in reputable peer-reviewed journals. An initial screening based on titles and abstracts yielded a preliminary dataset of 485 articles. Subsequently, we conducted a manual filtering process to ensure that the studies specifically addressed data that are sampled more frequently than daily intervals, with particular attention to critical issues in high-frequency financial data. This refinement process reduced the dataset to 165 articles. Finally, we performed a detailed review of these selected articles, categorizing them according to the major issues addressed and the proposed solutions. Detailed procedures are outlined in Figure 1.

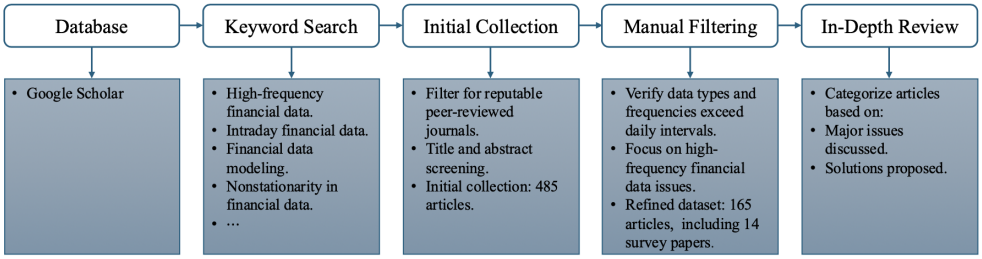

Figure 1. Flowchart depicting the process of article collection, screening, and categorization.

Overall, we have thoroughly surveyed 151 articles and 14 survey papers, mainly on forecasting using high-frequency financial data, as shown in Table A1. Figure 2 presents the distribution of articles in different years of publication. Meanwhile, it depicts the distribution of cited articles categorized by market types, primary issues, and various forecasting targets, respectively.

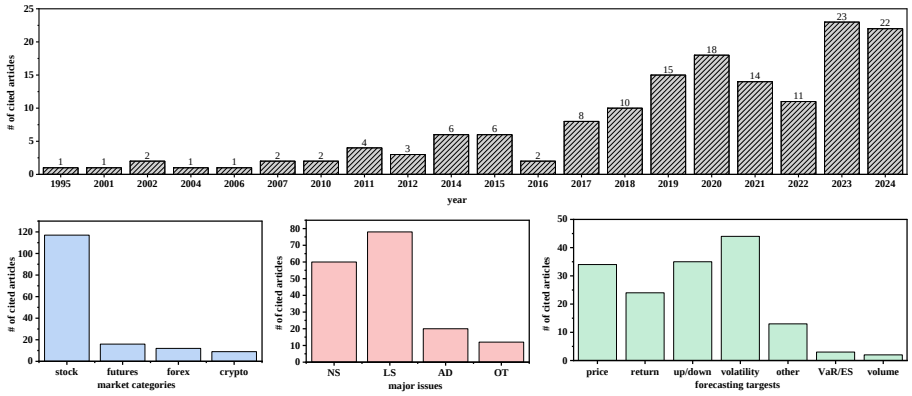

Figure 2. The distribution of referenced articles by publication year (first row). The second row illustrates the distribution of cited articles categorized by market categories, major issues, and forecasting targets presented from left to right, respectively. NS, LS, AD, and OT denote nonstationarity, low signal-to-noise ratio, asynchronous data, and other high-frequency financial data issues, respectively.

Figure 3 presents a Sankey diagram that illustrates the relationships between three key categories: major issues, data types, and solutions. The height of each flow corresponds to the relative frequency of connections between these categories, providing a clear visualization of their interdependencies. Our analysis identified 91 instances addressing the challenge of low signal-to-noise ratio (LS), 64 focusing on nonstationarity, 21 tackling asynchronous data issues, and 12 cases discussing other challenges in high-frequency financial data. Among the data types examined, trade and quote data (TAQ) emerged as the most frequently used, with approximately 152 proposed solutions leveraging TAQ-based analysis. Furthermore, the diagram highlights various solutions designed to address these major issues. The most commonly employed approaches include noise reduction techniques, feature engineering (e.g., new feature preparation, label preparation, etc.), alternative data integration, and dimensionality reduction. This comprehensive mapping underscores the critical role of TAQ data and the various strategies developed to overcome challenges in high-frequency financial data analysis.

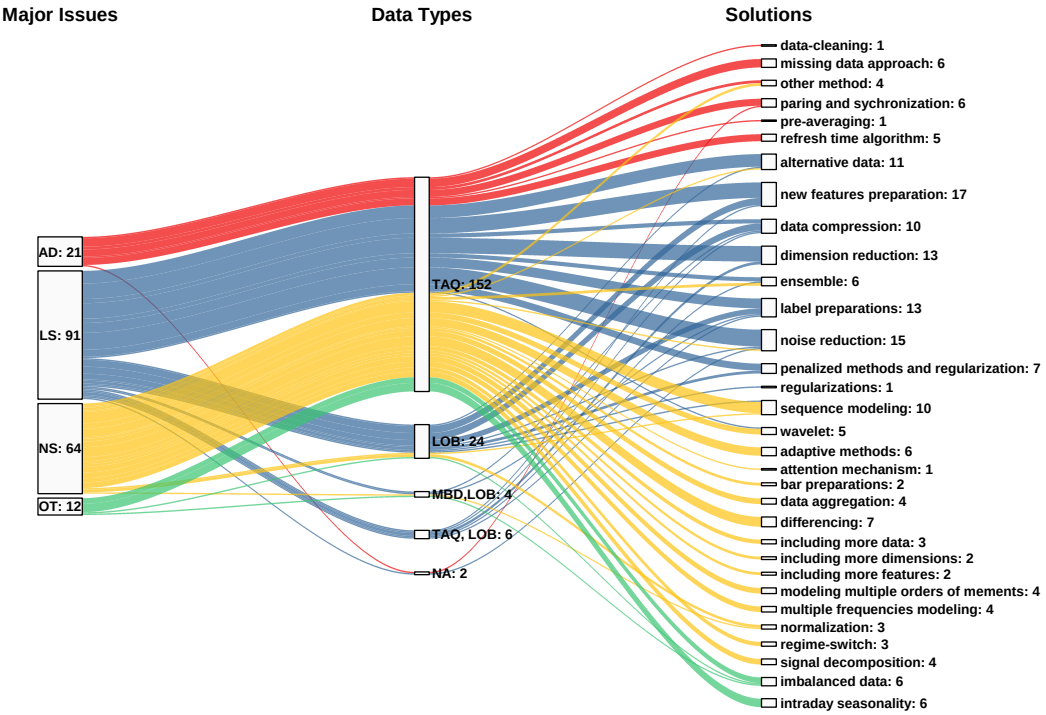

Figure 3. The Sankey diagram illustrates the relationships among major issues, data types, and solutions within the context of high-frequency financial data. The first column represents major issues, including asynchronous data (AD), low signal-to-noise ratio (LS), nonstationarity (NS), and other high-frequency financial data challenges (OT). The central column corresponds to data types, encompassing TAQ, LOB, and MBD—three common types of high-frequency financial data. Additionally, the category "NA" indicates cases where the data type was not specified in the surveyed papers. For a detailed discussion of high-frequency data types, refer to Section 2. The final column represents the solutions identified for addressing these challenges. The numbers in the diagram indicate the frequency of instances in which specific issues, data types, or solutions were encountered or applied.

The organization of the survey paper is the following. We first provide an overview of the primary types of high-frequency financial data, their distinctive characteristics, and the predominant utilization trends associated with each data type in Section 2. We then summarize the major issues for high-frequency financial data analysis and discuss the accompanying articles in Section 3. For each major issue, we conduct a comprehensive survey of the papers that have attempted to tackle the issue: Section 3.1 for the issue of nonstationarity, Section 3.2 for the issue of low signal-to-noise ratios, Section 3.3 for the

issue of asynchronous data, and Section 3.4 for other issues such as imbalanced data and intraday seasonality. Our survey on other relevant survey papers is in Section 4. Finally, we will conclude the article in Section 5 by providing some discussions on the current state of the literature and open questions that are worth further exploration.

2. High-Frequency Financial Data

High-frequency financial data refers to financial market data that are recorded and updated at short time intervals such as nanoseconds, microseconds, milliseconds, and seconds. In this survey, we only include articles that deal with financial data recorded with a higher frequency than daily data. During the whole process of trading financial assets in the modern era, numerous data have been generated such as orders placed and orders executed, and Table 1 is an example of three main types of high-frequency financial data. The following, we define types of data:

- Message book data (MBD): They are the most granular high-frequency financial data that provide a detailed record of every order event from an exchange. These events include order submission, modification, cancellation, execution, etc. Each event is timestamped at up to nanosecond level.
- Limit order book data (LOB): They consist of all outstanding buy and sell limit orders at each price level and are updated whenever there is an event that leads to a change of the limit order book. Depending on how limit orders are organized, there can be market-by-price (MBP) data or market-by-order (MBO) data [4]. Table 1 provides an example of the limit order book data with 10 levels on both the bid and the ask sides. The number of levels can be less than or more than 10, and we refer to https://lobsterdata.com/ (accessed on 20 January 2025) for details on the construction of LOB data.
- Trade and quote data (TAQ): They include information on every executed trade, including the price, volume, time of the transaction, the best bid and ask prices, and the corresponding filled volumes quantity/volumes. Table 1 provides an example of TAQ data, and the condition column indicates trade or quote conditions such as odd lots, regular, trades during extended hours, and the market opening price. The commonly used time bars with OHLCV are often constructed from these TAQ data.

According to our review of the literature, most proposed solutions utilize TAQ data, employing them in various forms such as prices (e.g., [5–7]), returns (e.g., [8–10]), volatility and covariance matrices (e.g., [11–13]), technical indicators (e.g., [14–16]), and even 2D images (e.g., [17]) such as as input features for modeling future financial variations.

Limit order book (LOB) data have emerged as a valuable resource for deep learning models. Numerous studies utilize the top ten levels of ask and bid prices and associated volumes as input features, employing convolutional layers or transformer autoencoders for automated feature extraction instead of relying on manually constructed representations. Notable examples include [1,3,18,19], among others. Another prominent manually engineered feature derived from LOB data is the order flow imbalance (OFI), a technical indicator that quantifies demand discrepancies between the bid and ask sides. Studies such as [3,20] have utilized OFIs as input features to improve the predictive accuracy of return forecasts.

For message book data (MBD) data, to the best of our knowledge, there are only two studies that incorporate MBD as an input feature to enhance modeling performance. In these studies, ref. [1,21], both employ CNN layers to extract features from MBO data. More details on the applications of these three types of high-frequency financial data can be found in Table A1.

Table 1. Example of message book data (MBD), limit order book data (LOB), and trade and quote data (TAQ).

Example of Message Book Data (MBD)							
Date	Time Stamps	Order Number	Event Type	Ticker	Price	Quantity	Excanhge
2018-06-21	04:00:00.095	3018	ADD BID	AAPL	160.00	7	ARCA
2018-06-21	04:00:00.096	3017	ADD ASK	AAPL	230.00	35	ARCA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2018-06-21	20:00:00.071	248481533	DELETE ASK	AAPL	0	0	NASDAQ
Example of Limit Order Book Data (LOB)							
Date	Time Stamps	Ask Price 1	Ask Size 1	Bid Price 1	Bid Size 1	⋯	Bid Size 10
2018-06-21	09:00:00.004	585.94	200	585.33	18	⋯	300
2018-06-21	04:00:00.001	585.94	200	585.33	18	⋯	200
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2018-06-21	15:59:59.913	577.67	300	577.54	410	⋯	100
Example of Trade and Quote Data (TAQ)							
Date	Time Stamps	Event Type	Ticker	Price	Quantity	Excanhge	Conditions
2018-06-21	04:00:00.070	QUOTE BID	AAPL	228.00	50	ARCA	00000001
2018-06-21	04:00:00.070	QUOTE ASK	AAPL	230.00	100	ARCA	00000001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2018-06-21	19:59:59.754	QUOTE ASK	AAPL	185.31	200	NASDAQ	00000001

3. Major Issues and Solutions

In this section, we conduct a comprehensive review of research articles to address common issues of analyzing high-frequency financial data. The following are the major issues covered: nonstationarity (NS), low signal-to-noise ratio (LS), asynchronous data (AD), and others (OT), including imbalanced data and intraday seasonality.

3.1. Nonstationarity (NS)

The nonstationarity issue discussed here has a loose definition, and we aim to indicate the issue as the phenomenon that any results obtained from one period of time cannot be generalized into another period of time. This is probably one of the most difficult issues with financial data analysis.

Efforts for addressing the issue can be mainly categorized into the following two approaches: one is to preprocess the raw data to have more stationary data for further analysis and the other is to use certain quantitative methods to tackle the problem directly. Here, we use the term quantitative method to indicate any method related to econometrics, statistics, machine learning, etc.

3.1.1. Data Preprocessing for Nonstationarity

Figure 4 is an overview of the data preprocessing approach to address the nonstationarity issue and their accompanying articles. The following are methods used to address high-frequency financial data analysis issues:

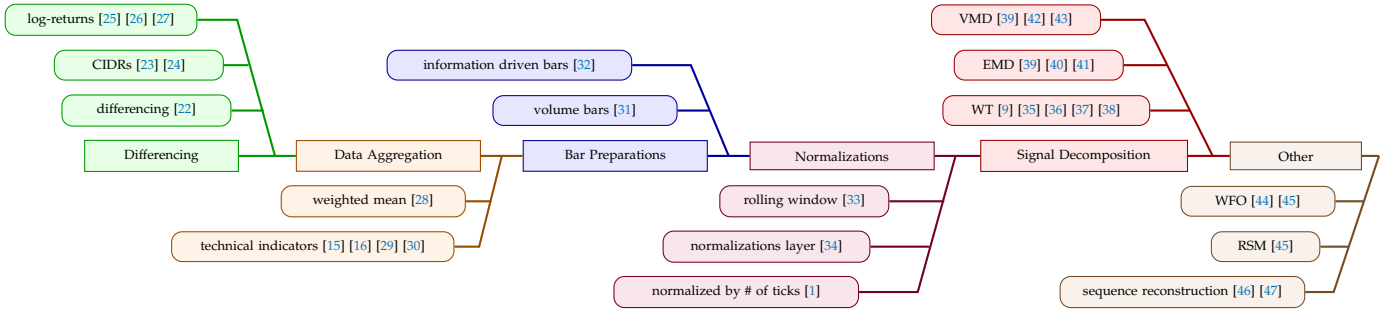

Figure 4. Data preprocessing methods for addressing nonstationarity in high-frequency financial data [1,9,15,16,22–47].

- **Differencing:** Differencing is a commonly used traditional method to address nonstationarity in financial time series. The process entails calculating the increment between two successive or nonsuccessive observations in the temporal sequence so that the increments of the time series are more stationary than the original time series. There are numerous papers on using this technique, and here we only list a very few of them as follows: [22] preprocessed the features using differencing to make them stationary before they were included in time series forecasting models. Refs. [23,24] calculated cumulative intraday returns (CIDRs) by the difference of logarithmic prices at the beginning ($\text{CIDR}_i = \ln p_i - \ln p_1$) to generate a more stationary functional time series. More commonly, returns are calculated using two successive logarithmic prices ($\text{returns}_i = \ln p_i - \ln p_{i-1}$) in predetermined time intervals (e.g., 5 min), as shown by studies such as [25–27] and others.
- **Data aggregation**: The problem of nonstationarity was addressed in [28] by aggregating the intraday sequence into a daily series using a weighted mean, and the authors suggested a linear regression approach that aggregates intraday high-frequency data into daily data. Meanwhile, they provided a data-driven technique to select the weight to achieve a reduced sum of asymptotic variance for parameter estimators. The suggested methods have the advantage of not being limited to certain parametric forms and can be produced using straightforward, constrained quadratic programming. Technical indicators play a vital role in simplifying complex market data into practical indications. Essentially, various technical indicators are simply aggregation methods that transform the price and volume data into relatively more stationary time series. Ref. [48] utilized technical indicators, which capture the trend, momentum, volume, and sentiment of the market to predict the returns of Bitcoin. Several other papers used technical indicators to aggregate information from high-frequency financial data, including [15,16,29,30], etc.
- **Bar preparations**: To apply algorithms to unstructured data, we often need to extract useful information and store it in a structured manner. Most machine learning algorithms assume that the extracted data are represented as a table. Financial professionals often refer to each row of such tables as a "bar" and each bar contains information such as open and closed prices [49]. Bars are sometimes referred to as binned data. Depending on how they are formed, there are time bars, volume bars, etc. Compared to time bars, volume bars and information-driven bars tend to be more stationary (see [49] for a discussion on different types of bars). Ref. [31] claimed that the measurement of volume-based transaction time is more relevant than the use of clock time in high-frequency trading. The authors introduced a novel "bulk volume" categorization method that combined transactions in specific time or volume intervals. Then, the standardized price difference between each interval's start and end was applied to estimate the proportion of buying and selling volume. Ref. [32]

created thirteen information-driven bars in three categories to forecast the return of stock transactions and determine the direction of the price movement. The three groups were associated with the stock's trading volume, recent trading imbalance, and measures of trading speed and cost, respectively.

- **Normalization**: Normalization is another approach that may reduce the issue of nonstationarity. Ref. [33] used data from the previous five days to normalize the input characteristics related to the bid and ask prices (size) in the data from the limit order book (LOB) to forecast fluctuations in the mid-price. Ref. [34] proposed deep adaptive input normalization layer to adaptively normalize input time series. The study conducted by [1] used market-by-order (MBO) data to generate normalized price (size) or normalized price (size) changes with 100 tick sizes, which were then used to forecast future market movements.
- **Signal decomposition:** Through frequency domain analysis, one may find more stationary patterns appearing at different frequencies. Such information can then be further used for analyzing the original financial time series. Here we survey some commonly used signal decomposition methods for financial time series analysis: Wavelet Transformation (WT), Empirical Mode Decomposition (EMD), and Variational Mode Decomposition (VMD).

  WT: The wavelet transform is a method used to decompose a time series $s(t)$ into low-frequency coefficients and high-frequency coefficients, often known as smooth and detail coefficients correspondingly [50]. For a comprehensive review of the wavelet transform, we refer to [51,52]. These coefficients were then used as features in conjunction with backpropagation neural networks to forecast stock values. Many academic researchers have integrated wavelet transformation (WT) with machine learning techniques to address the challenges posed by nonstationary and nonlinear factors in predicting high-frequency financial data. Ref. [35] investigated wavelet approximation methods and related computational algorithms to address the nonstationary issue in high-frequency financial time series. The results demonstrate that using wavelet dictionaries for data preprocessing could successfully capture concealed periodic elements, hence enabling us to achieve and enhance the capability of feature extraction. The study by [9] showed that it is possible to accurately predict future short-term stock returns using a dynamic recurrent neural network with wavelet transformed features. This network takes as input the current and lagged details and smooth coefficients obtained from a non-decimated Haar wavelet decomposition of the high and low stock prices. Refs. [36,37] introduced a deep learning architecture that integrates wavelet transformation to acquire knowledge on financial trading. Ref. [38] utilized wavelet multiresolution analysis to tackle the "calendar effect" observed in intraday high-frequency data, dissecting financial time series into various hierarchical time scales.

  EMD: Empirical Mode Decomposition (EMD) is an efficient method to analyze nonlinear and nonstationary financial time series data [39]. EMD is a flexible signal decomposition technique introduced by [53] as an essential component of the Hilbert-Huang transformation (HHT). The EMD is a data-driven approach that breaks down a signal into a residual component and a limited number of intrinsic mode functions (IMFs), each representing a distinct signal characteristic. The residual component has the ability to decrease the complexity of the time series by distinguishing trends and oscillations at various scales. This ultimately improves the accuracy of forecasting at certain time intervals [40]. Ref. [41] employed complete ensemble empirical mode decomposition alongside an adaptive noise (CEEMDAN) algorithm and the

Savitzky-Golay (SG) filter to denoise and improve the data prior to its input into deep neural networks.

VMD: Variational Mode Decomposition (VMD) is an approach that decomposes signals into a combination of oscillatory modes with well-defined instantaneous frequencies. This method is based on variational concepts and optimization approaches [42], and the objective of VMD is to identify the minimum specified cost function to extract the modes. VMD is claimed to be a better method than EMD for predicting Bitcoin prices in [39]. Both methods can decompose nonstationary and nonlinear series into stationary components known as IMFs depending on their frequency characteristics, and [39] used such IMFs as predictors and Bitcoin prices as the response variable in generalized additive models. Ref. [43] employed the Kullback-Leibler divergence optimized Variational Mode Decomposition (KL-VMD) technique to decompose crude oil price data into IMFs, subsequently applying the Fuzzy Entropy technique to classify the IMFs prior to fitting them into deep and statistical models.

- **Walk forward optimization method**: Refs. [44,45] mentioned walk forward optimization (WFO) method to alleviate the nonstationarity resulting from the data assessed under distinct regimes. The WFO algorithm trains a model using limited data points inside a designated training window. Then it evaluates the performance of the model within a separate test window. Subsequently, both windows were moved to the right, and the training and testing process was repeated.
- **Random sampling method**: Ref. [45] introduced a random sampling method (RSM) to tackle the nonstationary nature of cryptocurrency time series by sampling sequences only observed recently and assuming that the similarity of a pair of sequences can be characterized by the class, such as price movement directions, which they belong to. Specifically, given the input sequence $\mathbf{x}_t$, the RSM sampling scheme involves the sampling sequence of the closed interval $[t - m - n, t - m]$, where m is the window size of the simple moving average, and n is a hyperparameter that determines the size of this interval. The authors presented experimental information on the optimal selection of values for variables m and n in the article.
- **Sequence reconstruction**: Ref. [46] reconstructed financial time series using a set of representative motifs. Each motif represents a typical pattern of a segment of the time series. Then, the derived motifs were stitched together to reconstruct the time series. Subsequently, a convolutional neural network (CNN) model was used to extract complex and advanced characteristics from these sequences of motifs that may have improved ability to convey information from the original time series data. Ref. [47] employed the EMD with adaptive noise method to decompose the original data into multiple IMFs. Using sample entropy, they further reconstructed the IMFs into a simplified stationary high-frequency component and a low-frequency component.

3.1.2. Quantitative Methods for Nonstationarity

Figure 5 is an overview of the quantitative methods to address the nonstationarity issue and the accompanying articles.

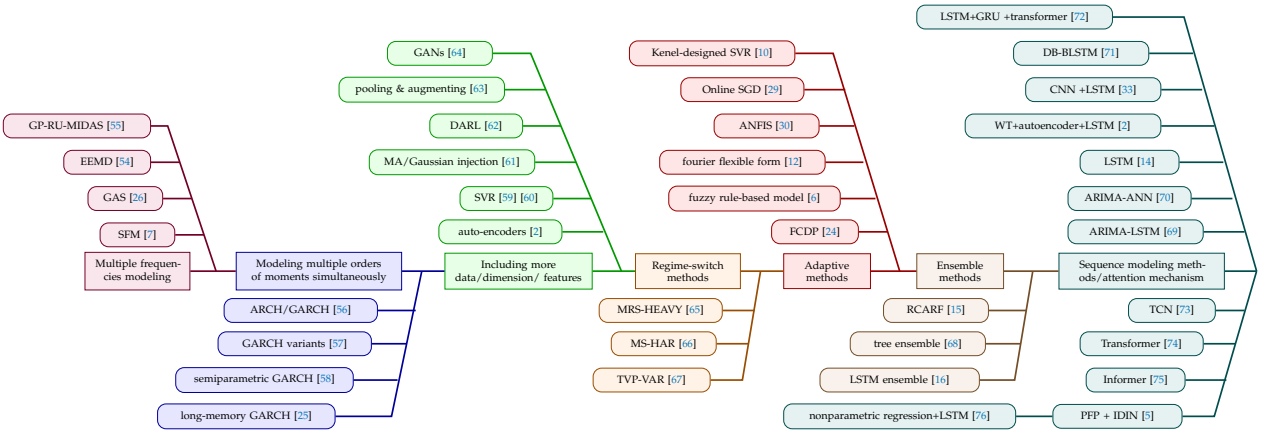

Figure 5. Quantitative methods for addressing nonstationarity in high-frequency financial data [2,5–7,10,12,14–16,24–26,29,30,33,54–76].

- **Multiple frequencies modeling**: To tackle the problem of nonstationarity and nonlinearity in stock price time series. Ref. [7] introduced a new recurrent deep neural network called State Frequency Memory (SFM). This network is designed to collect multifrequency trading patterns from historical market data, enabling it to generate accurate long- and short-term forecasts over time. SFM is similar to LSTM, but it decomposes the hidden states of the memory cells into several frequency components. Each component represents a certain frequency of trading patterns that influence stock price fluctuations. Future stock prices are then forecasted by applying an inverted Fourier transform (IFT) to a composite of these components. Ref. [26] introduced a dynamic coupla-based periodic mixed frequency generalized autoregressive (GAS) framework to forecast intraday risk measurements. The suggested GAS models overcome the nonstationarity issue by accounting for many forms of dependency in asset returns, such as long memory, periodicity, asymmetric nonlinear dependency structures, fat-tailed conditional distributions, and intraday jump processes. Ref. [54] utilized EEMD for sequence decomposition, calculating fuzzy entropy values for the decomposed components, which facilitates their reconstruction into high-frequency, medium-frequency, and low-frequency components. Ref. [55] introduced the group penalized reverse unrestricted mixed data sampling model (GP-RU-MIDAS) to efficiently leverage high-dimensional and mixed-frequency data from various sources to explain and predict high-frequency responsive variables.
- **Modeling multiple orders of moments simultaneously**: Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models provide great flexibility in representing higher moments, such as variance. This characteristic makes them well-suited for analyzing nonstationary time series data. The distribution of zeros often causes financial returns to be nonstationary. Ref. [56] introduced an ARCH/GARCHmodel that can handle nonstationary zero processes. They developed a zero-adjusted quasi-maximum likelihood method to estimate the model's parameters and demonstrated its consistency and asymptotic normality under mild assumptions. Ref. [57] proposed various GARCH models on log-returns to detect intraday volatility spillover in cryptocurrencies. Ref. [58] introduced a semiparametric scaling model to consider high-frequency financial data information by normalizing the trading day into a unit time interval. Then, to adequately use market information, a semiparametric GARCH model was proposed to improve the estimation of volatility functions by combining them with the various frequency volatility proxies. Ref. [25] presented extended periodic long-memory filters based on Gegenbauer polynomials, a proper mathematical technique for capturing repetitive patterns, to describe the time-varying

volatility of typical GARCH models. The suggested long-memory GARCH models can detect periodic long-range dependencies in the fluctuating volatility of intraday financial returns.

- **Including more data/dimensions/features**: Augmenting data or including more data or features can enrich the representation of the underlying data-generating process, making the model more flexible and capable of capturing a more comprehensive range of patterns and relationships. Thus, such data preparations can help models better adapt to changing conditions and mitigate the effect of nonstationarity. Ref. [2] used stacked autoencoders to systematically produce more abstract and high-level features hierarchically. Refs. [59,60] used linear, radial, and polynomial kernels of support vector regression (SVR) to transform the data into various dimensional feature sequences in order to capture the linear and nonlinear correlations. Ref. [61] applied exponential moving average and Gaussian noise injection, two data augmentation techniques, to augment high-frequency financial time series data by considering the high SNR and momentum effect. The approach described by [62], referred to as data-augmentation-based reinforcement learning (DARL), utilizes minute-candle data instead of daily-candle data to train the agent. Subsequently, the agent is employed to direct daily stock trading. Ref. [63] demonstrated that pooling stock data and augmenting predictor space by adding market volatility as an additional predictor often enhances the out-of-sample performance for most investigated ML models. Ref. [64] introduced a technique to augment data using generative adversarial networks (GANs) in artificial market simulations. This approach aims to address the challenge of nonstationarity in financial time series while training machine learning models.
- **Regime-switch methods**: Nonstationarity in time series data is often caused by structural interruptions or changes in the underlying data generation process. Regime-based approaches are ideal for detecting such breakdowns because they partition the data into multiple regimes, minimizing the impact of nonstationarity. Ref. [65] proposed a multivariate Markov regime-switch high-frequency-based volatility model (MRS-HEAVY) to predict the covariance structures of returns of spot and futures prices. The regime-switch characteristics allow us to handle abrupt changes in the covariance structure of the spot and futures markets during periods of market volatility. Ref. [66] proposed Markov switching heterogeneous autoregressive model (MS-HAR) to forecast the volatility of high-frequency crude oil futures. In addition to capturing regime transitions, modeling time-varying characteristics may also solve the problem of nonstationarity. Ref. [67] utilized the time-varying parameter vector autoregressions (TVP-VAR) approach to capture the time-varying dynamics dependencies among energy and other ETFs' intraday volatility. The significant advantage of the TVP-VAR model is that there is no need for arbitrariness when choosing the rolling window size. Besides, regime-based approaches can alter regimes to respond to changing data circumstances. This flexibility enables the model to alter its predictions to data structure changes, thus addressing the nonstationarity issue.
- **Adaptive methods**: Adaptive methods are designed to constantly update and modify their parameters based on new data, enabling them to capture and adapt to changes in high-frequency financial data. This real-time adjustment capability enables adaptive techniques to remain responsive to nonstationarity. Ref. [24] proposed functional classification and dynamic prediction (FCDP) techniques to predict the cumulative intraday returns of crude oil futures. The aim was to account for the unique features of high-frequency data and the intricate patterns of change in the crude oil futures market. The advantage of FCDP is seen in its ability to update dynamically. This

model continuously incorporates new data into the prediction model, assigning more weight to the most recent data. This feature is beneficial in improving the accuracy of the forecast outcomes. Ref. [6] proposed an online evolving fuzzy rule-based model to predict the trend of high-frequency financial data. This approach autonomously constructs and eliminates its rules, adapts its parameters, and chooses inputs acquiring knowledge from the data collected from previous and current scenarios. Despite the frequent occurrence of abrupt changes in high-frequency trading financial data, the suggested model has the ability to quickly adjust and align with the new trend. Ref. [12] introduced a novel framework for representing trends and cyclical patterns in high-frequency financial data. To better respond to constantly changing market circumstances, the Fourier flexible form was extended to a more advanced functional class, where the smooth trend and seasonality are allowed to change in real time and modeled nonparametrically. The authors provide the corresponding estimators and demonstrate that the suggested model performs well when there are sudden changes, as well as when there are unequal variabilities and heavy tails in the data. Ref. [30] improved the precision of stock price prediction by optimizing the parameters of an adaptive neuro-fuzzy inference system (ANFIS), and both the genetic algorithm and the particle swarm optimization algorithms were studied. The approach described by [29] aims to address the instability of financial time series, such as seasonal and irregular patterns, by using an online stochastic gradient descent algorithm (SGD) to identify seasonal patterns and flexible sliding windows that are updated based on detected distribution changes using stationarity. Ref. [10] introduced a new kernel-designed SVR, which used decaying cosine waves, to account for the decaying trend reversal process triggered by news.

- **Ensemble methods**: Ensemble methods have several advantages, such as the capacity of reducing model variance and misspecification, robustness, adaptation to changing environments. By leveraging these advantages, ensemble methods can improve predictions' accuracy, stability, and robustness under the nonstationarity and dynamic market conditions. Ref. [15] proposed recurring concepts adaptive random forests (RCARF) to enhance the accuracy of predicting stock prices that exhibit cyclical patterns and nonstationarity. This method improves the functionality of random forest to adapt to changing data streams. It incorporates a way to maintain and manage a shared library of inactive trees, which retains information on how market operators responded in similar situations. Ref. [68] used a decision tree-based ensemble technique to forecast the direction of stock closing prices. The authors highlighted that the gradient-boosting decision tree is particularly effective in handling information obtained from intraday and interday input characteristics. Ref. [16] introduced an ensemble of LSTM neural networks for the purpose of predicting the classification of the stock market. The suggested ensembles function assigns weights to individual models based on their recent performance, and enables us to effectively handle nonstationarity in a creative manner.
- **Sequence modeling methods/attention mechanisms**: High-frequency financial data are often treated as a sequence of data along time and thus many traditional and modern deep learning methods developed for sequence data are suitable for high-frequency financial data. Ref. [69] integrated the conventional ARIMA model with the deep learning model LSTM to predict high-frequency financial time series. The suggested model addresses the nonstationary problem and the ability to describe the nonlinear dependency of the error terms. Ref. [70] showed that the hybrid ARIMA-ANN model outperforms individual models such as ARIMA and ANN in terms of prediction accuracy. LSTM networks are a valuable tool for modeling and predicting

high-frequency financial data of nonstationary behavior in the area of deep learning. Ref. [14] examined the use of LSTM networks in forecasting future trends of stock prices by analyzing historical price data and technical indicators. Ref. [2] integrated wavelet transforms, stacked autoencoders, and LSTM to predict stock prices. Initially, the raw data undergo a wavelet transform. Subsequently, the denoised data are fed into stacked autoencoders to provide high-order features, and finally, LSTM networks are used for forecasting. Ref. [33] constructed a sophisticated deep learning model on a wide scale to forecast price fluctuations using the limit order book data. The design employs convolutional filters to capture the spatial structure of the limit order books and LSTM modules to capture temporal dependencies. Ref. [71] proposed a new hybrid wavelet deep learning model, known as Daubechies wavelet with the bidirectional LSTM model (DB-BLSTM), to address the challenges caused by the complicated periodicity and nonlinearity of high-frequency data. To tackle the issues of nonstationarity, nonlinearity, and low SNR in high-frequency stock trading to enable accurate and super-fast forecasting, ref. [72] developed online hybrid neural networks by integrating LSTM, Gate Recurrent Unit (GRU), and transformers. A temporal convolution network (TCN) is an effective tool for modeling sequence data, as it has shown the capability to retain a longer and sufficient memory. The study by [73] used a TCN to predict intraday trading directions using financial sequence data. The authors concluded that the TCN outperformed traditional recurrent neural networks such as LSTMs. Recent studies have shown the potential of transformer and informer-based models in forecasting long-sequence time series data [77]. The attention mechanism on both methods can effectively capture long-range connections between input and output to enhance prediction capability and therefore tackle the nonstationarity problem in modeling high-frequency financial data. For example, ref. [74] leveraged the transformer to predict the OHLC price of Shanghai crude oil futures. Ref. [75] utilized an informer based on the attention mechanism to predict the intraday stock price. Empirical results indicated that the performance of the informer-based model was significantly better than LSTM. Ref. [5] incorporated a phase fix procedure (PFP) to predict high-frequency financial time series. The main idea is to retrain the model by feeding back the predictions from the previous prediction. A hybrid forecasting model termed the increasing–decreasing linear neuron (IDIN) was developed to account for linear and nonlinear components in price processes. Ref. [76] utilized a kernel-based nonparametric regression model to capture nonlinear relationships between a nonstationary target series and its lags. An LSTM model was then applied to capture high volatility and frequent fluctuations. Empirical results showed that this hybrid approach significantly improved forecasting accuracy.

3.2. Low Signal-to-Noise Ratios (LS)

We use the term "low signal-to-noise ratios (SNRs)" of high-frequency financial data to actually mean that high-frequency financial returns, asset prices, and price movement directions are very hard to predict if they are at all predictable. In short, the signal here means factors that have some predictable power. To address the issue of low SNRs in high-frequency financial data while improving predictability, we discuss some commonly used and new methods below.

3.2.1. Data Preprocessing for Low Signal-to-Noise Ratios

In this section, we focus on techniques that are particularly helpful in improving signals and predictability while reducing noise from high-frequency financial time series,

although the data preprocessing techniques discussed in Section 3.1.1 on nonstationarity may also be relevant to address the issue of low SNRs.

Figure 6 is an overview of the data preprocessing approach to address the problem of low signal-to-noise ratios and the accompanying articles. The following describe methods to address low signal-to-noise ratios:

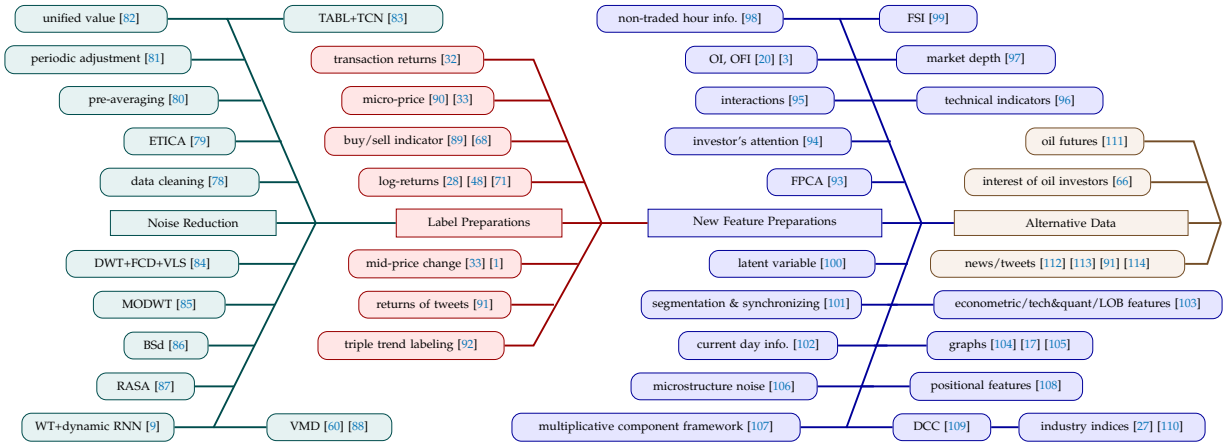

Figure 6. Data preprocessing methods for addressing low signal-to-noise ratios in high-frequency financial data [1,3,9,17,20,27,28,32,33,48,60,66,68,71,78–114].

- **Noise reduction**: Noise reduction may mitigate the low SNR problem by selectively eliminating random variations and isolating the more significant patterns or signals. For instance, data cleaning procedures, filtering algorithms, and sophisticated quantitative methods are valuable tools for identifying and removing noise components. Ref. [78] examined several data handling techniques for high-frequency financial data generated by the New York Stock Exchange. The author also detailed the reasoning behind these suggested algorithms and discussed their potential consequences. Subsequently, the authors provided precise methodologies for generating the time series of interest from the raw data. Ref. [79] assumed that the returns of each stock have two components: returns due to its own and returns from the market (i.e., $r = r_{\text{stock}} + r_{\text{market}}$), and they used External Trend and Internal Component Analysis (ETICA), a signal decomposition method based on independent component analysis, to distinguish the individual trend of the stock from the overall trend of the market, wherein they and claimed that r_{stock} is more predictable. Ref. [80] suggested that the preaveraging technique, which defines new observations by averaging latent observations on previous predetermined time intervals (see [115]) to estimate volatility in the presence of noise, offers a straightforward way to eliminate microstructure noise. This approach can mitigate the edge effect (see [115,116]) and is also robust to jumps after suitable operations. The study conducted by [81] used an intraday periodic adjustment approach, based on point process theory and established by [117], to mitigate the influence of market microstructure noise. Ref. [82] suggested the conversion of the returns from both long and short positions into a unified value for the prediction model so that it will filter the spike return while preserving the relative magnitude of the returns from the long and short positions. Ref. [83] developed a feature extractor layer that combines a temporal attention-augmented bilinear layer (TABL) and TCN modules to extract features and remove noise from the data. Ref. [9] demonstrated that a dynamic recurrent neural network is useful in predicting future stock returns. This network takes as input the current and lagged details, as well as the smoothed coefficients obtained from a nondecimated Haar wavelet decomposition of the high and low stock prices. Mathematical transformations, including the Fourier transfor-

mation, wavelets, and model decomposition, are often used as methods to remove noise from financial time series. The authors of [84] used a mix of discrete wavelet transformation (DWT), filter cycle decomposition (FCD), and variable length sampling (VLS) to eliminate noise in raw financial time series data. Ref. [85] proposed a novel nonlinear filtering algorithm based on the linear maximal overlap discrete wavelet transform (MODWT) to denoise high-frequency financial data. Ref. [86] utilized BSd (B-spline wavelet of high-order d) variant to forecast high-frequency financial time series. BSd has the ability to break down noisy time series into many smooth signals. The efficacy of implementing BSd is shown by contrasting it with other conventional models, such as the Harr and Daubechies wavelet transforms. Ref. [87] proposed a recurrently adaptive separation algorithm (RASA), which is based on the MODWT, to filter out noise caused by abnormality, irregularity, and heterogeneity in financial data. Variational mode decomposition (VMD) is a technique that eliminates noises and random fluctuations in price data by breaking the data down into variational model functions.Ref. [60] suggested combining VMD with technical analysis to anticipate Bitcoin prices. Ref. [88] introduced ensemble empirical model decomposition with adaptive noise and intrinsic sample entropy methodology to remove noise, thus improving the predictability of financial time series.

- **Label preparations**: The process of label preparation for high-frequency financial data is to identify those targets that are relatively more predictable facing the issue of low SNRs. A commonly used approach is to directly use the log returns, as shown by [28,48,71], among others. In addition, we provide a comprehensive list of additional techniques for target preparation in what follows.
  The buy or sell indicators were created by [89] based on the desired risk–reward ratio and the comparison of the price with a predefined risk threshold. The concept of microprice, as described by [90], is derived from the mid-price and incorporates information about the imbalance and bid–ask spread in limit order book data. According to the author's argument, the use of high-frequency data to estimate microprice is a more accurate predictor of short-term pricing compared to the mid-price or the weighted mid-price. For example, ref. [33] used microprice to construct feature maps to predict price movements using data from the limit order book. Ref. [32] proposed a definition of transaction returns based on the average return of transactions in a forward window to reduce noise, and they defined the directions of price movement by comparing such transaction returns with the average past returns. Refs. [1,33,83] used a mid-price change, together with a threshold value, to generate a label of market movement. Ref. [68] utilized a predetermined buy/sell threshold to compare anticipated values and determine whether to purchase, sell, or maintain holdings. Ref. [103] generated category labels by comparing the ratio of the average of the future mid-price and the present mid-price.
- **New feature preparations**: By introducing new features into the modeling process, we can improve the low SNR in high-frequency financial data, thereby improving the accuracy and robustness of forecasting models. Ref. [93] proposed three novel modeling strategies: utilizing information from usually discarded data, combining sampling and ensemble methods, and using functional principle component analysis (FPCA) to incorporate long-term historical data for processing raw data and preparing input features. Ref. [94] developed additional variables, such as the investor's attention factor, in conjunction with temporal convolutional networks (TCNs) to forecast intraday volatility. Ref. [95] extracted interactions from input characteristics to create new features and used a sub-industry index as supplementary information to forecast movements of stock prices. Ref. [96] used candlestick data and technical

indicators as input characteristics to predict the movement of the stock price using a deep neural network model. Ref. [20] developed the order imbalance (OI) indicator and the order flow imbalance (OFI) indicator as tools to improve the accuracy of future return predictions. Ref. [3] employed the OFI to forecast high-frequency returns using deep learning methodologies. Ref. [97] argued that the depth of the market has an essential impact on pricing in the limited order book market. Ref. [98] mentioned that the inclusion of non-traded hour information significantly influences the accuracy of volatility forecasts. Ref. [99] used financial stress indices (FSI) as additional variables to predict volatility in oil prices. Ref. [100] utilized latent efficient returns instead of observed returns to model high-frequency covariance dynamics to avoid market noise. Ref. [101] proposed a novel feature identification scheme to obtain trades and LOB features using time series segmentation and synchronizing the state of the order book. Ref. [103] provided a comprehensive collection of manually designed features, including econometric features, Tech and Quant features, and LOB features. These features were used to predict mid-price movements. In addition, they performed an experimental assessment that included directly comparing their proposed features with other manually designed features mentioned in previous research, as well as features obtained from an LSTM autoencoder. Unlike most studies that use historical data only as input features, ref. [102] examined the use of the first 3/5/10 min of stock price data from the current day and stock data from the previous day to forecast the movement of stock prices on the current day. Ref. [104] first established a correlation graph among stocks using a multi-Hawkes process. They then used an attention-based LSTM to learn the weighting matrix that underlies the dynamic graph. Ref. [106] showed that including microstructure noise in the two scales realized volatility significantly enhances the accuracy of volatility estimators. Ref. [17] transformed historical volatilities data into two-dimensional pictures and then used a filterbank CNN that includes 15 kinds of domain knowledge-based filters to extract superior features from the resulting images. Likewise, ref. [105] transformed LOB data into a sequence of standard images represented as 2D matrices and forecasted mid-price fluctuations using an image-based CNN. The multiplicative component framework was established by [107] in their study. This method takes advantage of the robust correlation between volatility and trade volume. They used a state-space method to incorporate a prior framework for modeling and predicting volatility. Empirical findings have shown that the suggested approach surpasses conventional models by offering more extensive information and achieving superior accuracy in predicting intraday volatility. Ref. [108] introduced positional features such as remaining time in the current trading day and the agent's current position, to collect contextual information, hence enhancing the overall performance of deep reinforcement learning in intraday trading. Ref. [109] demonstrated that dilated causal convolutional filters (DCCs) excel at extracting critical features from intraday financial time series, and this architecture effectively harnesses the predictive information embedded in high-frequency data for predicting day-ahead volatility. Refs. [27,110] employed U.S. industry indices as novel attributes to forecast stock market volatility and established that the majority of industries provide valuable insights for estimating future aggregate market volatility and returns.

- **Alternative data**: Integrating alternative data in high-frequency financial data analysis may effectively address the issues caused by low SNRs. We conducted an extensive analysis of the alternative data used to solve this problem. Ref. [66] showcased that the level of interest of oil investors, evaluated by the amount of Google searches, provides additional valuable information to predict the volatility of crude oil futures. The

study by [111] suggested that the high-frequency crude oil futures data had valuable information to forecast volatility in the US stock market. Ref. [112] showed that both macroeconomic and firm-specific news had a substantial impact on the volatility levels of intraday returns. Similarly, ref. [113] showed that macroeconomic news affects the returns and volatility of the indices. Ref. [91] designed a machine learning system to extract useful features from tweets as alternative information to predict stock price movements. Ref. [114] used public sentiment on Twitter to forecast the return and volumes of cryptocurrency prices.

3.2.2. Quantitative Methods for Low Signal-to-Noise Ratios

Figure 7 is an overview of quantitative methods to address the issue of low signal-to-noise ratios and their accompanying articles. Following that figure, we describe more methods:

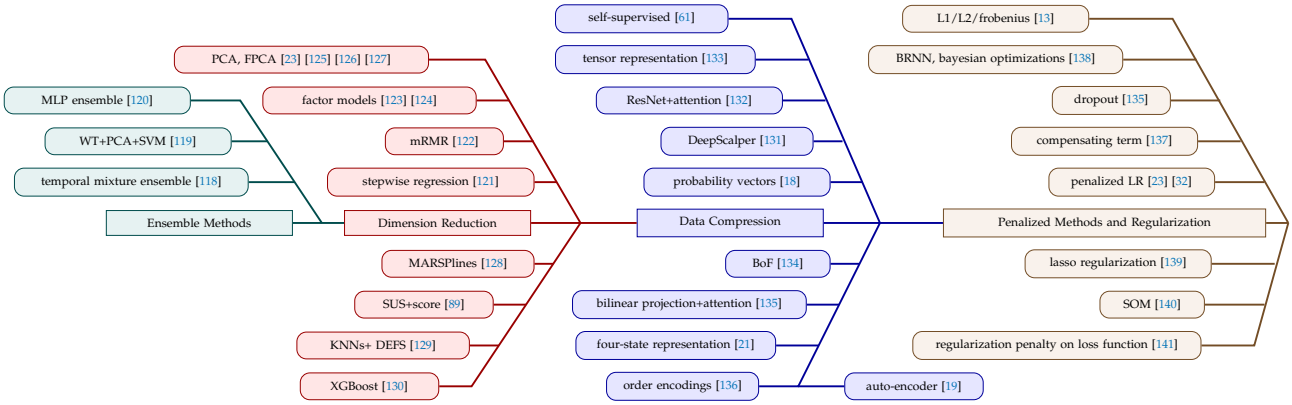

Figure 7. Quantitative methods for addressing low signal-to-noise ratio in high-frequency financial data [13,18,19,21,23,32,61,89,118–141].

- **Ensemble methods**: Ensemble techniques are well suited for tackling the low SNR problem in high-frequency financial fields by combining the knowledge of numerous weak learners to construct a more powerful and resilient prediction model. In addition to the ensemble models discussed in the nonstationarity section, we also include some additional models. For example, ref. [118] proposed a temporal mixture ensemble that is capable of adaptively using information to predict both the volume point estimate and its uncertainty. Ref. [119] introduced an ensemble method that incorporates wavelet analysis, phase space reconstruction technique, independent component analysis, and support vector machine to denoise high-frequency financial data. Ref. [120] created a binary ensemble classifier using two multilayer perception (MLP) frameworks to predict the market-making process.
- **Dimension reduction**: Dimension reduction approaches may mitigate the problem of low SNRs by identifying the most relevant features of the dataset. Ref. [121] utilized stepwise regression to determine the most important features to predict the stock price. Based on the principle that an optimal collection of features should be highly informative while minimally redundant, ref. [122] combined the minimum redundancy-maximum relevance (mRMR) feature selection criteria to assess the information included in the individual features after eliminating redundancy. Refs. [23,125,127] used (functional) principal component analysis to reduce the dimensionality of high-frequency financial data. Ref. [126] proposed a high-frequency principle component analysis (PCA) method that uses a correlation matrix to handle jumps, microstructure noise, and asynchronous observation periods in a robust manner. Ref. [123] used a matrix-valued factor model to decrease the dimensionality of high-dimensional functional time series. This was achieved using latent factors to

represent the information contained in the data. To fight the curse of dimensionality, ref. [124] proposed three specifications for factor models. They chose to use time series regression (TSR) or cross-sectional regression (CSR) to determine factor loadings, a correlation between observed and latent variables, depending on whether the factors are known or unknown. Alternatively, PCA was used if both factors and their loadings were unknown. Ref. [128] utilized multivariate adaptive regression splines (MARSPlines) to identify the key factors that influence crude oil prices. Ref. [89] gave a numerical score to each feature based on the measurement of autocorrelations and mutual information. Subsequently, they used a technique known as stochastic universal sampling (SUS) to choose features based on their respective scores, with a higher score indicating a higher probability of being selected. Ref. [129] introduced an innovative hybrid method to reduce dimensions. First, they utilized Pearson's correlation as a simple filter to remove linearly dependent features. Then, they combined proposed classifiers such as the *K*-nearest neighbor classifier (KNN) and Fuzzy KNN with differential evolution feature selection (DEFS), a method to determine the best parameters for each classifier proposed by [142], to select the most relevant features to predict intraday returns. Ref. [130] employed XGBoost for feature selection and integrated it with a deep learning approach to forecast price spread in high-frequency pairs trading.

- **Data compression**: Data compression is a method used to decrease the size of data by encoding them in more efficient forms while maintaining crucial information. This includes eliminating noise, extracting significant features, and reducing computational complexity. Thus, apart from the papers specifically addressing "dimension reduction" and "filter, noise reduction", this collection includes other works that use some techniques of data compression. Ref. [18] introduced a novel approach to transform raw LOB data into probability vectors. This conversion allows for compressing raw tick data while preserving crucial information necessary for modeling neural networks. The DeepScalper architecture, a reinforcement learning method proposed by [131], aims to capture short-lived intraday trading opportunities. An innovative aspect of their study is the use of the encoder approach to generate market representations. These representations are derived from micro embeddings, which include historical limit order book data and private observations, as well as macro embeddings, which include OHLCV data and technical indicators. The technique suggested by [132] integrates ResNet [143] with the variable-wise attention mechanism to address multicollinearity, resource constraints on computing hardware, and the problem of vanishing gradients in RNN series while analyzing complicated high-frequency financial data. Ref. [133] used tensor representations to compress the LOB data and introduced multilinear methods to predict the directions of the mid-price movement. Ref. [61] proposed a technique called self-supervised learning to extract feature representations from unlabeled financial data using data augmentation and neural network-based encoders. This method aims to overcome the bias caused by manually crafted features. Ref. [134] proposed a novel temporal-aware neural bag-of-features (BoF) model to convert 144 value feature vectors into a temporal BoF (T-BoF) histogram with only 32 values. The proposed model used the radial basis function and accumulation layers to capture both the short- and long-term dynamics of time series. Ref. [135] integrated the concept of bilinear projection and an attention mechanism to enhance the layer's ability to identify and concentrate on important temporal information to predict changes in financial time series. Initially, an attention technique was utilized to erase the impact of unimportant features. The authors then extracted higher-level feature representations after the bias shift and nonlinearity trans-

formation through bilinear projections. Ref. [21] utilized four-state representation to cover the inherent information of the market microstructures of the limit order books. Ref. [136] used order encoding methods to create input embeddings that represent all relevant features (e.g., market type, price, time, etc.) to predict the movement of the mid-price. In their experiment, convolutional neural networks (CNNs) combined with average pooling techniques were proposed to improve the prediction power. Ref. [19] utilized a bottlenecked transformer autoencoder to acquire concise and informative representations of limit order book patterns for subsequent anomaly identification in high-frequency markets.

- **Penalized methods and regularization**: Penalized approaches and regularization techniques are useful in mitigating the problem of low SNRs in high-frequency financial data by managing the complexity of the models and avoiding overfitting. Ref. [23] used penalized least squares and function-on-function linear regression as dynamic approaches to improve the accuracy of predicting cumulative intraday returns (CIDRs). Ref. [32] utilized penalized linear regression to reduce the impact of less informative variables by pushing their coefficients toward zero when forecasting high-frequency stock returns. Ref. [137] introduced a compensating term in the objective function to penalize data on the wrong side of the hyperplane when applying SVM to analyze the dynamics of the limit order book. Ref. [135] combined dropout and max norm as regularization techniques to improve the generalization capacity of networks in analyzing the data from the limit order book. The study conducted by [138] assessed the performance of different artificial intelligence models, such as Support Vector Regression (SVR), Gaussian Poisson regression (GPR), regression tree (RT), *K*-nearest neighbors (KNN), Feedforward Neural Network (FFNN), Bayesian Regularized Neural Network (BRNN), and Radial Basis Function Neural Network (RBFNN), to predict high-frequency Bitcoin price series. Bayesian optimizations were utilized to fine-tune key parameters in the proposed models. Ref. [13] proposed regularization approach (L_1, L_2, and Frobenius) to estimate the volatility matrix of high-frequency stock returns with the aim of achieving lower prediction errors. The study conducted by [139] used the Lasso method to select important features. Ref. [140] proposed a hybrid self-organizing map (SOM), which uses *K*-means clustering to cluster stocks and uses SVR to predict future price and volatility. The regularization term was applied in SVR to control the function's capacity. Ref. [141] showcased the capacity of RNN to capture the nonlinear correlation between short-term price movements and a spatio-temporal representation of the limit order book. The regularization penalty was applied on the loss function to find the optimal parameters.

3.3. Asynchronous Data (AD)

High-frequency financial data are recorded in irregularly spaced time. However, most well-developed methods are created for regularly temporally spaced and synchronous data. There are different ways to handle the common problem of asynchronous high-frequency financial data. The commonly used method is to specify a sequence of evenly spaced time points such as each minute and then use the available data that are right before the specified time point as the data for this specified time point. Although easy to handle, this method may not accurately reflect the actual time at which the trades were executed. There are several methods studied in the literature, and we summarize them as follows.

Figure 8 is an overview of the data preprocessing and quantitative approaches to address the issue of asynchronous data and their accompanying articles.

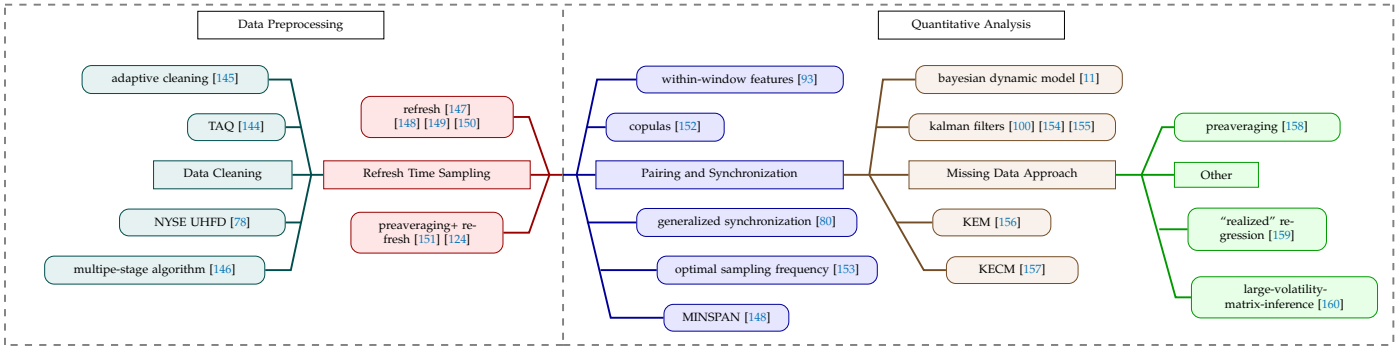

Figure 8. Data preprocessing and quantitative methods for addressing asynchronous issue in high-frequency financial data [11,78,80,93,100,124,144–160].

3.3.1. Data Preprocessing for Asynchronous Data

- **Data cleaning:** Ref. [144] has a detailed procedure for cleaning TAQ data. Ref. [145] introduced the adaptive cleaning method for filtering high-frequency financial data. This method employs a sequential approach to clean a time series and continually update its knowledge base in real time while simultaneously learning from the data. Ref. [78] examined the challenges related to gathering and managing high-frequency financial data from the NYSE. The authors illustrated typical issues associated with data errors and explained how they can be addressed using fundamental outlier identification techniques. Ref. [146] developed a multiple-stage algorithm for identifying and removing outliers in high-frequency financial data.
- **Refresh time sampling:** The algorithm was first proposed by [148] and included in the R package [149] highfrequency 1.0.1. The concept involves coordinating nonsynchronized transactions for many securities by selecting the most recent trades in each security after new trades appeared in all stocks. This approach successfully deals with the Epps effect [161], where the sample correlation tends to exhibit a significant bias toward zero as the sampling interval decreases. Following [149], we illustrate how the refresh time sampling algorithm works in Figure 9. Ref. [147] utilized the refresh time algorithm to synchronize irregularly spaced time series data. As an extension of refresh time sampling, ref. [150] proposed the "pairwise-refresh" and "all-refresh" schemes to handle nonsynchronized trading. Refs. [124,151] used the preaveraging approach with a refresh time to address the issue of microstructure asynchrony.

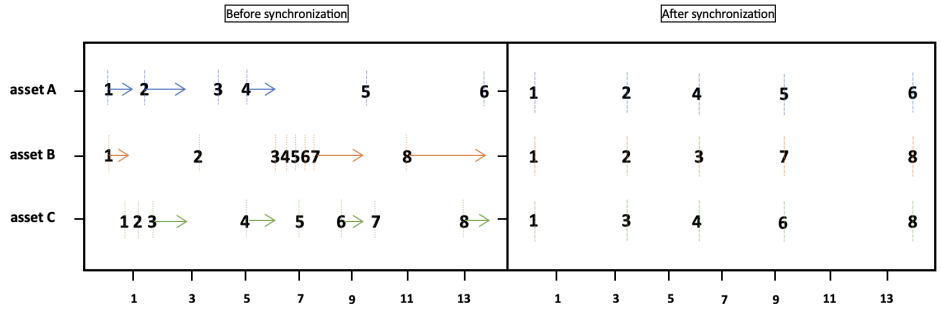

Figure 9. Refresh time algorithm visualization. The x axis denotes the time interval. The dashed vertical lines reflect the occurrence of trading in assets A, B, and C. The numbers indicate the sequence of transactions, with 1 being the initial trade, 2 the second trade, and so forth. The right arrows show the timestamp at which the observations are shifted.

3.3.2. Quantitative Methods for Asynchronous Data

- **Pairing and synchronization methods**: Ref. [152] addressed the issue of asynchronicity on several measures of association using copulas. The pairing method used in

the paper is similar to the refresh time sampling method, and, in addition, the trading time of the reference ticker is kept in their method. The paper proposed some improved estimators for the dependence measures of asynchronous financial data and shows that ignoring the asynchronous issue may generate serious problems in assessing the dependence in terms of correlations and Kendall's τ. Ref. [93] employed within-window features to enrich the information contained in binned data while reducing data synchronicity issues. The Generalized Synchronization approach suggested by [162] aims to synchronize high-frequency financial data to avoid the Epps effect. This approach is based on the concept of generalized sampling time, in which an arbitrary observation is selected within a certain time period. Ref. [80] implemented the generalized synchronization approach to address the issue of nonsynchronicity. Ref. [153] provided the optimal sampling frequency that achieves a balance between bias and other kinds of random errors, such as asynchronous trading, microstructure noise, and temporal discretization. Ref. [148] used the minimizing the tuple span technique (MINSPAN), a data collection approach aimed at reducing the tuple span, to improve the synchronization of financial data.

- **Missing data approach**: Refs. [100,154] treated the asynchronicity problem as a missing data problem that can be easily handled by Kalman filters. The benefit of the method is that the data are not partially discarded; therefore, all the information in the data is kept. Ref. [155] applied the Kalman-filter-filling method to analyze emerging markets stocks. Ref. [11] adopted a Bayesian dynamic linear model in which asynchronous trading and microstructure noises are interpreted as missing observations and measurement errors, respectively, in an otherwise synchronous series. Ref. [156] modeled noisy and asynchronous high-frequency asset prices in a state-space framework with missing data. Then, a covariance matrix of the latent states was subsequently estimated utilizing the Kalman smoother and expectation maximization (KEM) method. Ref. [157] extended the KEM algorithm to price models that include jumps. They developed a Kalman Expectation Conditional Maximization (KECM) algorithm to estimate the unknown covariance and identify jumps.
- **Preaveraging**: Preaveraging refers to calculating the weighted average of returns within a predetermined time window. Ref. [158] extensively examined the process of selecting the weight function and the length of the time window. In addition, ref. [158] used the modulated realized covariance (MRC) estimate to represent the covariance of financial asset returns. This estimator utilizes preaveraging to produce the asymptotic distribution, taking into account asynchronous trading and microstructure noises. Meanwhile, the authors improved the Hayashi–Yoshida (HY) estimator by using preaveraging. This estimator has a significant benefit by retaining information that is generally discarded during a synchronization process.
- **Other methods**: Ref. [159] proposed a "realized" regression based on a time-varying factor model to estimate the spot beta from asynchronous and noisy high-frequency and high-dimensional data. Ref. [160] investigated an innovative methodology that merges low-frequency dynamic models with high-frequency volatility matrix estimation to address the challenges associated with asynchronization and high-dimensional problems in estimating large volatility matrix.

3.4. Others Issues (OT)

3.4.1. Imbalanced Data

A major topic in high-frequency financial data analysis is predicting the direction of the price movement of financial assets, and it is often a classification question in machine learning. Depending on the performance of the financial market and how one defines the

price movement directions such as upward, level, and downward, the numbers of observations among different price movement directions could be very imbalanced. For example, if the stock market performs well during the study period, then there would be many more data labeled upward than downward. However, machine learning algorithms tend to generate bias towards certain classes when trained on datasets with imbalanced class distributions, particularly when the sample size of a given class far outweighs the others. We have compiled a list of solutions as follows to address this problem:

- **SMOTE, oversampling, undersampling**: In the literature, many methods have been studied to deal with imbalanced labels for classification questions, but only a few articles address such an issue in the field of quantitative finance. Ref. [122] used the random undersampling and synthetic minority oversampling technique (SMOTE) [163] to mitigate the problem of class imbalance when forecasting the upward and downward movements of the intraday stock market. Ref. [164] utilized the SMOTE in conjunction with a fuzzy approximative classifier to forecast movement in several financial domains, such as the stock index. To overcome the imbalance issue on forecasting foreign exchange jump movements, ref. [165] first used the SMOTE to oversample the minority classes. Subsequently, they employed a random undersampling technique on the majority class to achieve equal class sizes in the training sets.
- **Threshold**: Using carefully selected thresholds to create a more balanced training set is a widely used method to address the problem of class imbalance in classification that predicts price movement using TAQ or limit order book data, such as [1,33,83], and others.
- **Interval-valued fuzzy rule-based classification systems**: Ref. [164] proposed a system that uses fuzzy rules with interval values to model and forecast financial applications. The suggested solution addresses the issue of financially imbalanced assets by removing the need for preprocessing or resampling methods and thus ensures that no unintentional noises are introduced into the data during the learning process.
- **SVM preprocessor**: Ref. [166] applied the support vector machine (SVM) to address the issue of class imbalance. The concept involves using the SVM as a preprocessor, where the predictions made by the trained SVM are used to substitute the actual values of the training data. Ref. [167] used the proposed preprocessor SVM to create a balanced dataset, mitigating the significant learning biases associated with unbalanced training sets and thus improving the detection of intraday stock price manipulation.

3.4.2. Intraday Seasonality

Intraday seasonality refers to patterns that emerge throughout a single trading day in financial markets. These patterns may include different volatilities and trading activities at specified times, such as market opening, lunchtime, and market closing. The following are some articles that account for such intraday seasonality issues.

The multiplicative component GARCH model [168] is an appropriate approach to address seasonality concerns in the prediction of high-frequency volatility because it allows for the breakdown of volatility into several components, with each indicating a different source of volatility. To account for intraday seasonality in high-frequency log-returns, ref. [169] utilized the multiplicative component GARCH model to estimate the univariate margin and then used a full-range tail dependence copula to capture intraday dynamic tail dependence patterns. Similarly, ref. [170] used the multiplicative component GARCH to forecast risk measures (e.g., VaR) on intraday foreign exchange returns. To eliminate intraday seasonality in raw return data, ref. [171] applied the multiplicative component GARCH model for deseasonalized returns, intraday returns divided by an estimated seasonality term [172], to forecast the value-at-risk (VaR) using high-frequency stock data.

Ref. [173] revealed that intraday returns of S&P 500 index Futures have the characteristics of heteroscedasticity and time-varying autocorrelation. To overcome the aforementioned intraday seasonality problem, they utilized a vector autoregression heteroskedasticity and autocorrelation consistent (VARHAC) estimator, which was first proposed by [174], that utilized vector autoregression to create a heteroskedasticity and autocorrelation covariance matrix to estimate the realized volatility. Ref. [175] proposed self-organizing maps (SOMs) that use neural network learning and nonlinear projections to tackle the problem of intraday seasonality. A straightforward method for extracting intraday seasonality, utilizing a wavelet multiscaling approach and excluding model selection parameters, was presented in a study by [8].

4. Recent Survey Papers

This section provides an overview of the most recent survey papers (published since 2020) relevant to the analysis of high-frequency financial data. Although these surveys are not exclusively focused on high-frequency financial data, they are arranged in decreasing order according to the percentage of articles in each survey that focus on high-frequency data. A comprehensive summary of these recent survey articles is presented in Table 2.

The study reviewed by [176] included a comprehensive analysis, comparison, and assessment of the latest research in the field of forecasting price movements using limit order book (LOB) data. Ref. [177] examined the importance and need for using deep neural networks (DNNs) to predict stock prices and trends. This article explored the suitability of several types of DNNs to analyze temporal stock market data. Ref. [178] reviewed the most recent machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance. They summarized the articles using different categories of stock market data/tasks/prediction models/model performance evaluation/portfolio evaluation. Ref. [179] conducted an extensive literature study on the use of deep learning (DL) to implement financial time series forecasting. The studies were categorized according to their intended areas, such as forecasting stocks, market indices, and foreign exchanges. Furthermore, they were grouped according to their DL model choices, which included Deep Multilayer Perceptron (DMLP), Recurrent Neural Network (RNN), Long-Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), Autoencoders (AEs), etc. Ref. [180] discussed various techniques for predicting stock prices that incorporate the ARIMA, LSTM, hybrid LSTM, CNN, and hybrid CNN. This survey also discussed the constraints and accuracy of the variant hybrid LSTMs and CNNs. Ref. [181] explored machine learning methods to predict the stock market. The primary contribution of this survey consists of a comprehensive analysis of market data, predictors, and various machine learning methods used for stock market predictions. Ref. [182] comprehensively surveyed scenarios where deep learning was used in the stock market. Their primary emphasis was on publications that incorporate the practice of backtesting. Ref. [183] provided three types of data processing methods, including feature selection, order reduction, and feature representations, to formulate the data before fitting the raw stock data to machine learning models. In the meantime, a comprehensive survey on the use of machine learning for stock market prediction has been conducted. Ref. [184] reviewed the latest progress made in the use of machine learning and deep learning techniques in the field of finance. The primary contributions consisted of a detailed analysis of data characteristics and input features, validation evaluation, and model performance. Ref. [185] provided an extensive review of the advancements and applications of deep learning (DL) techniques, reinforcement learning (RL) and deep reinforcement learning (DRL) in the context of information-based decision making within the financial sector. Ref. [186] performed a thorough literature review and meta-analysis

of machine learning algorithms used to forecast the foreign exchange market. Ref. [187] critically examined the most recent research on machine learning in finance. The author identified three distinct categories of ML applications: (1) the development of advanced and innovative metrics, (2) the minimization of prediction inaccuracies, and (3) the expansion of the conventional econometric toolkit. Ref. [188] provided a bibliographic review on applying artificial intelligence and machine learning to predict stock and oil prices, manage investment portfolios, and analyze behavior finance. Ref. [189] examined the application of machine learning and text mining methodologies in the context of news-driven stock market forecasting.

Table 2. Summary of recent survey papers.

Article	Data Frequency	Reviewed Models	Implementation Areas	Data Type
[176]	tick-by-tick, second, daily, weekly, monthly	Linear, Nonlinear, Multilinear, Image Classification, Dimensionality Reduction, Neural Network, Deep Learning	stock price changes prediction on LOB	TAQ, LOB, MBO
[177]	mins, daily, weekly	CNN, Deep Q-network, RNN, LSTM, GRU, echo state network, restricted Boltzmann machine, deep belief network	stock price, trend prediction	TAQ
[178]	intraday and interday	Statistical Method (AR, MA, ARMA, ARIMA), Machine Learning (ML, LR, NB, DT, RF, SVM, Bagging Boosting, KNN), Deep Learning (ANN, MLP, RNN, CNN, LSTM), Reinforcement Learning (RL-Q-Learning, RL-Deep-Q-Learning)	quantitative finance and the stock market	TAQ
[179]	mins, daily, weekly, monthly	Deep Multi-Layer Perceptron, RNN, LSTM, Convolutional Neural Networks, Restricted Boltzmann Machines, Deep Belief Networks and Autoencoders.	stock, index, commodity, cryptocurrency price forecasting, stock price forecasting with text mining techniques, volatility forecasting, trend forecasting, trend forecasting with text mining techniques	TAQ, Textual Data
[180]	mins, daily, weekly	ARIMA, SLTM, Hybrid LSTM, CNN, Hybrid CNN, LSTM Variants	stock price prediction	TAQ
[181]	mins, hourly, daily, weekly, monthly, quarterly, yearly	Machine Learning	stock market prediction	TAQ
[182]	intraday and interday	Deep Learning	stock market prediction	TAQ
[183]	intraday, daily, weekly, monthly, quarterly	ANN, SVM, Naïve Bayes, Genetic Algorithms, Fuzzy Algorithms, Deep Neural Networks, Regression Algorithms, Hybrid Approaches	pre-processing techniques, stock market prediction	TAQ, Textual Data
[184]	intraday and interday	Machine Learning and Deep Learning	stock markets, portfolio management, cryptocurrency, forex markets, financial crisis, bankruptcy and insolvency	TAQ
[185]	intraday and interday	Reinforcement learning, Deep Learning, Deep Reinforcement Learning	decision making in finance	TAQ
[186]	mins, daily, montly, quartely	Machine Learning	foreign exchange market forecasting	TAQ
[187]	intraday and interday	Artificial Intelligence	the application of ML in finance	TAQ
[188]	intraday and interday	Artificial Intelligence and Machine Learning	bankruptcy prediction, stock, oil price prediction, portfolio management, etc.	TAQ
[189]	intraday and interday	Machine Learning and Text Mining	news-driven stock market forecasting	TAQ, Textual Data

5. Discussion

The primary objective of this survey is to identify the critical and unique challenges associated with high-frequency financial data and to highlight relevant studies that propose

potential solutions, improvements, or offer insightful perspectives on addressing these challenges. Through an extensive review of recent literature, we have observed a notable trend toward the application of deep learning and artificial intelligence techniques for analyzing high-frequency financial data. For a detailed introduction and discussion on deep learning architectures, we refer readers to [179,180,190]. In addition, there is a growing emphasis on using increasingly granular datasets, which underscores the evolving nature of research in this domain.

The inherent complexity of high-frequency financial data, coupled with the major challenges identified in our analysis, necessitates the adoption of more advanced computational algorithms and the incorporation of higher-resolution data. Although such advances provide promising opportunities compared to traditional methods that utilize lower-frequency data, they also present significant challenges. Specifically, the use of powerful algorithms and extensive datasets does not inherently translate into improved predictive accuracy or meaningful insights. In fact, there is a risk that sophisticated black-box models may overfit the data, resulting in suboptimal performance, particularly when applied to out-of-sample data.

As such, it is crucial to investigate strategies for effectively leveraging these advanced computational capabilities, algorithms, and detailed comprehensive datasets in a manner that enhances predictive power. Our documented data preprocessing and quantitative methods, presented in this survey, serve a dual purpose: they support advanced quantitative research in high-frequency financial data by systematically collecting, categorizing, and documenting classical and state-of-the-art methodologies, and they provide essential tools for evaluating fair market practices. These methods enable a precise analysis of market liquidity, volatility, and efficiency within high-frequency financial data contexts. In addition, they facilitate detailed examinations across various types of financial data, including stocks, cryptocurrencies, futures, and foreign exchange markets, thus improving financial market transparency and promoting regulatory compliance. We hope that this survey contributes to advance these efforts by providing a thorough examination of the current landscape and identifying key areas for future research.

Author Contributions: Conceptualization, L.H. and L.Z.; methodology, L.H. and L.Z.; validation, L.Z. and L.H.; formal analysis, L.Z. and L.H.; investigation, L.H. and L.Z.; resources, L.Z and L.H.; data curation, L.Z. and L.H.; writing—original draft preparation, L.Z.; writing—review and editing, L.H.; visualization, L.Z.; supervision, L.H.; project administration, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The authors extend their gratitude to the Information Delivery Services of Northern Illinois University for their prompt response to providing many requested articles. Their efficient support greatly facilitated the completion of the survey paper. The authors also thank the three anonymous reviewers whose suggestions have helped us improve the presentation of the work.

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

Abbreviations

The following abbreviations are used in this manuscript:

AD	Asynchronous data
AEs	Autoencoders
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
ARCH	Autoregressive conditional heteroskedasticity model

ARIMA	Autoregressive integrated moving average
ARIMA-ANN	Autoregressive integrated moving average with ANN
ARIMA-LSTM	Autoregressive integrated moving average with LSTM
BoF	Bag-of-features
BRNN	Bayesian regularized neural network
BSd	B-spline wavelet of high order d
CEEMDAN	Complete ensemble empirical mode decomposition with adaptive noise
CIDR	Cumulative intraday returns
CNN	Convolutional neural networks
CSR	Cross-sectional regression
DARL	Data-augmentation-based reinforcement learning
DB-BLSTM	Daubechies wavelet with the bidirectional LSTM mode
DBNs	Deep belief networks
DCC	dilated causal convolutional filters
DEFS	Differential evolution feature selection
DL	Deep learning
DMLP	Deep multi-layer perceptron
DNN	Deep neural networks
DRL	Deep reinforcement learning
DWT	Discrete wavelet transformatio
EEMD	Empirical ensemble mode decomposition
EMD	Empirical mode decomposition
ETICA	External trend and internal component analysis
FCD	Filter cycle decomposition
FCDP	Functional classification and dynamic prediction
FFNN	Feedforward neural network
FPCA	Functional principal component analysis
FSI	Financial stress indices
GANs	Generative adversarial networks
GARCH	Generalized autoregressive conditional heteroskedasticity
GAS	Generalized autoregressive
GP-RU-MIDAS	Group penalized reverse unrestricted mixed data sampling mode
GPR	Gaussian poisson regression
GRU	Gate recurrent unit
HHT	Hilbert-Huang transformation
HP	High-pass filter
HY	Hayashi-Yoshida
IDIN	Increasing-decreasing linear neuro
IFT	Inverted fourier transform
IMFs	Intrinsic mode functions
KECM	Kalman expectation conditional maximization
KEM	Kalman smoother and expectation maximization
KL-VMD	Kullback-Leibler divergence optimized variational mode decomposition
KNN	K-Nearest neighbor classifie
LOB	Limit order book data
LP	Low-pass filter
LR	Linear regression
LS	Low signal-to-noise
LSTM	Long short term memory
MA	Moving average
MARSPlines	Multivariate adaptive regression spline

MBD	Message book data
MBO	Market-by-order
MBP	Market-by-price
MINSPAN	Minimizing the tuple span
ML	Machine learning
MLP	Multilayer perception
MODWT	Maximal overlap discrete wavelet transform
MRC	Modulated realized covariance
mRMR	Minimum redundancy-maximum relevance
MRS-HEAVY	Markov regime-switch high-frequency-based volatility mode
MS-HAR	Markov switching heterogeneous autoregressive mode
NS	Nonstationarity
NYSE	New York Stock Exchange
OFI	Order flow imbalance
OHLC	Open, High, Low, Close
OHLCV	Open, High, Low, Close, Volume
OI	Order imbalance
OT	Others
PCA	Principal component analysis
PFP	Phase fix procedure
RASA	Recurrently adaptive separation algorithm
RBFNN	Radial basis function neural network
RBMs	Restricted boltzmann machines
RCARF	Recurring concepts adaptive random forest
ResNet	Residual learning neural networks
RL	Reinforcement learning
RNN	Recurrent neural network
RSM	Random sampling method
RT	Regression tree
SFM	State frequency memory
SG	Savitzky–Golay filter
SGD	Stochastic gradient descent algorithm
SMOTE	Synthetic minority oversampling technique
SNR	Signal-to-noise ratios
SOM	Self-organizingg map
SUS	Stochastic universal sampling
SVM	Support vector machine
SVR	Support vector regression
T-BoF	Temporal bag-of-features
TABL	Temporal attention-augmented bilinear layer
TAQ	Trade and quote data
TCN	Temporal convolution network
TSR	Time-series regression
TVP-VAR	Time-varying parameter vector autoregression
UHFD	Ultra high-frequency Data
VaR	Value-at-risk
VARHAC	Vector autoregression heteroskedasticity and autocorrelation consisten
VLS	Variable length sampling
VMD	Variational mode decomposition
WFO	Walk forward optimization
WT	Wavelet transformation

Appendix A. Overview of Articles on High-Frequency Financial Data

Table A1. Summary issue–solution tables for addressing nonstationarity (NS), low signal-to-noise (LS), asynchronous data (AD), and other issues (OT). The term "Data" refers to the data type.

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[155]	Turkish equity market Borsa Istanbul	TAQ	log price	Intraday conditional correlation	GAS framework, State-Space Modeling	AD—missing data approach
[156]	S&P 500	TAQ	price	covariance matrix	Kalman–EM algorithm	AD-missing data approach
[157]	Simulated data	TAQ	log price, returns, jumps	Covariance Matrix	KEM, pairwise refresh, GARCH	AD—missing data approach
[11]	NYSE	TAQ	return	covariance matrix	Bayesian Dynamic Linear model	AD—missing data approach
[154]	SPY	TAQ	volatility	volatility	QML, Step, RT, Pol, Kern, MRC, GARCH-X	AD—missing data approach
[159]	S&P 100 constituents.	TAQ	fixed year effects, earnings announcement dummies	spot beta	Smoothed TSRV estimator	AD—other method
[160]	Shenzhen and Shanghai Stock Exchanges	TAQ	log price process	volatility matrix	matrix factor model, vector autoregressive model	AD—other method
[162]	Foreign exchange future contracts	TAQ	log returns	covariance matrices	QMLE, generalized synchronization scheme	AD—pairing and synchronization methods
[152]	Stocks	TAQ	log price	return	copula	AD—pairing and synchronization methods
[153]	MSFT, GOOG	TAQ	price	covariance	Previous-tick covariance estimator, Epp effect remedies, TSCV	AD—pairing and synchronization methods
[158]	Equity data	TAQ	log-price	covariance matrix	MRC estimator, HY, ARMA	AD—preaveraging
[147]	US equity	TAQ	return	covariance matrices	Multivariate realized kernels	AD—refresh time alogirthm
[150]	Dow Jones index	TAQ	log price	volatility covariance	all/pairwise refresh	AD—refresh time alogirthm
[151]	simulated data	TAQ	log price process	volatility matrix	stationary bootstrapping method, MCMC simulation	AD—refresh time alogirthm
[148]	IBM	TAQ	historical price differentials	price differentials	regression	AD—refresh time alogirthm, paring and synchronization method
[17]	Taiwan futures	TAQ	images	arbitrage opportunities	Filterbank CNN	LS—alternative data
[113]	Brazilian and Mexican equity indices	TAQ	macroeconomic news , return	return, volatility	regressions	LS—alternative data
[112]	Dow Jones stocks	TAQ	return, public information	volatility	FIEGARCH models	LS—alternative data
[111]	S&P 500, crude oil futures	TAQ	volatility	volatility	MIDAS regression model	LS—alternative data
[104]	Chinese A-share	LOB	correlation graph	price	Multi-Hawkes Processes, GALSTM	LS—alternative data
[91]	S&P500	TAQ	directional indicators, relevance indicators, meta features	tweet returns	RF, FFNN, K-nearest, Naïve	LS—alternative data, label preparations
[21]	London Stock Exchange	LOB, MBD	input matrices	up/down	CNN	LS—data compression
[18]	FOREX market	LOB	raw tick time series	probabilities vector	CNN, LSTM, LSTM-CNN	LS—data compression

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[132]	S&P 500, KOSPI, DJIA	TAQ	price time series, mini batch	rise or fall of index direction	ResNet18, RNN, LSTM	LS—data compression
[134]	NASDAQ stock, FI-2010 data set	LOB	10-level ask/bid prices/volumes	up, down, static	MLP, PCA, LDA, AE, BoFs	LS—data compression
[19]	LOBSTER	LOB	LOB data	anomaly detection	Transformer autoencoder	LS—data compression
[131]	Stock index and Treasury bond	TAQ, LOB	OHLCV, 5-level LOB	price	BAH, MV, TSM, MLP, GRU, LGBM, DQN, variant of DeepScalper	LS—data compression
[136]	Tokyo Stock Exchange	LOB, MBO	order embeddings	up/down	Logistic, SVM, MLP, variant CNNs	LS—data compression
[133]	FI-2010 data set	LOB	tensor input	mid-price direction	WMTR, MTR, RR, SLFN, BoF, N-BoF	LS—data compression
[135]	NASDAQ, FI-2010 data set	LOB	10-level ask/bid prices/volumes	up/down	TABLs, SVM, MLP, CNN, LSTM	LS—data compression, regularizations
[126]	S&P 100 Index	TAQ	spot correlation matrix	eignvalue	high frequency PCA	LS—dimension reduction
[125]	NYSE, NASDAQ	TAQ	latent factors	volatility, correlation matrix	PCA,VAR, HAR	LS—dimension reduction
[139]	S&P constituent	TAQ	cross-section returns	returns	Lasso, EN, PCR, RF, GBRT, ANN	LS—dimension reduction
[121]	Korea Composite Stock Index	TAQ	various prices and volumes	prices	stepwise regression, ANN	LS—dimension reduction
[130]	Taiwan stock market	TAQ, LOB	basic features, LOB features and technical indicators	price spread	Deep Learning models	LS—dimension reduction
[129]	S&P 500	TAQ	technical indicators	returns	DEFS, (Fuzzy) KNN classifier, MLPM-FKNN	LS—dimension reduction
[128]	WTI and Brent crude oil	TAQ	technical indicators	prices	BPNNs	LS—dimension reduction
[123]	Dow Jones Index	TAQ	latent factors	return	Matrix-Valued Factor Model	LS—dimension reduction
[124]	Dow Jones, S&P	TAQ	various portfolios	covariance matrices	TSR, CSR, PCA	LS—dimension reduction, AD—refresh time sampling
[122]	Shenzhen Stock Exchange	TAQ	liquidity measures and technical indicators	jumps directions	SVM, ANN, RF, KNN	LS—dimension reduction, OT-imbalanced data
[118]	Cryptocurrency	TAQ, LOB	time-lagged features, transcation loggs	volume	ARMA-GARCH, GBM, TME	LS—ensemble
[120]	Brazilian Stock Exchange	TAQ	candlestick, technical indicators	0, 1	MLP	LS—ensemble
[119]	Chinese Shanghai Composite Index	TAQ	time series	price	Ensemble De-noising, SVM	LS—ensemble
[90]	Bank of America, Chevron	LOB	bid/ask price/size	microprice	mathematical computation	LS—label preparation
[89]	NASDAQ	TAQ, LOB	technical indicators	0, 1	GRU	LS—label preparation, dimension reduction
[92]	Cryptocurrencies	TAQ	close price	trend prediction	triple trend labeling method, attention-based CNN–LSTM model	LS—label preparation, ensemble
[110]	US industry indexes	TAQ	Industry and market volatility	Realized volatility	HAR, machine learning	LS—new feature preparations
[106]	Intel and Microsoft	TAQ	log return	volatility	TSRV	LS—new features preparation

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[20]	BIST30	LOB	OFIs, # of arrive/ canceled/initiated buy (sell) orders	return	time series regression, Fama–MacBeth regression	LS—new features preparation
[99]	Oil market	TAQ	realized volatility	volatility	variants of HAR model	LS—new features preparation
[108]	foreign exchange and commodity futures	TAQ	price-based features, positional features	Financial decision making	deep reinforcement learning	LS—new features preparation
[98]	Tokyo Stock Exchange	TAQ	return	volatility	HAR model	LS—new features preparation
[114]	Cryptocurrency	TAQ	twitter sentiment, message volume	price return, and trading volume	sentiment analysis, Granger causality testing,	LS—new features preparation
[94]	Shanghai Securities Composite Index	TAQ	technical indicators, inverstor attention index	volatility	TCN, LSTM, GARCH, FIGARCH, ARIMA, HAR-RV	LS—new features preparation
[3]	LOBSTER	LOB	OFI	return	CNN+inception LSTM	LS—new features preparation
[101]	Brazil Stock Exchange	TAQ, LOB	trade/LOB features, segment features	High, Medium, Low	XGBoost	LS—new features preparation
[109]	NASDAQ-100	TAQ	intraday return	volatility	Dilated Causal Convolutions (DCCs)-based neural network	LS—new features preparation
[96]	India stock	TAQ	technical indicators	up/down	ANN, SVM, proposed model	LS—new features preparation
[97]	Australian stock market	LOB	technical indicators	buys/sells	HARX, LX	LS—new features preparation
[102]	Shanghai Securities 50 Index	TAQ	previous daily data, index opening features	up/down	decision tree, XGBoost, RF, SVM, LSTM	LS—new features preparation
[107]	S&P500	TAQ	return, volume	volatility	State-Space, Kalman Filter and Smoother, mcsGARCH	LS—new features preparation
[105]	Stocks	LOB	LOB 2D images	up/down	CNN	LS—new features preparation
[95]	Stock index	TAQ	price, volume, technical indicators	up/down	LSTM, LSTM+DeepFM, ATT-CNN, MFNN, FA-CNN	LS—new features preparation
[100]	NYSE	TAQ	log price	covariance matrices	LL, g-GAS, DCC, EWMA	LS—new features preparation, AD—missing data approach
[93]	Dow Jones 30 component stocks.	TAQ	features extracted from proposed methods	up, down, static	SVM, Enet	LS—new features preparation, AD—pairing and synchronization method
[103]	US and Nordic stocks	LOB	Econometric features, Tech and Quant features, LOB features	1, −1, 0	MLP, CNN, LSTM	LS—new features preparation, label preparations
[37]	Currencies	TAQ	time series data	exchange rate	deep learning models	LS—noise reduction
[88]	S&P 500	TAQ	price	price	LSTM	LS—noise reduction
[87]	US DJIA 30 stocks	TAQ	return	return	WRASA, AR, ARMA, ARMA-GARCH	LS—noise reduction
[79]	S&P, Nasdaq, SSE and SZSE stocks	TAQ	returns	internal components	ETICA	LS—noise reduction
[86]	S&P500, NASDAQ, DJIA, NYSE	TAQ	historical volatility	volatility	BSd, db3/4, Haar-RNN, GARCH, ARIMA	LS—noise reduction
[84]	Stock indices	TAQ	price	price	ARIMA, SVR, LSTM, GRU	LS—noise reduction

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[82]	Stock exchange	TAQ	bid/ask price/size related features	long, short, holding	SVM/SVR, random forest, XGBoost, and neural network	LS—noise reduction
[85]	German equity market	TAQ	time series	price	LLSA+ARIMA, MODWT+ARIMA	LS—noise reduction
[81]	China stock market	TAQ	volume, return	IVaR	Copula, MCMC simulate	LS—noise reduction
[41]	CSI 300 Index	TAQ	price	close price	CEEMDDAN+SG, LSTM	LS—noise reduction
[78]	NYSE	TAQ	raw data	cleaned data	data cleaning procedures	LS—noise reduction, AD—data cleaning
[80]	Simulated data	NA	log price	covariance matrix	preaverging, multiscale	LS—noise reduction, AD—pairing and synchronization method
[83]	Nasdaq Stock Market, FI-2010	LOB	40 features for 10 level ask/bid price and size	up/down	DeepLOB, DeepLOB-Attention, DTNN	LS—noise reduction, label preparations OT—imbalanced data
[127]	S&P 100 Index	TAQ	return	covariance	PCA	LS—penalized methods
[140]	NSE stocks	TAQ	price, volume	price, volatility	cluster and SVR	LS—penalized methods and regularization
[141]	S&P500 E-mini futures	LOB	records of LOB observations	$-1, 0, 1$	RNN	LS—penalized methods and regularization
[137]	NASDAQ	LOB	LOB features, technical indicators	up, down, static	SVM	LS—penalized methods and regularization
[138]	Bitcoin	TAQ	previous five price observations	price	SVR, GPR, RT, kNN, RBFNN, FFNN, BRNN	LS—penalized methods and regularization
[13]	New York Stock Exchange	TAQ	price	volatility matrix	KRVM, KRVPM, PRVM, PRVPM	LS—penalized methods and regularization
[6]	NYSE, NASDAQ, AMEX	TAQ	OHLC price	high price	online evolving fuzzy rule-based prediction model	NS—adaptive methods
[10]	Chinese CSI 300 Index	TAQ	lagged returns	return	SVR with New/Radial Basis/Sigmoid Kernel	NS—adaptive methods
[12]	Foreign Exchange	TAQ	return	volatility	SST, rFFF, FFF	NS—adaptive methods
[75]	China's financial market	TAQ	date, normalized open/close/low/ high volume/money	close price	Informer, LSTM	NS—attention mechanism
[74]	Shanghai crude oil market	TAQ	normalized OHLC data	price	naïve, VAR, VECM, MLR, SVR, LSTM, transformer	NS—attention mechanism
[31]	S&P 500 futures	TAQ	technical indicators	flow toxicity	PIN model	NS—bar preparations
[32]	S&P100	TAQ	imformation driven bars	return	panel regression, RF	NS—bar preparations, LS—label preparation
[30]	NSE	TAQ	technical indicators	price	GA-ANFIS, PSO-ANFIS	NS—data aggregation, adaptive methods
[16]	US large-cap stocks	TAQ	technical indicators	buy/sell	LSTM ensembles, Lasso LR, Ridge LR	NS—data aggregation, ensemble
[15]	S&P 500 ETF	TAQ	technical indicators	price	RCARF, ARF, RCD, DWM, AHOEFT	NS—data aggregation, ensemble
[48]	Bitcoin	TAQ	technical indicators	log return	FF-D-ANN, SVM, RF, FW, ARMAX	NS—data aggregation, LS—label preparations
[27]	SPY, QQQ	TAQ	log returns	return	Bayesian linear regression with Student's t error	NS—differencing
[22]	NASDAQ	TAQ	open	HLC	LR, BPNN	NS—differencing
[24]	Crude oil futures	TAQ	CIDRs	CIDRs	weighted functional FPCA	NS—differencing, adaptive

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[28]	Shanghai Composite Index	TAQ	log return	log return	regression	NS—differencing, data aggregation, LS—label preparations
[23]	S&P/ASX	TAQ	price	CIDRs	FPCR + bootstrapping, FPCA	NS—differencing, LS—dimension reduction
[26]	S&P500	TAQ	log return	volatility	V-MF-GAS	NS—differencing, multiple frequencies modeling
[25]	Euro/Dollar exchange rate	TAQ	log return	volatility	long-memory GARCH	NS—differencing, modeling multiple orders of ments
[40]	SS&P 500 Index	TAQ	price	price	Naïve, ARIMA, SVR	NS—EMD
[39]	Bitcoin	TAQ	price	price	GAM	NS—EMD, VMD
[68]	China stock market	TAQ	statistical indicators	buy, sell, keep	logistic regression, SVM, GBDT, RF, LSTM	NS—ensemble, LS—label preparation
[64]	Tokyo Stock Exchange	TAQ	historical market event, noise	synthetic data	GANs	NS—including more data
[62]	Stocks	TAQ	OHLC	return	SLM-Lab, Deep Q-Learning	NS—including more data
[61]	Stocks	TAQ	encorders	up/down	Self-FIS, CNN, TCN	NS—including more data, LS—data compression
[59]	Brazilian, American, and Chinese stocks	TAQ	technical indicators	return	SVR	NS—including more dimensions
[60]	Bitcoin	TAQ	technical indicators	price	TI-SVR, rVMF-SVR, TI-rVMF-SVR	NS—including more dimensions, LS—noise reduction
[63]	S&P500 index	TAQ	volatility	volatility	ARIMA, HAR, OLS, LASSO, XGBoost, MLP, LSTM	NS—including more features
[2]	Stocks	TAQ	technical indicators	price	WSAEs-LSTM, WLSTM, LSTM, RNN	NS—including more features, sequence modeling methods
[57]	Cryptocurrency	TAQ	return	volatility	GARCHs	NS—modeling multiple orders of ments
[58]	Shanghai Stock Index.	TAQ	log return	volatility	Semiparametric GARCH Model	NS—modeling multiple orders of ments
[56]	NYSE, USD/EUR returns	TAQ	return	volatility	ARCH, GARCH	NS—modeling multiple orders of moments simultaneously
[54]	Stocks	TAQ	investor sentiments, stock return	return	text sentiment analysis, RR-MIDAS, rolling EMD decomposition	NS—multiple frequencies modeling
[7]	Stocks	TAQ	price	price	AR, LSTM, SFM recurrent network	NS—multiple frequencies modeling
[55]	Chinese stock market	TAQ	risk factors	return	GP-RU-MIDAS	NS—multiple frequencies modeling
[47]	Gold futures prices	TAQ	reconstructed components by IMFs	price	CEEMDAN-SE, ARIMA-CNN-LSTM	NS—nonstationarity
[38]	Crude oil futures	TAQ	time series data	price	Wavelet Multiresolution Analysis, VAR-BEKK-GARCH	NS—nonstationarity
[34]	NASDQA	LOB	144-D feature vectors	up/down	MLP, CNN, RNN	NS—normalization
[33]	London Stock Exchange	LOB	10 level ask/bid price and size	up/down	ML benchmarks, DeepLOB	NS—normalization, LS—sequence modeling, label preparations
[1]	London Stock Exchange	LOB, MBD	normalized price/size/change price/change size, side and action	up/down	LM, MLP, LSTM, Attention, CNN, DeepLOB, Ensemble	NS—normalization, LS—label preparations, OT—imbalanced data

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[65]	S&P 500	TAQ	return	covariance structure of returns	GARCH, HEAVY, MRS-GARCH, MRS-HEAVY	NS—regime-switch methods
[67]	US ETFs, energy	TAQ	returs	volatility	TVP-VAR	NS—regime-switch methods
[66]	Crude oil futures	TAQ	return	volatility	MS-HAR	NS—regime-switch, LS—alternative data
[76]	Shanghai Shenzhen 300 Index	TAQ	close price	return	Hybrid model of nonparametric regression and LSTM	NS—sequence modeling
[70]	Stocks	TAQ	log price	log price	ARIMA, ANN, ARIMA-ANN	NS—sequence modeling
[5]	Brazilian stock market index	TAQ	price time series	price	ARIMA, MLP, IMP, SHIF, IDLN	NS—sequence modeling
[69]	CSI 300 index	TAQ	price	price, return	ARIMA, SVM, LSTM, ARIMA-SVM, ARIMA-LSTM	NS—sequence modeling
[72]	CSI-500	LOB	ask/bid price/size	stock price percentage change	Linear, XGBoost, DeepLOB, DeepAcc, LGT, O-LGT	NS—sequence modeling
[14]	BM&F Bovespa stock exchange	TAQ	technical indicators, price history	0, 1	LSTM, MLP, RF, Random	NS—sequence modeling
[73]	CSI 300 index	TAQ	intraday return, return related to closing and opeing price	0, 1	TCN, LSTM, RF	NS—sequence modeling
[71]	Chinese stock index futures	TAQ	OCHLV and time indicators	log-returns	wavelets+machine learning methods (ANN, GRU, LSTM)	NS—sequence modeling, LS-label preparation
[43]	Crude oil	TAQ	econometric variables	price	Deep learning, VMD	NS—signal decomposition
[35]	Nikkei stock return index	TAQ	price	volatility	MA, ARCH, GARCH	NS—wavelet
[50]	S&P 500	TAQ	price	price	AW, ARMA, BPNN+AC, BPNN+DC	NS—wavelet
[36]	S&P500, DJI, and KOSPI200 indices	TAQ	Open, Close, Volume, MA5, MA10	buy/sell/hold action	LSTM, RNN, RL	NS—wavelet
[44]	Commodity and FX future	TAQ	mini batch	−1, 0, 1	DNN	NS-WFO
[45]	Cryptocurrency	TAQ	OHLC	up, down, static	MLP, LSTM, RSM	NS—WFO, RSM
[9]	AAPL stock	TAQ	log return	log return	Elman/Jordan neuro-wavelet net	NS/LS-wavelet
[165]	Foreign exchange	TAQ	liquidity, volatility, technical indicators	up/down	KNN, decision tree, SVM, RF, ANN, ARIM+AdaBoost	OT—Imbalanced data
[167]	Stock index	TAQ	trade-based/characteristic features	1, 0	RNN-based ensemble learning	OT—imbalanced data
[164]	Various datasets, including financial data	TAQ	interval values	up, down, static	Fuzzy Rule-Based Classification System	OT—imbalanced data
[171]	Stock data	TAQ	deseasonalized filtered return	VaR, CVaR	CGARCH-EVT-Copula	OT—intraday seasonality
[170]	EUR/USD exchange rate	TAQ	return	VaR, ES	Multiplicative component GARCH	OT—intraday seasonality

Table A1. *Cont.*

Art.	Market	Data	Input	Target	Methodology	Issue–Solution
[173]	S&P 500	TAQ	return	volatility	VARHAC	OT—intraday seasonality
[8]	DEM-USD, JPY-USD	TAQ	return	return	Wavelet smooth, AR+Seasonality	OT—intraday seasonality
[169]	US stock	TAQ	log returns, Fama–French five factors	tail dependence	ARMA-mcsGARCH, PPPP copula, regression	OT—intraday seasonality
[175]	Foreign exchange	TAQ	return	volatility	IAOM, kernel smoothing approache, Nadaraya–Watson kernel method	OT—intraday seasonality

References

1. Zhang, Z.; Lim, B.; Zohren, S. Deep learning for market by order data. *Appl. Math. Financ.* **2021**, *28*, 79–95. [CrossRef]
2. Bao, W.; Yue, J.; Rao, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE* **2017**, *12*, e0180944. [CrossRef]
3. Lucchese, L.; Pakkanen, M.S.; Veraart, A.E. The short-term predictability of returns in order book markets: A deep learning perspective. *Int. J. Forecast.* **2024**, *40*, 1587–1621. [CrossRef]
4. Available online: https://www.cmegroup.com/education/market-by-order-mbo.html (accessed on 20 January 2025).
5. Araújo, R.d.A.; Oliveira, A.L.; Meira, S. A hybrid model for high-frequency stock market forecasting. *Expert Syst. Appl.* **2015**, *42*, 4081–4096. [CrossRef]
6. Gu, X.; Angelov, P.P.; Ali, A.M.; Gruver, W.A.; Gaydadjiev, G. Online evolving fuzzy rule-based prediction model for high frequency trading financial data stream. In Proceedings of the 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Natal, Brazil, 23–25 May 2016; IEEE: New York, NY, USA, 2016; pp. 169–175.
7. Zhang, L.; Aggarwal, C.; Qi, G.J. Stock price prediction via discovering multi-frequency trading patterns. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 2141–2149.
8. Gençay, R.; Selçuk, F.; Whitcher, B. Differentiating intraday seasonalities through wavelet multi-scaling. *Phys. A Stat. Mech. Its Appl.* **2001**, *289*, 543–556. [CrossRef]
9. Ortega, L.; Khashanah, K. A neuro-wavelet model for the short-term forecasting of high-frequency time series of stock returns. *J. Forecast.* **2014**, *33*, 134–146. [CrossRef]
10. Qu, H.; Zhang, Y. A new kernel of support vector regression for forecasting high-frequency stock returns. *Math. Probl. Eng.* **2016**, *2016*, 4907654. [CrossRef]
11. Peluso, S.; Corsi, F.; Mira, A. A Bayesian high-frequency estimator of the multivariate covariance of noisy and asynchronous returns. *J. Financ. Econom.* **2014**, *13*, 665–697.
12. Vatter, T.; Wu, H.T.; Chavez-Demoulin, V.; Yu, B. Non-parametric estimation of intraday spot volatility: Disentangling instantaneous trend and seasonality. *Econometrics* **2015**, *3*, 864–887. [CrossRef]
13. Zou, J.; An, Y.; Yan, H. Volatility matrix inference in high-frequency finance with regularization and efficient computations. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; IEEE: New York, NY, USA, 2015; pp. 2437–2444.
14. Nelson, D.M.; Pereira, A.C.; De Oliveira, R.A. Stock market's price movement prediction with LSTM neural networks. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; IEEE: New York, NY, USA, 2017; pp. 1419–1426.
15. Suárez-Cetrulo, A.L.; Cervantes, A.; Quintana, D. Incremental market behavior classification in presence of recurring concepts. *Entropy* **2019**, *21*, 25. [CrossRef]
16. Borovkova, S.; Tsiamas, I. An ensemble of LSTM neural networks for high-frequency stock market classification. *J. Forecast.* **2019**, *38*, 600–619. [CrossRef]
17. Chen, Y.Y.; Chen, W.L.; Huang, S.H. Developing arbitrage strategy in high-frequency pairs trading with filterbank CNN algorithm. In Proceedings of the 2018 IEEE International Conference on Agents (ICA), Singapore, 28–31 July 2018; IEEE: New York, NY, USA, 2018; pp. 113–116.
18. Milke, V.; Luca, C.; Wilson, G.B. Reduction of financial tick big data for intraday trading. *Expert Syst.* **2024**, *41*, e13537. [CrossRef]
19. Poutré, C.; Chételat, D.; Morales, M. Deep unsupervised anomaly detection in high-frequency markets. *J. Financ. Data Sci.* **2024**, *10*, 100129. [CrossRef]

20. Akyildirim, E.; Sensoy, A.; Gulay, G.; Corbet, S.; Salari, H.N. Big data analytics, order imbalance and the predictability of stock returns. *J. Multinatl. Financ. Manag.* **2021**, *62*, 100717. [CrossRef]
21. Doering, J.; Fairbank, M.; Markose, S. Convolutional neural networks applied to high-frequency market microstructure forecasting. In Proceedings of the 2017 9th Computer Science and Electronic Engineering (CEEC), Colchester, UK, 27–29 September 2017; IEEE: New York, NY, USA, 2017; pp. 31–36.
22. Lam, K. Predictability of intraday stock index. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290), Honolulu, HI, USA, 12–17 May 2002; IEEE: New York, NY, USA, 2002; Volume 3, pp. 2156–2161.
23. Shang, H.L.; Ji, K. Forecasting intraday financial time series with sieve bootstrapping and dynamic updating. *J. Forecast.* **2023**, *42*, 1973–1988. [CrossRef]
24. Li, X.; Liu, X. Functional classification and dynamic prediction of cumulative intraday returns in crude oil futures. *Energy* **2023**, *284*, 129355. [CrossRef]
25. Bordignon, S.; Caporin, M.; Lisi, F. Generalised long-memory GARCH models for intra-daily volatility. *Comput. Stat. Data Anal.* **2007**, *51*, 5900–5912. [CrossRef]
26. Eckernkemper, T.; Gribisch, B. Intraday conditional value at risk: A periodic mixed-frequency generalized autoregressive score approach. *J. Forecast.* **2021**, *40*, 883–910. [CrossRef]
27. Zhang, L.; Hua, L. Market Predictability Before the Closing Bell Rings. *Risks* **2024**, *12*, 180. [CrossRef]
28. Feng, W.; Zhang, X.; Chen, Y.; Song, Z. Linear regression estimation using intraday high frequency data. *AIMS Math.* **2023**, *8*, 13123–13133. [CrossRef]
29. Bousbaa, Z.; Sanchez-Medina, J.; Bencharef, O. Financial time series forecasting: A data stream mining-based system. *Electronics* **2023**, *12*, 2039. [CrossRef]
30. Chandar, S.K. Forecasting intraday stock price using ANFIS and bio-inspired algorithms. *Int. J. Netw. Virtual Organ.* **2021**, *25*, 29–47. [CrossRef]
31. Easley, D.; López de Prado, M.M.; O'Hara, M. Flow toxicity and liquidity in a high-frequency world. *Rev. Financ. Stud.* **2012**, *25*, 1457–1493. [CrossRef]
32. Aït-Sahalia, Y.; Fan, J.; Xue, L.; Zhou, Y. *How and When Are High-Frequency Stock Returns Predictable?* Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2022.
33. Zhang, Z.; Zohren, S.; Roberts, S. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Trans. Signal Process.* **2019**, *67*, 3001–3012. [CrossRef]
34. Passalis, N.; Tefas, A.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Deep adaptive input normalization for time series forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 3760–3765. [CrossRef] [PubMed]
35. Capobianco, E. Multiscale analysis of stock index return volatility. *Comput. Econ.* **2004**, *23*, 219–237. [CrossRef]
36. Lee, J.; Koh, H.; Choe, H.J. Learning to trade in financial time series using high-frequency through wavelet transformation and deep reinforcement learning. *Appl. Intell.* **2021**, *51*, 6202–6223. [CrossRef]
37. Arslan, M.; Hunjra, A.I.; Ahmed, W.S.; Ben Zaied, Y. Forecasting multi-frequency intraday exchange rates using deep learning models. *J. Forecast.* **2024**, *43*, 1338–1355. [CrossRef]
38. Pan, J.; Yin, C.; Zhou, Y. The Spillover Effects among Crude Oil Future Markets under High-Frequency Data Environment: Trivariate VAR-BEKK-GARCH Model Based on Wavelet Multiresolution Analysis. *Procedia Comput. Sci.* **2024**, *242*, 758–765. [CrossRef]
39. Gyamerah, S.A. On forecasting the intraday Bitcoin price using ensemble of variational mode decomposition and generalized additive model. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1003–1009. [CrossRef]
40. Nava, N.; Di Matteo, T.; Aste, T. Financial time series forecasting using empirical mode decomposition and support vector regression. *Risks* **2018**, *6*, 7. [CrossRef]
41. Zhu, R.; Zhong, G.Y.; Li, J.C. Forecasting price in a new hybrid neural network model with machine learning. *Expert Syst. Appl.* **2024**, *249*, 123697. [CrossRef]
42. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2013**, *62*, 531–544. [CrossRef]
43. Lu, W.; Huang, Z. Crude Oil Prices Forecast Based on Mixed-Frequency Deep Learning Approach and Intelligent Optimization Algorithm. *Entropy* **2024**, *26*, 358. [CrossRef] [PubMed]
44. Dixon, M.; Klabjan, D.; Bang, J.H. Classification-based financial markets prediction using deep neural networks. *Algorithmic Financ.* **2017**, *6*, 67–77. [CrossRef]
45. Shintate, T.; Pichl, L. Trend prediction classification for high frequency bitcoin time series with deep learning. *J. Risk Financ. Manag.* **2019**, *12*, 17. [CrossRef]
46. Wen, M.; Li, P.; Zhang, L.; Chen, Y. Stock market trend prediction using high-order information of time series. *IEEE Access* **2019**, *7*, 28299–28308. [CrossRef]
47. Dong, Z.; Zhou, Y. A Novel Hybrid Model for Financial Forecasting Based on CEEMDAN-SE and ARIMA-CNN-LSTM. *Mathematics* **2024**, *12*, 2434. [CrossRef]

48. Gradojevic, N.; Kukolj, D.; Adcock, R.; Djakovic, V. Forecasting Bitcoin with technical analysis: A not-so-random forest? *Int. J. Forecast.* **2023**, *39*, 1–17. [CrossRef]
49. De Prado, M.L. *Advances in Financial Machine Learning*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
50. Lahmiri, S. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. *J. King Saud Univ.-Comput. Inf. Sci.* **2014**, *26*, 218–227. [CrossRef]
51. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
52. Daubechies, I. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 961–1005. [CrossRef]
53. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
54. Cai, Y.; Tang, Z.; Chen, Y. Can real-time investor sentiment help predict the high-frequency stock returns? Evidence from a mixed-frequency-rolling decomposition forecasting method. *N. Am. J. Econ. Financ.* **2024**, *72*, 102147. [CrossRef]
55. Zhuo, X.; Luo, S.; Cao, Y. Exploring Multisource High-Dimensional Mixed-Frequency Risks in the Stock Market: A Group Penalized Reverse Unrestricted Mixed Data Sampling Approach. *J. Forecast.* **2024**. [CrossRef]
56. Francq, C.; Sucarrat, G. Volatility estimation when the zero-process is nonstationary. *J. Bus. Econ. Stat.* **2022**, *41*, 53–66. [CrossRef]
57. Ampountolas, A. Cryptocurrencies intraday high-frequency volatility spillover effects using univariate and multivariate GARCH models. *Int. J. Financ. Stud.* **2022**, *10*, 51. [CrossRef]
58. Chai, F.; Li, Y.; Zhang, X.; Chen, Z. Daily Semiparametric GARCH Model Estimation Using Intraday High-Frequency Data. *Symmetry* **2023**, *15*, 908. [CrossRef]
59. Henrique, B.M.; Sobreiro, V.A.; Kimura, H. Stock price prediction using support vector regression on daily and up to the minute prices. *J. Financ. Data Sci.* **2018**, *4*, 183–201. [CrossRef]
60. Gyamerah, S.A. Two-stage hybrid machine learning model for high-frequency intraday bitcoin price prediction based on technical indicators, variational mode decomposition, and support vector regression. *Complexity* **2021**, *2021*, 1–15. [CrossRef]
61. Sun, J.; Qing, Y.; Liu, C.; Lin, J. Self-fts: A self-supervised learning method for financial time series representation in stock intraday trading. In Proceedings of the 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), Perth, Australia, 25–28 July 2022; IEEE: New York, NY, USA, 2022; pp. 501–506.
62. Yuan, Y.; Wen, W.; Yang, J. Using data augmentation based reinforcement learning for daily stock trading. *Electronics* **2020**, *9*, 1384. [CrossRef]
63. Zhang, C.; Zhang, Y.; Cucuringu, M.; Qian, Z. Volatility forecasting with machine learning and intraday commonality. *J. Financ. Econom.* **2024**, *22*, 492–530. [CrossRef]
64. Naritomi, Y.; Adachi, T. Data augmentation of high frequency financial data using generative adversarial network. In Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Melbourne, Australia, 14–17 December 2020; IEEE: New York, NY, USA, 2020; pp. 641–648.
65. Lai, Y.S.; Sheu, H.J.; Lee, H.T. A multivariate Markov regime-switching high-frequency-based volatility model for optimal futures hedging. *J. Futur. Mark.* **2017**, *37*, 1124–1140. [CrossRef]
66. Liu, Y.; Niu, Z.; Suleman, M.T.; Yin, L.; Zhang, H. Forecasting the volatility of crude oil futures: The role of oil investor attention and its regime switching characteristics under a high-frequency framework. *Energy* **2022**, *238*, 121779. [CrossRef]
67. Naeem, M.A.; Karim, S.; Yarovaya, L.; Lucey, B.M. COVID-induced sentiment and the intraday volatility spillovers between energy and other ETFs. *Energy Econ.* **2023**, *122*, 106677. [CrossRef] [PubMed]
68. Zhang, Q.; Zhang, P.; Zhou, F. Intraday and interday features in the high-frequency data: Pre-and post-Crisis evidence in China's stock market. *Expert Syst. Appl.* **2022**, *209*, 118321. [CrossRef]
69. Li, Z.; Han, J.; Song, Y. On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning. *J. Forecast.* **2020**, *39*, 1081–1097. [CrossRef]
70. Alshawarbeh, E.; Abdulrahman, A.T.; Hussam, E. Statistical Modeling of High Frequency Datasets Using the ARIMA-ANN Hybrid. *Mathematics* **2023**, *11*, 4594. [CrossRef]
71. Liang, D.; Xu, Y.; Hu, Y.; Du, Q. Intraday return forecasts and high-frequency trading of stock index futures: A hybrid wavelet-deep learning approach. *Emerg. Mark. Financ. Trade* **2023**, *59*, 2118–2128. [CrossRef]
72. Li, C.; Shen, L.; Qian, G. Online Hybrid Neural Network for Stock Price Prediction: A Case Study of High-Frequency Stock Trading in the Chinese Market. *Econometrics* **2023**, *11*, 13. [CrossRef]
73. Xiang, X.; Wang, W. Predicting Intraday Trading Direction of CSI 300 Based on TCN Model. In Proceedings of the 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Jiuzhaigou, China, 25–29 July 2023; IEEE: New York, NY, USA, 2023; pp. 293–299.

74. Huang, W.; Gao, T.; Hao, Y.; Wang, X. Transformer-based forecasting for intraday trading in the Shanghai crude oil market: Analyzing open-high-low-close prices. *Energy Econ.* **2023**, *127*, 107106. [CrossRef]
75. Ding, Y.; Zhang, C.; Zhang, C. An Informer Based Method for Stock Intraday Price Prediction. In Proceedings of the 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Harbin, China, 29–31 July 2023; IEEE: New York, NY, USA, 2023; pp. 1–6.
76. Song, Y.; Cai, C.; Ma, D.; Li, C. Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model. *Expert Syst. Appl.* **2024**, *237*, 121527. [CrossRef]
77. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.
78. Brownlees, C.T.; Gallo, G.M. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Comput. Stat. Data Anal.* **2006**, *51*, 2232–2245. [CrossRef]
79. Dioubi, F.; Khurshid, A. How External Trends and Internal Components Decomposition Method Improve the Predictability of Financial Time Series? *J. Financ. Risk Manag.* **2022**, *11*, 621–633. [CrossRef]
80. Liu, Z.; Xia, X.; Zhou, G. Pre-averaging estimate of high dimensional integrated covariance matrix with noisy and asynchronous high-frequency data. *Random Matrices Theory Appl.* **2018**, *7*, 1850005. [CrossRef]
81. Wang, K.; Liu, X.; Ye, W. Intraday VaR: A copula-based approach. *J. Empir. Financ.* **2023**, *74*, 101419. [CrossRef]
82. Peng, Y.L.; Lee, W.P. Data selection to avoid overfitting for foreign exchange intraday trading with machine learning. *Appl. Soft Comput.* **2021**, *108*, 107461. [CrossRef]
83. Lai, S.; Wang, M.; Zhao, S.; Arce, G.R. Predicting high-frequency stock movement with differential transformer neural network. *Electronics* **2023**, *12*, 2943. [CrossRef]
84. Li, X.; Tang, P. Stock index prediction based on wavelet transform and FCD-MLGRU. *J. Forecast.* **2020**, *39*, 1229–1237. [CrossRef]
85. Sun, E.W.; Meinl, T. A new wavelet-based denoising algorithm for high-frequency financial data mining. *Eur. J. Oper. Res.* **2012**, *217*, 589–599. [CrossRef]
86. Hajiabotorabi, Z.; Kazemi, A.; Samavati, F.F.; Ghaini, F.M.M. Improving DWT-RNN model via B-spline wavelet multiresolution to forecast a high-frequency time series. *Expert Syst. Appl.* **2019**, *138*, 112842. [CrossRef]
87. Chen, Y.T.; Lai, W.N.; Sun, E.W. Jump detection and noise separation by a singular wavelet method for predictive analytics of high-frequency data. *Comput. Econ.* **2019**, *54*, 809–844. [CrossRef]
88. Chacón, H.D.; Kesici, E.; Najafirad, P. Improving financial time series prediction accuracy using ensemble empirical mode decomposition and recurrent neural networks. *IEEE Access* **2020**, *8*, 117133–117145. [CrossRef]
89. Radojičić, D.; Kredatus, S. The impact of stock market price fourier transform analysis on the gated recurrent unit classifier model. *Expert Syst. Appl.* **2020**, *159*, 113565. [CrossRef]
90. Stoikov, S. The micro-price: A high-frequency estimator of future prices. *Quant. Financ.* **2018**, *18*, 1959–1966. [CrossRef]
91. Schnaubelt, M.; Fischer, T.G.; Krauss, C. Separating the signal from the noise–financial machine learning for twitter. *J. Econ. Dyn. Control* **2020**, *114*, 103895. [CrossRef]
92. Peng, P.; Chen, Y.; Lin, W.; Wang, J.Z. Attention-based CNN–LSTM for high-frequency multiple cryptocurrency trend prediction. *Expert Syst. Appl.* **2024**, *237*, 121520. [CrossRef]
93. Zhang, X.; Huang, Y.; Xu, K.; Xing, L. Novel modelling strategies for high-frequency stock trading data. *Financ. Innov.* **2023**, *9*, 39. [CrossRef] [PubMed]
94. Lei, B.; Zhang, B.; Song, Y. Volatility forecasting for high-frequency financial data based on web search index and deep learning model. *Mathematics* **2021**, *9*, 320. [CrossRef]
95. Zhang, X.; Liu, S.; Zheng, X. Stock price movement prediction based on a deep factorization machine and the attention mechanism. *Mathematics* **2021**, *9*, 800. [CrossRef]
96. Naik, N.; Mohan, B.R. Intraday stock prediction based on deep neural network. *Natl. Acad. Sci. Lett.* **2020**, *43*, 241–246. [CrossRef]
97. Pham, M.C.; Anderson, H.M.; Duong, H.N.; Lajbcygier, P. The effects of trade size and market depth on immediate price impact in a limit order book market. *J. Econ. Dyn. Control* **2020**, *120*, 103992. [CrossRef]
98. Jayawardena, N.I.; Todorova, N.; Li, B.; Su, J.J. Volatility forecasting using related markets' information for the Tokyo stock exchange. *Econ. Model.* **2020**, *90*, 143–158. [CrossRef]
99. Gkillas, K.; Gupta, R.; Pierdzioch, C. Forecasting realized oil-price volatility: The role of financial stress and asymmetric loss. *J. Int. Money Financ.* **2020**, *104*, 102137. [CrossRef]
100. Buccheri, G.; Bormetti, G.; Corsi, F.; Lillo, F. A Score-Driven Conditional Correlation Model for Noisy and Asynchronous Data: An Application to High-Frequency Covariance Dynamics. *J. Bus. Econ. Stat.* **2021**, *39*, 920–936. [CrossRef]
101. Mantilla, P.; Dormido-Canto, S. A novel feature engineering approach for high-frequency financial data. *Eng. Appl. Artif. Intell.* **2023**, *125*, 106705. [CrossRef]

102. Tang, P.; Tang, X.; Yu, W. Intraday trend prediction of stock indices with machine learning approaches. *Eng. Econ.* **2023**, *68*, 60–81. [CrossRef]
103. Ntakaris, A.; Mirone, G.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Feature engineering for mid-price prediction with deep learning. *IEEE Access* **2019**, *7*, 82390–82412. [CrossRef]
104. Yin, T.; Liu, C.; Ding, F.; Feng, Z.; Yuan, B.; Zhang, N. Graph-based stock correlation and prediction for high-frequency trading systems. *Pattern Recognit.* **2022**, *122*, 108209. [CrossRef]
105. Ye, W.; Yang, J.; Chen, P. Short-term stock price trend prediction with imaging high frequency limit order book data. *Int. J. Forecast.* **2024**, *40*, 1189–1205. [CrossRef]
106. Aït-Sahalia, Y.; Mykland, P.A.; Zhang, L. Ultra high frequency volatility estimation with dependent microstructure noise. *J. Econom.* **2011**, *160*, 160–175. [CrossRef]
107. Xiu, S.; Palomar, D.P. Intraday Volatility-Volume Joint Modeling and Forecasting: A State-Space Approach. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; IEEE: New York, NY, USA, 2023; pp. 1395–1399.
108. Goluža, S.; Kovačević, T.; Bauman, T.; Kostanjčar, Z. Deep reinforcement learning with positional context for intraday trading. *Evol. Syst.* **2024**, *15*, 1865–1880. [CrossRef]
109. Moreno-Pino, F.; Zohren, S. Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions. *Quant. Financ.* **2024**, *24*, 1105–1127. [CrossRef]
110. Niu, Z.; Demirer, R.; Suleman, M.T.; Zhang, H.; Zhu, X. Do industries predict stock market volatility? Evidence from machine learning models. *J. Int. Financ. Mark. Institutions Money* **2024**, *90*, 101903. [CrossRef]
111. Wang, J.; Huang, Y.; Ma, F.; Chevallier, J. Does high-frequency crude oil futures data contain useful information for predicting volatility in the US stock market? New evidence. *Energy Econ.* **2020**, *91*, 104897. [CrossRef]
112. Shi, Y.; Ho, K.Y. News sentiment and states of stock return volatility: Evidence from long memory and discrete choice models. *Financ. Res. Lett.* **2021**, *38*, 101446. [CrossRef]
113. Hussain, S.M.; Omrane, W.B.; Al-Yahyaee, K. US macroeconomic news effects around the US and European financial crises: Evidence from Brazilian and Mexican equity indices. *Glob. Financ. J.* **2020**, *46*, 100482. [CrossRef]
114. Kraaijeveld, O.; De Smedt, J. The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *J. Int. Financ. Mark. Institutions Money* **2020**, *65*, 101188. [CrossRef]
115. Jacod, J.; Li, Y.; Mykland, P.A.; Podolskij, M.; Vetter, M. Microstructure noise in the continuous case: The pre-averaging approach. *Stoch. Processes Their Appl.* **2009**, *119*, 2249–2276. [CrossRef]
116. Bandi, F.M.; Russell, J.R. Market microstructure noise, integrated variance estimators, and the accuracy of asymptotic approximations. *J. Econom.* **2011**, *160*, 145–159. [CrossRef]
117. Wu, Z. On the intraday periodicity duration adjustment of high-frequency data. *J. Empir. Financ.* **2012**, *19*, 282–291. [CrossRef]
118. Antulov-Fantulin, N.; Guo, T.; Lillo, F. Temporal mixture ensemble models for probabilistic forecasting of intraday cryptocurrency volume. *Decis. Econ. Financ.* **2021**, *44*, 905–940. [CrossRef]
119. Wang, C.; Sun, Y. An Ensemble De-noising Method for High Frequency Financial Data. In Proceedings of the 2014 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014), Singapore, 5–6 August 2014; Atlantis Press: Dordrecht, The Netherlands, 2014; pp. 27–33.
120. Silva, E.; Brandão, H.; Castilho, D.; Pereira, A.C. A binary ensemble classifier for high-frequency trading. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; IEEE: New York, NY, USA, 2015; pp. 1–8.
121. Jeon, S.; Hong, B.; Chang, V. Pattern graph tracking-based stock price prediction using big data. *Future Gener. Comput. Syst.* **2018**, *80*, 171–187. [CrossRef]
122. Kong, A.; Zhu, H.; Azencott, R. Predicting intraday jumps in stock prices using liquidity measures and technical indicators. *J. Forecast.* **2021**, *40*, 416–438. [CrossRef]
123. Tang, C.; Shi, Y. Forecasting high-dimensional financial functional time series: An application to constituent stocks in Dow Jones index. *J. Risk Financ. Manag.* **2021**, *14*, 343. [CrossRef]
124. Dai, C.; Lu, K.; Xiu, D. Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. *J. Econom.* **2019**, *208*, 43–79. [CrossRef]
125. Dong, Y.; Tse, Y.K. Forecasting large covariance matrix with high-frequency data using factor approach for the correlation matrix. *Econ. Lett.* **2020**, *195*, 109465. [CrossRef]
126. Chen, D. High frequency principal component analysis based on correlation matrix that is robust to jumps, microstructure noise and asynchronous observation times. *J. Econom.* **2024**, *240*, 105701. [CrossRef]
127. Aït-Sahalia, Y.; Xiu, D. Principal component analysis of high-frequency data. *J. Am. Stat. Assoc.* **2019**, *114*, 287–303. [CrossRef]
128. Manickavasagam, J.; Visalakshmi, S.; Apergis, N. A novel hybrid approach to forecast crude oil futures using intraday data. *Technol. Forecast. Soc. Chang.* **2020**, *158*, 120126. [CrossRef]

129. Mailagaha Kumbure, M.; Lohrmann, C.; Luukka, P. A Study on Relevant Features for Intraday S&P 500 Prediction Using a Hybrid Feature Selection Approach. In Proceedings of the International Conference on Machine Learning, Optimization, and Data Science, Grasmere, UK, 4–8 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 93–104.
130. Liou, J.H.; Liu, Y.T.; Cheng, L.C. Price spread prediction in high-frequency pairs trading using deep learning architectures. *Int. Rev. Financ. Anal.* **2024**, *96*, 103793. [CrossRef]
131. Sun, S.; Xue, W.; Wang, R.; He, X.; Zhu, J.; Li, J.; An, B. DeepScalper: A risk-aware reinforcement learning framework to capture fleeting intraday trading opportunities. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, 17–21 October 2022; pp. 1858–1867.
132. Noh, Y.; Kim, J.M.; Hong, S.; Kim, S. Deep Learning Model for Multivariate High-Frequency Time-Series Data: Financial Market Index Prediction. *Mathematics* **2023**, *11*, 3603. [CrossRef]
133. Tran, D.T.; Magris, M.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Tensor representation in high-frequency financial data for price change prediction. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November–1 December 2017; IEEE: New York, NY, USA, 2017; pp. 1–7.
134. Passalis, N.; Tefas, A.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *4*, 774–785. [CrossRef]
135. Tran, D.T.; Iosifidis, A.; Kanniainen, J.; Gabbouj, M. Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 1407–1418. [CrossRef]
136. Tashiro, D.; Matsushima, H.; Izumi, K.; Sakaji, H. Encoding of high-frequency order information and prediction of short-term stock price by deep learning. *Quant. Financ.* **2019**, *19*, 1499–1506. [CrossRef]
137. Kercheval, A.N.; Zhang, Y. Modelling high-frequency limit order book dynamics with support vector machines. *Quant. Financ.* **2015**, *15*, 1315–1329. [CrossRef]
138. Lahmiri, S.; Bekiros, S. Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market. *Chaos Solitons Fractals* **2020**, *133*, 109641. [CrossRef]
139. Huddleston, D.; Liu, F.; Stentoft, L. Intraday market predictability: A machine learning approach. *J. Financ. Econom.* **2023**, *21*, 485–527. [CrossRef]
140. Choudhury, S.; Ghosh, S.; Bhattacharya, A.; Fernandes, K.J.; Tiwari, M.K. A real time clustering and SVM based price-volatility prediction for optimal trading strategy. *Neurocomputing* **2014**, *131*, 419–426. [CrossRef]
141. Dixon, M. Sequence classification of the limit order book using recurrent neural networks. *J. Comput. Sci.* **2018**, *24*, 277–286. [CrossRef]
142. Khushaba, R.N.; Al-Ani, A.; Al-Jumaily, A. Feature subset selection using differential evolution and a statistical repair mechanism. *Expert Syst. Appl.* **2011**, *38*, 11515–11526. [CrossRef]
143. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
144. Barndorff-Nielsen, O.E.; Hansen, P.R.; Lunde, A.; Shephard, N. Realized kernels in practice: Trades and quotes. *Econom. J.* **2009**, *12*, C1–C32. [CrossRef]
145. Gençay, R.; Dacorogna, M.; Muller, U.A.; Pictet, O.; Olsen, R. *An Introduction to High-frequency Finance*; Elsevier: Amsterdam, The Netherlands, 2001.
146. Verousis, T.; Ap Gwilym, O. An improved algorithm for cleaning ultra high-frequency data. *J. Deriv. Hedge Funds* **2010**, *15*, 323–340. [CrossRef]
147. Barndorff-Nielsen, O.E.; Hansen, P.R.; Lunde, A.; Shephard, N. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *J. Econom.* **2011**, *162*, 149–169. [CrossRef]
148. Harris, F.H.d.; McInish, T.H.; Shoesmith, G.L.; Wood, R.A. Cointegration, error correction, and price discovery on informationally linked security markets. *J. Financ. Quant. Anal.* **1995**, *30*, 563–579. [CrossRef]
149. Boudt, K.; Kleen, O.; Sjørup, E. Analyzing Intraday Financial Data in R: The highfrequency Package. *J. Stat. Softw.* **2022**, *104*, 1–36. [CrossRef]
150. Fan, J.; Li, Y.; Yu, K. Vast volatility matrix estimation using high-frequency data for portfolio selection. *J. Am. Stat. Assoc.* **2012**, *107*, 412–428. [CrossRef]
151. Hwang, E.; Shin, D.W. Two-stage stationary bootstrapping for bivariate average realized volatility matrix under market microstructure noise and asynchronicity. *J. Econom.* **2018**, *202*, 178–195. [CrossRef]
152. Chakrabarti, A.; Sen, R. Copula estimation for nonsynchronous financial data. *Sankhya B* **2023**, *85*, 116–149. [CrossRef]
153. Zhang, L. Estimating covariation: Epps effect, microstructure noise. *J. Econom.* **2011**, *160*, 33–47. [CrossRef]
154. Shephard, N.; Xiu, D. Econometric analysis of multivariate realised QML: Estimation of the covariation of equity prices under asynchronous trading. *J. Econom.* **2017**, *201*, 19–42. [CrossRef]
155. Bahcivan, H.; Karahan, C.C. High frequency correlation dynamics and day-of-the-week effect: A score-driven approach in an emerging market stock exchange. *Int. Rev. Financ. Anal.* **2022**, *80*, 102008. [CrossRef]

156. Corsi, F.; Peluso, S.; Audrino, F. Missing in asynchronicity: A Kalman-em approach for multivariate realized covariance estimation. *J. Appl. Econom.* **2015**, *30*, 377–397. [CrossRef]
157. Ho, M.; Xin, J. Sparse Kalman filtering approaches to realized covariance estimation from high frequency financial data. *Math. Program.* **2019**, *176*, 247–278. [CrossRef]
158. Christensen, K.; Kinnebrock, S.; Podolskij, M. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *J. Econom.* **2010**, *159*, 116–133. [CrossRef]
159. Chen, D.; Mykland, P.A.; Zhang, L. Realized regression with asynchronous and noisy high frequency and high dimensional data. *J. Econom.* **2024**, *239*, 105446. [CrossRef]
160. Tao, M.; Wang, Y.; Yao, Q.; Zou, J. Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Am. Stat. Assoc.* **2011**, *106*, 1025–1040. [CrossRef]
161. Epps, T.W. Comovements in stock prices in the very short run. *J. Am. Stat. Assoc.* **1979**, *74*, 291–298.
162. Aït-Sahalia, Y.; Fan, J.; Xiu, D. High-frequency covariance estimates with noisy and asynchronous financial data. *J. Am. Stat. Assoc.* **2010**, *105*, 1504–1517. [CrossRef]
163. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
164. Sanz, J.A.; Bernardo, D.; Herrera, F.; Bustince, H.; Hagras, H. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Trans. Fuzzy Syst.* **2014**, *23*, 973–990. [CrossRef]
165. Uzun, S.; Sensoy, A.; Nguyen, D.K. Jump forecasting in foreign exchange markets: A high-frequency analysis. *J. Forecast.* **2023**, *42*, 578–624. [CrossRef]
166. Farquad, M.A.H.; Bose, I. Preprocessing unbalanced data using support vector machine. *Decis. Support Syst.* **2012**, *53*, 226–233. [CrossRef]
167. Wang, Q.; Xu, W.; Huang, X.; Yang, K. Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning. *Neurocomputing* **2019**, *347*, 46–58. [CrossRef]
168. Engle, R.F.; Sokalska, M.E. Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *J. Financ. Econom.* **2012**, *10*, 54–83. [CrossRef]
169. Hua, L. Discovering Intraday Tail Dependence Patterns via a Full-Range Tail Dependence Copula. *Risks* **2023**, *11*, 195. [CrossRef]
170. Summinga-Sonagadu, R.; Narsoo, J. Risk model validation: An intraday VaR and ES approach using the multiplicative component GARCH. *Risks* **2019**, *7*, 10. [CrossRef]
171. Karmakar, M.; Paul, S. Intraday portfolio risk management using VaR and CVaR: A CGARCH-EVT-Copula approach. *Int. J. Forecast.* **2019**, *35*, 699–709. [CrossRef]
172. Taylor, S.J.; Xu, X. The incremental volatility information in one million foreign exchange quotations. *J. Empir. Financ.* **1997**, *4*, 317–340. [CrossRef]
173. Bollen, B.; Inder, B. Estimating daily volatility in financial markets utilizing intraday data. *J. Empir. Financ.* **2002**, *9*, 551–562. [CrossRef]
174. Den Haan, W.J.; Levin, A.T. *Inferences from Parametric and Non-Parametric Covariance Matrix Estimation Procedures*; National Bureau of Economic Research: Cambridge, MA, USA, 1996.
175. Omrane, W.B.; de Bodt, E. Using self-organizing maps to adjust for intra-day seasonality. *J. Bank. Financ.* **2007**, *31*, 1817–1838. [CrossRef]
176. Zaznov, I.; Kunkel, J.; Dufour, A.; Badii, A. Predicting stock price changes based on the limit order book: A survey. *Mathematics* **2022**, *10*, 1234. [CrossRef]
177. Thakkar, A.; Chaudhari, K. A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Syst. Appl.* **2021**, *177*, 114800. [CrossRef]
178. Sahu, S.K.; Mokhade, A.; Bokde, N.D. An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Appl. Sci.* **2023**, *13*, 1956. [CrossRef]
179. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
180. Shah, J.; Vaidya, D.; Shah, M. A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intell. Syst. Appl.* **2022**, *16*, 200111. [CrossRef]
181. Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Syst. Appl.* **2022**, *197*, 116659. [CrossRef]
182. Olorunnimbe, K.; Viktor, H. Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artif. Intell. Rev.* **2023**, *56*, 2057–2109. [CrossRef] [PubMed]
183. Rouf, N.; Malik, M.B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H.C. Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics* **2021**, *10*, 2717. [CrossRef]

184. Nazareth, N.; Reddy, Y.V.R. Financial applications of machine learning: A literature review. *Expert Syst. Appl.* **2023**, *219*, 119640. [CrossRef]
185. Singh, V.; Chen, S.S.; Singhania, M.; Nanavati, B.; Gupta, A. How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100094. [CrossRef]
186. Ayitey Junior, M.; Appiahene, P.; Appiah, O.; Bombie, C.N. Forex market forecasting using machine learning: Systematic Literature Review and meta-analysis. *J. Big Data* **2023**, *10*, 9. [CrossRef]
187. Hoang, D.; Wiegratz, K. Machine learning methods in finance: Recent applications and prospects. *Eur. Financ. Manag.* **2023**, *29*, 1657–1701. [CrossRef]
188. Ahmed, S.; Alshater, M.M.; El Ammari, A.; Hammami, H. Artificial intelligence and machine learning in finance: A bibliometric review. *Res. Int. Bus. Financ.* **2022**, *61*, 101646. [CrossRef]
189. Ashtiani, M.N.; Raahemi, B. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Syst. Appl.* **2023**, *217*, 119509. [CrossRef]
190. Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; Wang, J. Deep time series models: A comprehensive survey and benchmark. *arXiv* **2024**, arXiv:2407.13278.