LOB-Based Deep Learning Models for Stock Price Trend Prediction: A Benchmark Study

Matteo Prata[*], **Giuseppe Masi**[*], **Leonardo Berti**[*], **Viviana Arrigoni**[*], **Andrea Coletta**[†], **Irene Cannistraci**[*], **Svitlana Vyetrenko**[†], **Paola Velardi**[*], **Novella Bartolini**[*]
[*]Department of Computer Science, Sapienza University of Rome, Italy.
[†]J.P. Morgan AI Research, New York, USA.
{prata, masi.g, arrigoni, cannistraci, velardi, bartolini}@di.uniroma1.it
{andrea.coletta, svitlana.s.vyetrenko}@jpmchase.com

Abstract

The recent advancements in Deep Learning (DL) research have notably influenced the finance sector. We examine the robustness and generalizability of fifteen state-of-the-art DL models focusing on Stock Price Trend Prediction (SPTP) based on Limit Order Book (LOB) data. To carry out this study, we developed LOBCAST, an open-source framework that incorporates data preprocessing, DL model training, evaluation and profit analysis. Our extensive experiments reveal that all models exhibit a significant performance drop when exposed to new data, thereby raising questions about their real-world market applicability. Our work serves as a benchmark, illuminating the potential and the limitations of current approaches and providing insight for innovative solutions.

1 Introduction

Predicting stock market prices is a complex endeavour due to myriad factors, including macroeconomic conditions and investor sentiment [1]. Nevertheless, professional traders and researchers usually forecast price movements by understanding key market properties, such as volatility or liquidity, and recognizing patterns to anticipate future market trends [2]. Effective mathematical models are essential for capturing complex market dependencies. The recent surge in artificial intelligence has led to significant work in using machine learning algorithms to predict future market trends [3–5]. Recent Deep Learning (DL) models have achieved over 88% in F1-score in predicting market trends in simulated settings using historical data [6]. However, replicating these performances in real markets is challenging, suggesting a possible *simulation-to-reality* gap [7, 8].

In this paper, we benchmark the most recent and promising DL approaches to Stock Price Trend Prediction (SPTP) based on Limit Order Book (LOB) data, one of the most valuable information sources available to traders on the stock markets. Our benchmark evaluates their robustness and generalizability [9–11]. In particular, we assess the models' robustness by comparing the stated performance with our reproduced results on the same dataset FI-2010 [12]. We also assess their generalizability by testing their performance on unseen market scenarios using LOBSTER data [13]. We focus on novel data-driven approaches from Machine Learning (ML) and DL that analyze the market at its finest resolution, using high-frequency LOB data. In this work, we formally define the SPTP problem considering a ternary trend classification. Our findings reveal that while best models exhibit robustness, achieving solid F1-scores on FI-2010, they show poor generalizability, as their performance significantly drops when applied to unseen LOBSTER market data.

Preprint. Under review.

The main contributions of our work are the following:

- We release a highly modular open-source framework called **LOBCAST**[1], to pre-process data, train, and test stock market models. Our framework employs the latest DL libraries to provide all researchers an easy, performing, and maintainable solution. Furthermore, to support future studies, we release two meta-learning models and a backtesting environment for profit analysis.
- We evaluate existing LOB-based stock market trend predictors, showing that most of them overfit the FI-2010 dataset, with remarkably lower performance on unseen stock data.
- We survey and discuss the financial performance of existing methods under different market scenarios to guide model selection in real-world applications[2].
- We discuss the strengths and limitations of existing methodology and identify areas for future research toward more reliable, robust, and reproducible approaches to stock market prediction.

2 Related Work

The increasing interest in DL for price trend prediction motivated several researchers to collect and analyze State-Of-the-Art (SOTA) solutions in benchmark surveys. The study by Jiang et al. [4] analyzes papers published between 2017 and 2019 that focused on stock price and market index prediction. In their literature review, the authors studied the Neural Network (NN) structures and evaluation metrics used in selected papers, as well as implementation and reproducibility. This work was extended in [14], including an in-depth analysis of the data (i.e., market indices, input variables used for stock market predictions). Ozboyoglu et al. [15] and Sezer et al. [5] provide a comprehensive overview of the SOTA DL models used for financial predictions. The work in [16] surveys 86 papers on stock and foreign exchange price prediction. The authors review the datasets, variables, models, and performance metrics used in each surveyed article. In contrast to these works, in this paper, we run experiments to study the robustness and generalizability of the selected approaches. Nti et al. [17] conducted a systematic and critical review of 122 papers. Their study also compares the self-stated accuracy, error metrics, and software packages used in the selected papers by means of experiments. In contrast to this, we focus on papers that use LOB data and DL algorithms for price trend predictions. We also evaluate the generalizability of the models by driving tests on a different dataset. Other studies [18, 19] also analyze solutions based on sentiment analysis through Natural Language Processing (NLP) to investigate the impact of social media on the stock market, showing that this combination improves the accuracy of stock prediction models. In [20], the authors presented a comprehensive overview of traditional and ML-based approaches for stock market prediction and highlighted some limitations of traditional approaches, showing that DL models outperform them in terms of accuracy. Similar findings are reported in [21]. Lim et al. [22] discussed recent developments in hybrid DL models, which combine statistical and learning components for both one-step-ahead and multi-horizon time-series forecasting. Similarly, Shah et al. [23] discussed hybrid approaches in their work on the state-of-the-art algorithms commonly applied to stock market prediction. Additionally, they provided a taxonomy of computational approaches for stock market analysis and prediction. Finally, Olorunnimbe et al. [24] focused on exploring applications of DL in the stock market that involve backtesting, with a particular emphasis on research papers that meet the requirements for real-world use. They reviewed various scenarios in which DL has been employed in finance, with a focus on trade strategy, price prediction, portfolio management, and others.

Our work adds to this literature by providing a benchmark of recent deep learning approaches based on LOB data, evaluating their robustness and generalizability, and releasing an open-source framework for pre-processing data, training, and testing models.

3 Stock Price Trend Prediction

The common ground that unifies the models studied in this paper is the goal of solving the SPTP problem via Deep Neural Networks (DNNs) trained on LOB data. LOB data are particularly enlightening as they provide raw and granular information on stocks' trades. By observing the LOB in a fixed period of time, SPTP models return a distribution over the possible future market movements.

[1] The code is included in the supplementary material and will be publicly available upon acceptance

[2] The details are reported in the supplementary materials for space reasons

Limit Order Book A stock exchange employs a matching engine for storing and matching the orders issued by the trading agents. This is achieved by updating the so-called Limit Order Book (LOB) data structure. Each security (tradable asset) has a LOB, recording all the outstanding bid and ask orders currently available on an exchange or a trading platform. The shape of the order book gives traders a simultaneous view of the market demand and supply.

There are three major types of orders. *Market orders* are executed immediately at the best available price. *Limit orders*, instead, include the specification of a desired target price: a limit sell [buy] order will be executed only when it is matched to a buy [sell] order whose price is greater [lower] than or equal to the target price. Finally, a *Cancel order* removes a previously submitted limit order.

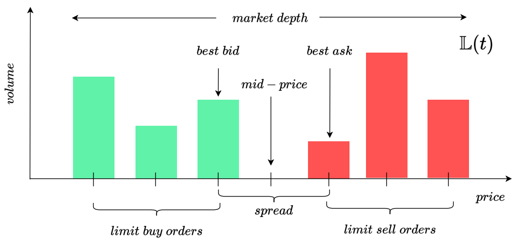

Figure 1: An example of LOB.

Figure 1 depicts an example of a LOB snapshot, characterized by *buy* orders (*bid*) and *sell* orders (*ask*) of different prices. A *level*, shown on the horizontal axis, represents the number of shares with the same price either on the bid or ask side. In the example of Figure 1, there are three bid and three ask levels. The *best bid* is the price of the shares with the highest price on the buy side; analogously, the *best ask* is the price of the shares with the lowest price on the bid side. When the former exceeds or equals the latter, the corresponding limit ask and bid orders are executed. The LOB is updated with each event (order insertion/modification/cancellation) and can be sampled at regular time intervals.

We represent the evolution of a LOB as a time series $\mathbb{L}$, where each $\mathbb{L}(t) \in \mathbb{R}^{4L}$ is called a LOB record, for $t = 1, \ldots, N$, being N the number of LOB observations and L the number of levels. In particular, $\mathbb{L}(t) = \{P^s(t), V^s(t)\}_{s \in \{\texttt{ask},\texttt{bid}\}}$, where $P^{\texttt{ask}}(t), P^{\texttt{bid}}(t) \in \mathbb{R}^L$ represent the prices of levels 1 to L of the LOB, on the *ask* ($s = \texttt{ask}$) side and *bid* ($s = \texttt{bid}$) side, respectively, at time t. Analogously, $V^{\texttt{ask}}(t), V^{\texttt{bid}}(t) \in \mathbb{R}^L$ represent the volumes. This means that for each t and every $j \in \{1, \ldots, L\}$ on the *ask* side, $V_j^{\texttt{ask}}(t)$ shares can be sold at price $P_j^{\texttt{ask}}(t)$. The *mid-price* $m(t)$ of the stock at time t, is defined as the average value between the best bid and the best ask, $m(t) = \frac{P^{\texttt{ask}}(t)+P^{\texttt{bid}}(t)}{2}$. Mid-prices are synthetic values that are commonly used as indicators of the stock price trend. In average, if most of the executed orders are on the ask [bid] side, the mid-price increases [decreases] accordingly.

Trend Definition We use a ternary classification for trends: U ("upward") if the price trend is increasing; D ("downward") for decreasing prices; and S ("stable") for prices with negligible variations. Thanks to their informativeness, mid-prices are well-suited to drive this classification. Nevertheless, because of the market's inherent fluctuations and shocks, they can exhibit highly volatile trends. For this reason, using a direct comparison of consecutive mid-prices, i.e., $m(t)$ and $m(t+1)$, for stock price labelling would result in a noisy labelled dataset. As a result, labelling strategies typically employ smoother mid-price functions instead of raw mid-prices. Such functions consider mid-prices over arbitrarily long time intervals, called *horizons*. Our experiments adopt the labelling proposed in [12] and repurposed in several other state-of-the-art solutions we selected for benchmarking. The adopted labelling strategy compares the current mid-price to the average mid-prices $a^+(k,t)$ in a future *horizon* of k time units, formally:

$$a^+(k,t) = \frac{1}{k}\sum_{i=1}^{k} m(t+i). \tag{1}$$

The average mid-prices are used to define a static threshold $\theta \in (0,1)$ that is used to identify an interval around the current mid-price and define the class of the trend at time t as follows:

$$\texttt{U}\colon a^+(k,t) > m(t)(1+\theta), \ \texttt{D}\colon a^+(k,t) < m(t)(1-\theta), \ \texttt{S}\colon a^+(k,t) \in [m(t)(1-\theta), m(t)(1+\theta)]. \tag{2}$$

With this labelling, we beat the effect of mid-price fluctuations by considering their average over a desired horizon k and considering a trend to be stable when the average mid-price variations do not change significantly, thus avoiding over-fitting. We highlight that time stamps t can come either from a homogeneous or an event-based process. In our experiments, we consider an event-based process.

Models I/O Given the time series of a LOB $\mathbb{L}$ and a temporal window $T = [t - h, t]$, $h \in \mathbb{N}$, we can extract *market observations* on T, $\mathbb{M}(T)$, by considering the sub-sequence of LOB observations starting from time $t - h$ up to t. In Section 1 of the Supplemental Material (SUP), we give a representation of a market observation $\mathbb{M}(T) \in \mathbb{R}^{h \times 4L}$. The market observation over the window $[t - h, t]$ is associated with the label computed through Equations 1 and 2 at time t. An SPTP predictor takes as an input a market observation and outputs a probability distribution over the trend classes U, D, and S.

4 Experiments

We conducted an extensive evaluation to assess the ***robustness*** and ***generalizability*** of 15 DL models to solve the SPTP task, as presented in Section 3. Among these, 13 were SOTA models, and 2 DL baseline models commonly used in the literature. More details on the models are given in Section 4.2.

In line with many other studies, we adopt the definition of robustness and generalizability introduced by J. Pineau et al. in their work [9]. Robustness is evaluated by testing the proposed models on **FI-2010**, a benchmark dataset employed in all surveyed papers. In some cases, the authors of the considered works have not provided crucial information, such as the code or the hyperparameters of their models, making reimplementation and hyperparameter search necessary. We refer to Section 5.1 in SUP for a complete description of the hyperparameters search. To evaluate the generalizability, we created two datasets called **LOB-2021** and **LOB-2022**, extrapolated from the LOBSTER dataset [13]. We describe these datasets in Section 4.1.

Our experiments were carried out using **LOBCAST** [25], the open-source framework we make available online. The framework allows the definition of new price trend predictors based on LOB data. More details on the framework are given in Section 4.3.

4.1 Datasets

LOB data are not often publicly available and very expensive: stock exchanges (e.g., NASDAQ) provide fine-grained data only for high fees. The high cost and low availability restrict the application and development of DL algorithms in the research community.

The most widely spread public LOB dataset is **FI-2010** which is licensed under *Creative Commons Attribution 4.0 International (CC BY 4.0)* and was proposed in 2017 by Ntakaris et al. [12] with the objective of evaluating the performance of machine learning models on the SPTP task. The dataset consists of LOB data from five Finnish companies: Kesko Oyj, Outokumpu Oyj, Sampo, Rautaruukki, and Wärtsilä Oyj of the NASDAQ Nordic stock market. Data spans the time period between June 1st to June 14th, 2010, corresponding to 10 trading days (trading happens only on business days). About 4 million limit order messages are stored for 10 levels of the LOB. The dataset has an event-based granularity, meaning that the time series records are not uniformly spaced in time. LOB observations are sampled at intervals of 10 *events*, resulting in a total of 394,337 events. This dataset has the intrinsic limitation of being already pre-processed (filtered, normalized, and labelled) so that the original LOB cannot be backtracked, thus hampering thorough experimentation. Additionally, the labelling method employed is found to be prone to instability, as demonstrated by Zhang et al. in [26]. Moreover, the dataset is unbalanced at varying prediction horizons. Varying the horizon $k \in \mathcal{K} = \{1, 2, 3, 5, 10\}$, the stationary class S is progressively less predominant in favour of the upward and downward classes. For instance, the class composition for different values of k is $k = 1$, U: 18%, S: 63%, D:19%; $k = 5$, U: 32%, S: 35%, D:33%; $k = 10$, U: 37%, S: 25%, D:38%.

To test the generalizability of the models in a more realistic scenario, we used data extracted from **LOBSTER** [13], an online LOB data provider for order book data, which is not available for free, as is often the case for critical applications such as health and finance [9]. The data are reconstructed from NASDAQ traded stocks and are publicly available for the research community with an annual fee. To compare the performance of the algorithms in a wide range of scenarios, we have created a

	Tsantekidis et al. [27] MLP (2017)	Tsantekidis et al. [27] LSTM (2017)	Tsantekidis et al. [28] CNN1 (2017)	Tran et al. [29] CTABL (2018)	Zhang et al. [26] DEEPLOB (2019)	Passalis et al. [30] DAIN (2019)	Tsantekidis et al. [31] CNNLSTM (2020)	Tsantekidis et al. [31] CNN2 (2020)	Wallbridge et al. [32] TRANSLOB (2020)	Passalis et al. [33] TLONBoF (2020)	Tran et al. [6] BINCTABL (2021)	Zhang et al. [34] DEEPLOBATT (2021)	Guo et al. [35] DLA (2022)	Tran et al. [36] ATNBoF (2022)	Kisiel et al. [37] AXIALLOB (2021)
temporal shape (h)	100	100	100	10	100	15	300	300	100	15	10	50	5	100	40
features shape	40	40	40	40	40	144	42	40	40	144	40	40	144	40	40
code available	✗	✗	✗	TF	PT	PT	✗	✗	TF	PT	✗	TF	✗	PT	✗
n. trainable parameters	$1.0 \cdot 10^6$	$1.6 \cdot 10^4$	$3.5 \cdot 10^4$	$1.1 \cdot 10^4$	$1.4 \cdot 10^5$	$2.1 \cdot 10^6$	$5.3 \cdot 10^4$	$2.8 \cdot 10^5$	$1.1 \cdot 10^5$	$6.5 \cdot 10^5$	$1.1 \cdot 10^4$	$1.8 \cdot 10^5$	$1.2 \cdot 10^5$	$1.3 \cdot 10^7$	$2.0 \cdot 10^4$
inference time (*ms*)	0.08	0.21	0.36	0.48	1.31	0.15	0.50	0.49	2.40	0.43	0.71	1.73	0.23	3.90	1.91

Table 1: Relevant characteristics of the selected models.

large LOB dataset, including several stocks and time periods. The chosen pool of stocks includes those from the top 50% more liquid stocks of NASDAQ. To create a challenging evaluation scenario, we selected six stocks, namely: SoFi Technologies (SOFI), Netflix (NFLX), Cisco Systems (CSCO), Wing Stop (WING), Shoals Technologies Group (SHLS), and Landstar System (LSTR). The periods in consideration are *July 2021* (2021-07-01 to 2021-07-15, 10 trading days) making up **LOB-2021**, and *February 2022* (2022-02-01 to 2022-02-15, 10 trading days) making up **LOB-2022**. The selection of these two periods aimed to capture data from periods with different levels of market volatility. February 2022 exhibited higher volatility compared to July 2021, largely influenced by the Ukrainian crisis. This allows for an assessment of models across varying market conditions. We describe in detail our stock selection procedure in Section 3 in SUP.

Datasets for the Generalizability Study Due to copyright reasons, we are unable to release the LOB-2021 and LOB-2022 datasets. However, in Section 4 in SUP, we provide detailed insights into how they are generated, ensuring transparency and replicability in future research. The approach we adopt to generate both datasets closely follows the creation process presented for FI-2010 in [12]. In summary, for each considered stock s, we construct a *stock time series* of LOB records $\mathbb{L}_s(t) \in \mathbb{R}^{4L}$, with $L = 10$. To resemble the FI-2010 structure, we sample the market observation every 10 events and split records into *training*, *validation*, and *testing* sets using a 6-2-2 days split[3]. Normalization is performed on stock time series using a z-score approach, and the dataset is labelled by leveraging the trend definitions described in Equation (2). Lastly, both LOB-2021 and LOB-2022 contain prediction labels for each one of the considered horizons $\mathcal{K}$.

4.2 Models

We have selected 13 SOTA models based on DL for the SPTP task. These models were proposed in papers published between 2017 and 2022 and utilized datasets LOB data for training and testing. In addition to the models proposed in the selected papers, we also included two classical DL algorithms, namely Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN), which were used as a benchmark in [27] and in [31], respectively. All proposed models are based on DNNs and were originally trained and tested on the FI-2010 dataset.

A comprehensive summary of the benchmarked models can be found in Table 1, while for additional details, we refer the reader to Section 2 in SUP. In Table 1, the *temporal shape* represents the length of the input market observation for the model. The *features shape* refers to the number of features used by the models to infer the trend in the original papers. In the Table, we also indicate whether the authors released the code, and if so, whether they have used PyTorch (PT) [38] or Tensorflow (TF) [39]. This is relevant because to ensure consistency and compatibility within our proposed framework, based on PyTorch Lightning, we found it necessary to re-implement models for which the code was not available or was only available in Tensorflow. To improve the reproducibility of the results, it is advisable for the research community to publish the code developed.

In High-Frequency Trading (HFT) and algorithmic trading in general, minimizing latency between model querying and order placement is of utmost importance [40]. To explore this aspect, we analyzed the inference time in milliseconds of all models, based on the experiments reported in Sec-

[3] For the experiments on FI-2010 we followed the same data splitting procedure as the 13 SOTA papers. We split the dataset using the first 7 days for the train set and validation set (80% / 20%) and the last three days as the test set.

tion 4.4. As shown in Table 1, DEEPLOB, DEEPLOBAT, AXIALLOB, TRANSLOB, and ATNBoF had inference times in the order of milliseconds, potentially unsuitable for HFT applications compared to other models with shorter times. Finally, we have reported the number of trainable parameters for each model. A noteworthy observation is that the average number of parameters is very low compared to other classical fields, such as computer vision [41] and natural language processing [42, 43]. This leads us to conjecture that current systems are inadequate in effectively handling the complexity of LOB data, as we will verify in the rest of this paper.

To explore the possibility of achieving new SOTA performance by combining the predictions of all 15 models, we have implemented two ensemble methods: *MAJORITY*, which performs a majority voting weighted by the F1-Score of the predictions made by all the models, and *METALOB*, which is trained with the predictions made by the individual models to learn the most appropriate aggregation function. A detailed description of these ensemble methods can be found in Section 2.1 in SUP.

4.3 LOBCAST Framework for SPTP

We present **LOBCAST** [25], a Python-based framework developed for stock market trend forecasting using LOB data. LOBCAST is an open-source framework that enables users to test DL models for the SPTP task. The framework provides data pre-processing functionalities, which include normalization, splitting, and labelling. LOBCAST also offers a comprehensive training environment for DL models implemented in PyTorch Lightning [38]. It integrates interfaces with the popular hyperparameter tuning framework WANDB [44], which allows users to tune and optimize model performance efficiently. The framework generates detailed reports for the trained models, including performance metrics regarding the learning task (F1, Accuracy, Recall, etc.). LOBCAST supports backtesting for profit analysis, utilizing the Backtesting.py [45] external library. This feature enables users to assess the profitability of their models in simulated trading scenarios. We plan to add new features such as (i) training and testing with different LOB representations [46, 47], and (ii) test on adversarial perturbations to evaluate the representations' robustness [48]. We believe that LOBCAST, along with the advancements in DL models and the utilization of LOB data, has the potential to improve the state of the art on trend forecasting in the financial domain.

4.4 Performance, Robustness and Generalizability

To test robustness and generalizability, we conducted our experiments for each model using five different seeds to ensure reliable results and mitigate the impact of random initialization of network weights and training dataset shuffling. The training process involved training the 15 models for each seed on each of the considered prediction horizons ($\mathcal{K} = \{1, 2, 3, 5, 10\}$). More details on the setting of the experiments are provided in the SUP Section 5. On average over all 5 runs, the training process for all the models took approximately 155 hours for FI-2010 and 258 hours for LOB-2021/2022, utilizing a cluster comprised of 8 GPUs (1 NVIDIA GeForce RTX 2060, 2 NVIDIA GeForce RTX 3070, and 5 NVIDIA Quadro RTX 6000).

In Table 2, we summarize the results of our experiments. As the datasets are not well balanced, we focused on F1-score; other performance metrics are reported in the SUP. The Table compares the claimed performance of each system (column F1 Claim) with those measured in the robustness (FI-2010) and generalizability (LOB-2021 and 2022) experiments. For each dataset, we show the average performance and the standard deviation achieved by each model in all the horizons, along with its rank.

To evaluate the robustness and the generalizability of the models, we compute the **robustness** and the **generalizability scores**, a value ≤ 100 that is computed as $100 - (|A| + S)$, where A and S are defined as follows. A is the average difference between the F1-score reported in the original paper and the one that we observed in our experiments on FI-2010 for robustness, and on LOB-2021 and LOB-2022 for generalizability. S is the standard deviation of these differences. The score penalizes models that demonstrate higher variability in their performance by subtracting the standard deviation. The average and standard deviation were computed over the declared horizons for each model and considering all five seeds.

Table 2 clearly highlights the following:

	FI-2010				LOB-2021			LOB-2022		
Model	F1 Claim	F1 LOBCAST	F1 Rank	Robustness Score (%)	F1 LOBCAST	F1 Rank	General. Score (%)	F1 LOBCAST	F1 Rank	General. Score (%)
MLP	51.8 ± 3.2	↓ 48.0 ± 2.6	14	91.8	↑ 55.5±3.9	14	95.0	↑ 53.1±2.5	13	96.6
LSTM	63.4 ± 2.1	↓ 63.4 ± 3.6	7	95.5	↓ 56.9±4.1	11	85.9	↓ 56.1±2.8	9	88.8
CNN1	57.9 ± 1.9	↑ 58.1±13.1	10	80.9	↓ 57.5±3.0	8	97.0	↓ 57.1±2.7	6	99.3
CTABL	74.3 ± 5.2	↓ 69.6 ± 4.3	5	91.3	↓ 59.7±2.7	3	78.4	↓ 58.1±3.3	5	78.4
DEEPLOB	78.9 ± 4.4	↓ 71.4 ± 5.3	4	87.6	↓ 59.5±3.0	4	73.7	↓ **59.5 ± 2.9**	**1**	74.7
DAIN	66.8 ± 1.5	↓ 55.6 ± 5.9	11	81.4	↓ 55.9±4.4	12	79.5	↓ 54.1±2.1	12	83.9
CNNLSTM	47.0 ± 0.0	↑ 63.2 ± 8.4	8	75.7	↑ 57.0±3.3	10	87.8	↑ 56.8±2.5	7	90.3
CNN2	45.0 ± 0.8	↑ 50.5±17.3	12	70.5	↑ 55.5±3.5	13	86.6	↑ 55.8±3.2	10	88.6
TRANSLOB	**87.3 ± 4.0**	↓ 59.4 ± 2.6	9	69.9	↓ 57.7±2.9	7	64.2	↓ 50.4±6.1	14	56.4
TLONBoF	53.0 ± 0.0	↓ 49.7±10.5	13	81.5	↑ 57.3±2.9	9	**99.1**	↑ 54.2±3.1	11	**99.9**
BINCTABL	80.1 ± 6.9	↑ **82.6 ± 7.0**	**1**	**99.7**	↓ **61.2 ± 2.7**	**1**	73.5	↓ 59.2±3.3	2	72.3
DEEPLOBATT	78.8 ± 3.1	↓ 67.3 ± 9.0	6	81.2	↓ 60.1±3.0	2	75.7	↓ 58.9±2.8	3	74.5
DLA	78.7 ± 0.7	↓ 73.4±12.1	2	93.2	↓ 57.7±3.7	6	74.9	↓ 56.6±2.4	8	76.9
ATNBoF	67.1 ± 5.5	↓ 40.9 ± 7.7	15	66.1	↓ 53.1±3.7	15	80.9	↓ 48.0±6.9	15	81.2
AXIALLOB	82.0 ± 3.7	↓ 73.4 ± 5.7	3	88.2	↓ 59.5±3.3	5	71.3	↓ 58.6±2.6	4	70.7
METALOB	–	82.2 ± 7.3	–	–	55.9 ± 2.6	–	–	53.2 ± 1.5	–	–
MAJORITY	–	60.0 ± 12.7	–	–	55.5 ± 2.3	–	–	47.9 ± 2.0	–	–

Table 2: Robustness, generalizability, and performance scores of the models. Arrows indicate whether the measured F1 of a system is higher or lower than stated in the original paper. Colour saturation highlights systems with best (green) and worst (red) robustness and generalizability scores.

1. Except for a few systems, there is a considerable difference between the claimed performances and those measured in both robustness and generalizability experiments. Note that while the performance gap is negative on average and considerably negative in the scenario of LOB-2021 and 2022, a few systems outperform the claimed results, as highlighted by the arrows in Table 2.
2. All models are very sensitive to hyperparameters, in fact, they diverged (F1-score $\leqslant 33\%$) during the hyperparameters search for about half of the runs.
3. The ranking of systems changes considerably if we compare the declared performances with those measured in our experiments. On the other hand, the best six systems in FI-2010 remain the same in LOB-2021 and 2022.
4. The best-ranked systems do not consistently hold the lead in terms of robustness and generalizability - except for BINCTABL. On the contrary, some of them obtained poor generalizability scores, suggesting that they overfitted the FI-2010 dataset.
5. Five of the best six models incorporate attention mechanisms. In particular, the best-performing model is BINCTABL, which enhances the original CTABL model by adding an Adaptive Bilinear Normalization layer, enabling joint normalization of the input time series along both temporal and feature dimensions. On average, BINCTABL improves the F1-score by up to 9.2% compared to DLA, i.e., the second-best model, and up to 13% compared to CTABL.
6. Regrettably, ensemble models (the last two rows in Table 2) do not exceed the performance of the top-performing models, which is probably due to the relatively high agreement rate among systems, as shown in Section 6 in SUP.

Robustness on FI-2010 As far as the robustness experiments are concerned, it is important to note that some models discussed in the literature incorporate additional market observation features for predictions. This is the case for models such as DAIN, CNNLSTM, TLOBOF, and DLA. To ensure a fair comparison among the models, we included them in our study but reduced their feature set to only the 40 raw LOB features. Due to the presence of these additional features, a strict robustness study could not be conducted for these models. However, the reduction of features did not necessarily cause a deterioration in performance: of particular interest is the case of CNNLSTM, for which the authors used stationary features derived from the LOB, stating that they were better than the raw ones. Impressively, CNNLSTM achieves the greatest average improvement of 20.9% among all the models, proving that, for this model, the raw LOB features are better suited to forecast the mid-price movement than the features proposed by the original authors. This is also the case for DLA, which originally uses 144 input features. In fact, with the only raw features, DLA exhibited remarkable performance, ranking second best in terms of F1-score.

Based on these experiments (summarized in Table 2), the BINCTABL model demonstrates the **highest F1-score** when averaged over the seeds and prediction horizons, achieving an average of $82.6\% \pm 7.0$. Notably, the BINCTABL model also exhibits the strongest robustness score of 99.7, ranking as the best in terms of robustness. For a more comprehensive analysis, Figure 2 provides

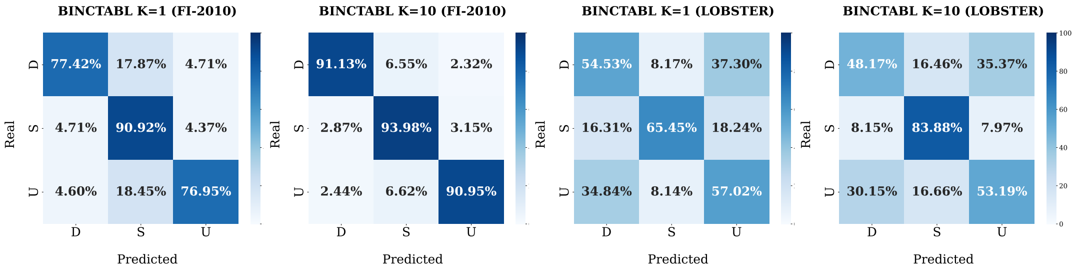

Figure 2: Confusion matrices for BINCTABL $(k = 1, 10)$ on FI-2010 and LOB-2021 datasets.

the confusion matrices of the BINCTABL model's predictions for two horizons ($k = 1$ and $k = 10$). The confusion matrices demonstrate that the model is slightly biased toward the stationary class. This pattern is consistent across all the models, especially for the first three horizons, reflecting the imbalance of the dataset towards the stationary class, as specified in Section 4.1.

Remarkably, a significant number of models in our study failed to achieve the claimed performance levels. Two possible reasons are the lack of the original code and the missing hyperparameters declaration. Among the models, TRANSLOB and ATNBOF exhibit the largest discrepancies, ranking as the second and first worst performers, respectively. Notably, ATNBoF performs the poorest among all models, both in terms of robustness score and F1-score.

We observed that CNN1, CNN2, CNNLSTM, TLONBOF, and DLA are the most sensitive models in terms of network weight initialization and dataset shuffling, in fact, these models exhibit a standard deviation over the runs that exceeds 5 points, indicating a high degree of variability in their performance.

Finally, we highlight that none of the top three models in our study utilize $h = 100$ long market observations as input, despite it being a common practice in the literature [26–28, 32, 36], meaning that they are able to achieve good results without relying on a large historical context. This suggests that the most influential and relevant dynamics impacting their predictions tend to occur within a short time frame. In Section 6 in SUP, we analyze in more detail the robustness results of our benchmark study when varying the horizons.

Generalizability on LOB-2021/2022 When comparing the performance of models on the FI-2010 and LOB-2021/2022 datasets, we observe that models showing high performance on the FI-2010 dataset demonstrate a deterioration in performance. Conversely, some of the models that performed poorly on the FI-2010 dataset show an improvement in performance on the LOB-2021/2022 datasets. However, the overall performance of all models on the LOB-2021/2022 dataset is still significantly lower than on the FI-2010 dataset, ranging 48-61% in F1-score. Furthermore, we conjecture that the overall performance is worse in LOB-2022 than in LOB-2021 due to the higher stocks' volatility. We mention two potential factors contributing to this observed phenomenon. Firstly, the LOB-2021/2022 datasets present a higher level of complexity than the FI-2010 dataset, despite having been generated with a similar approach. Indeed, NASDAQ is a more efficient and liquid market than the Finnish one, as evidenced by the fact that LOB-2021/2022 datasets have approximately three times the size of FI-2010 in terms of events for the same period length. Secondly, the best-performing models may overfit the FI-2010 dataset, leading to a decrease in their performance when applied to LOB-2021/2022 datasets. In particular, BINCTABL experiences an average decrease of approximately 19.6% in F1-Score across all horizons, resulting in a generalizability score of 73.5%. For a more detailed analysis of our generalizability results, we refer to Section 6 in SUP, where we also illustrate the substantial performance variation across different stocks. Among the tested models, CSCO stands out as yielding the highest performance. This may be attributed to the high stationarity of CSCO (balance 18-65-17% in the train set), indicating more stable and predictable behaviour. This hypothesis is supported by the confusion matrices, which consistently show the best performance in the stationary class across all models; for reasons of space, we reported only those of BINCTABL in Figure 2 while for the complete study, we refer the reader to Section 7 in SUP. Finally, as a final benchmark test, we conducted a trading simulation using LOB-2021. The results confirm the challenging nature of the task using the up-to-date LOB-2021 dataset, indicating that the models' profitability is far from guaranteed. For more detailed information about the simulation, please refer to Section 7 in SUP.

5 Discussion and Conclusions

Our findings highlight that price trend predictors based on DNNs using LOB data are not consistently reliable as they often exhibit non-robust and non-generalizable performance. Our experiments demonstrate that the existing models are susceptible to hyperparameter selection, randomization, and experimental context (stocks, volatility). In addition, the selection of datasets and the experimental setup fail to capture the intricacies of the real-world scenario. This lack of generalizability makes them inadequate for practical applications in real-world settings.

Models Our results lead to a crucial observation: on the LOBSTER dataset, SOTA DL models for LOB data exhibit low generalizability. We suggest that this phenomenon is due to two factors: the higher complexity of the LOBSTER dataset compared to the FI-2010 dataset, and the overfitting of the best-performing models to the FI-2010 dataset, which lowers their performance on the LOBSTER dataset. Another key finding of this study is that the top models with the highest performance on both datasets employ attention mechanisms. This suggests that the attention technique enhances the extraction of informative features and the discovery of patterns in LOB data. However, in general, it appears that current models cannot cope with the complexity of financial forecasting with LOB data. Future investigations should consider state-of-the-art approaches to multivariate time series forecasting, such as [49–51], which have not yet been adopted in the financial sector.

Dataset Financial movements can be influenced by geopolitical events, as political actions and decisions can significantly impact economic conditions, market sentiment, and investor confidence [1]. These factors are not captured by LOB data alone. For this reason, we believe that price predictors may benefit from integrating LOB data with additional information, for example, sentiment analysis relying on social media and press data, representing an easily accessible source of exogenous factors impacting the market [52]. This is particularly true for mid- and long-term price trend prediction, whereas it might not hold for HFT strategies [2]. We remark that micro and macroscopic market trends are fundamentally different, and the microscopic behaviour of the market is very much driven by HFT algorithms, making it almost exclusively dependent on financial movements rather than external factors. In this scenario, granular and raw LOBs may suffice to provide data for price trend prediction. Another weakness in dataset generation is the potential for training, validation, and test splits to have dissimilar distributions. This occurs due to the distinct characteristics of the historical periods covered by the stock time series. This can negatively affect the model's ability to generalize effectively and make reliable predictions on unseen data.

Labelling As we discussed in Sections 2, 4.1 and 4.4, the choice of the threshold for class definition in Equation 2 plays a crucial role in determining the trend associated to a market observation. We believe that current solutions present room for improvement. As discussed in Section 4.1, in FI-2010, the parameter θ was chosen to obtain a balanced dataset in the number of classes for the horizon $k = 5$ (which is the mean value of the considered interval in the set $\mathcal{K}$). Thus, θ is not chosen in accordance with its financial implication but rather serves the purpose of balancing the dataset. We recall that the dataset is made of different stocks. With such a labelling system, fixed θ, stocks with low returns become associated with stable trends, as their behaviour is overshadowed by stocks exhibiting higher returns. Good practices that could be investigated are to use a weighted look-behind moving average to absorb data noise instead of mid-prices as in Equation 2 or to define a dynamically adapting θ which accounts for changing trends of a stock's mid-price. Moreover, the labelling approach of Equation 2, used by all surveyed models, fails to leverage important aspects available in LOB data, including the volume, which directly influences stock volatility. Therefore, another possible improvement is the definition and use of other insightful features that can be extrapolated from a LOB in addition to the mid-price. Such values could encapsulate other peculiar and informative features, such as stocks' spread and volumes.

Profit In the context of stock prediction tasks, it is of utmost importance to go beyond standard statistical performance metrics such as accuracy and F1 score and incorporate trading simulations to assess the practical value of algorithms. SPTP predictors' ultimate measure of success lies in their ability to generate profits under real market conditions. It is essential to conduct trading simulations using real simulators that go beyond testing on historical data. Recent progress has been made in the context of reactive simulators [53–56].

We acknowledge that our study is subject to some limitations, which should be considered when interpreting our findings. First, we conducted a grid hyperparameter search for the models which did not specify them. Since hyperparameter search is not exhaustive, our chosen best hyperparameters could potentially undermine the quality of the original systems. Secondly, due to computational resource limitations, we could not train the benchmarked models on LOB datasets spanning longer periods, e.g., years rather than weeks. We recognize that doing so could have improved our generalizability results.

Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

Acknowledgements

This research was funded by JPMorgan Chase AI Research Faculty award "*Understanding interdependent market dynamics: vulnerabilities and opportunities*". We also thank Poste Italiane for funding a Ph.D. scholarship on Financial applications of Artificial Intelligence.

References

[1] R. F. Engle, E. Ghysels, and B. Sohn, "Stock market volatility and macroeconomic fundamentals," *Review of Economics and Statistics*, vol. 95, no. 3, pp. 776–797, 2013.

[2] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press, 2018.

[3] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.

[4] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *Expert Systems with Applications*, vol. 184, p. 115537, 2021.

[5] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied soft computing*, vol. 90, p. 106181, 2020.

[6] D. T. Tran, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Data normalization for bilinear structures in high-frequency financial time-series," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7287–7292.

[7] X.-Y. Liu, Z. Xia, J. Rui, J. Gao, H. Yang, M. Zhu, C. Wang, Z. Wang, and J. Guo, "Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1835–1849, 2022.

[8] I. Zaznov, J. Kunkel, A. Dufour, and A. Badii, "Predicting stock price changes based on the limit order book: a survey," *Mathematics*, vol. 10, no. 8, p. 1234, 2022.

[9] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 7459–7478, 2021.

[10] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[11] M. Baker, "Reproducibility crisis," *Nature*, vol. 533, no. 26, pp. 353–66, 2016.

[12] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods," *Journal of Forecasting*, vol. 37, no. 8, pp. 852–866, 2018.

[13] G. The Efficient Reconstructor at Humboldt Universität zu Berlin, "Lobster: Limit order book system." [Online]. Available: https://lobsterdata.com/

[14] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: a literature review," *Expert Systems with Applications*, p. 116659, 2022.

[15] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," *Applied Soft Computing*, vol. 93, p. 106384, 2020.

[16] Z. Hu, Y. Zhao, and M. Khushi, "A survey of forex and stock price prediction using deep learning," *Applied System Innovation*, vol. 4, no. 1, p. 9, 2021.

[17] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, 2020.

[18] J. Shah, D. Vaidya, and M. Shah, "A comprehensive review on multiple hybrid deep learning approaches for stock prediction," *Intelligent Systems with Applications*, p. 200111, 2022.

[19] A. I. Al-Alawi and Y. A. Alaali, "Stock market prediction using machine learning techniques: Literature review analysis," in *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, 2023, pp. 153–157.

[20] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato, "Machine learning for quantitative finance applications: A survey," *Applied Sciences*, vol. 9, no. 24, p. 5574, 2019.

[21] L. N. Mintarya, J. N. Halim, C. Angie, S. Achmad, and A. Kurniawan, "Machine learning approaches in stock market prediction: A systematic literature review," *Procedia Computer Science*, vol. 216, pp. 96–102, 2023.

[22] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200209, 2021.

[23] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," *International Journal of Financial Studies*, vol. 7, no. 2, p. 26, 2019.

[24] K. Olorunnimbe and H. Viktor, "Deep learning in the stock market—a systematic survey of practice, backtesting, and applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2057–2109, 2023.

[25] "LOBCAST," Link available upon acceptance. The code is included in the supplementary material.

[26] Z. Zhang, S. Zohren, and S. Roberts, "Deeplob: Deep convolutional neural networks for limit order books," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 3001–3012, 2019.

[27] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Using deep learning to detect price change indications in financial markets," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 2511–2515.

[28] ——, "Forecasting stock prices from the limit order book using convolutional neural networks," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, vol. 1. IEEE, 2017, pp. 7–12.

[29] D. T. Tran, A. Iosifidis, J. Kanniainen, and M. Gabbouj, "Temporal attention-augmented bilinear network for financial time-series data analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1407–1418, 2018.

[30] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Deep adaptive input normalization for time series forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3760–3765, 2019.

[31] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Using deep learning for price prediction by exploiting stationary limit order book features," *Applied Soft Computing*, vol. 93, p. 106401, 2020.

[32] J. Wallbridge, "Transformers for limit order books," *arXiv preprint arXiv:2003.00130*, 2020.

[33] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Temporal logistic neural bag-of-features for financial time series forecasting leveraging limit order book data," *Pattern Recognition Letters*, vol. 136, pp. 183–189, 2020.

[34] Z. Zhang and S. Zohren, "Multi-horizon forecasting for limit order books: Novel deep learning approaches and hardware acceleration using intelligent processing units," *arXiv preprint arXiv:2105.10430*, 2021.

[35] Y. Guo and X. Chen, "Forecasting the mid-price movements with high-frequency lob: A dual-stage temporal attention-based deep learning architecture," *Arabian Journal for Science and Engineering*, pp. 1–22, 2022.

[36] D. T. Tran, N. Passalis, A. Tefas, M. Gabbouj, and A. Iosifidis, "Attention-based neural bag-of-features learning for sequence data," *IEEE Access*, vol. 10, pp. 45 542–45 552, 2022.

[37] D. Kisiel and D. Gorse, "Axial-lob: High-frequency trading with axial attention," *arXiv preprint arXiv:2212.01807*, 2022.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, no. 2016. Savannah, GA, USA, 2016, pp. 265–283.

[40] P. Gomber and M. Haferkorn, "High frequency trading," in *Encyclopedia of Information Science and Technology, Third Edition*. IGI Global, 2015, pp. 1–9.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv:2005.14165*, 2020.

[44] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[45] "Backtesting.py." [Online]. Available: https://github.com/kernc/backtesting.py

[46] Y. Wu, M. Mahfouz, D. Magazzeni, and M. Veloso, "Towards robust representation of limit orders books for deep learning models," *arXiv preprint arXiv:2110.05479*, 2022.

[47] L. Lucchese, M. Pakkanen, and A. Veraart, "The short-term predictability of returns in order book markets: a deep learning perspective," *arXiv preprint arXiv:2211.13777*, 2022.

[48] Y. Wu, M. Mahfouz, D. Magazzeni, and M. Veloso, "How robust are limit order book representations under data perturbation?" *arXiv preprint arXiv:2110.04752*, 2021.

[49] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, “Temporal fusion transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[50] A. Drouin, É. Marcotte, and N. Chapados, “Tactis: Transformer-attentional copulas for time series,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 5447–5493.

[51] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[52] R. Ren, D. D. Wu, and T. Liu, “Forecasting stock market movement direction using sentiment analysis and support vector machine,” *IEEE Systems Journal*, vol. 13, no. 1, pp. 760–770, 2018.

[53] A. Coletta, M. Prata, M. Conti, E. Mercanti, N. Bartolini, A. Moulin, S. Vyetrenko, and T. Balch, “Towards realistic market simulations: A generative adversarial networks approach,” in *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF)*, New York, NY, USA, 2022. [Online]. Available: https://doi.org/10.1145/3490354.3494411

[54] A. Coletta, A. Moulin, S. Vyetrenko, and T. Balch, “Learning to simulate realistic limit order book markets from data as a world agent,” in *Proceedings of the Third ACM International Conference on AI in Finance*, 2022, pp. 428–436.

[55] T. Mizuta, “A brief review of recent artificial market simulation (agent-based model) studies for financial market regulations and/or rules,” *Available at SSRN 2710495*, 2016.

[56] Z. Shi and J. Cartlidge, “Neural stochastic agent-based limit order book simulation: A hybrid methodology,” *arXiv preprint arXiv:2303.00080*, 2023.