Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Fx-spot predictions with state-of-the-art transformer and time embeddings

Tizian Fischer [a,*], Marius Sterling [b], Stefan Lessmann [a,c]

[a] Chair of Information Systems, School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[b] Distributed Artificial Intelligence Laboratory, Technical University of Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany
[c] Bucharest University of Economic Studies - ASE, Piata Romana, No. 8, Bucharest, Romania

ARTICLE INFO

Keyword:
Transformer Financial forecasting Forex Deep Learning

ABSTRACT

The transformer architecture with its attention mechanism is the state-of-the-art deep learning method for sequence learning tasks and has achieved superior results in many areas such as NLP. Utilizing the transformer architecture for the prediction of sequential time series such as financial time series has hardly been investigated in previous studies. In this research paper, the transformer architecture with time embeddings is used in foreign exchange (FX) trading, the world's largest financial market, and tests its suitability. A systematic comparison is made between transformer and benchmark models. It also examined which influence multivariate, cross-sectional input data have on the forecasting performance of the various models. The goal of the paper is to contribute to the empirical literature on FX forecasting by introducing a transformer with time embeddings to the forecasting community and assessing the accuracy of corresponding models by forecasting exchange rate movements. Empirical results indicate the suitability of transformer models for FX-Spot forecasting in general but also evidence the need for transformer models for multivariate, cross-sectional input data to outperform other state-of-the-art neural networks such as LSTM.

1. Introduction

The largest financial market with a daily trading volume of USD 6.6 trillion is the foreign exchange (FX) market (Aslam et al., 2020). Due to its strong links with goods, labor, and financial markets, the FX market has a major impact on the real economy and is therefore of great importance for the economy and society (Davidson, 2003). This market has some peculiarities compared to stock and bond markets. Transactions do not take place via a regulated exchange but over the counter (OTC) and most of them via electronic trading platforms such as EBS. Independence from an exchange result in continuous market opening hours is one of the drivers of higher liquidity and tighter spreads (Frankel & Froot, 1990). Another special feature is the participating players and their market share. Most of the volume traded takes place in the interbank market. In contrast to stocks and bonds, private individuals contribute very little to the traded volume of currencies and most of the volume traded from them comes from indirect FX transactions such as payments. Another important group of players is central banks, which not only exert an influence through monetary policy measures such as key interest rates but also through direct interventions in the FX market (Frankel & Froot, 1990).

Given the enormous trading volume, potential profits, and effects on the complementary goods and labor markets, it is not surprising that the forecasting of financial markets receives a lot of attention from academic researchers and practitioners (Bradshaw, 2011). As a basis for the prediction of asset prices, the question of the efficiency of financial markets arises. The much-recognized work of Fama (Fama, 1970) serves as a milestone in this regard and distinguishes between three degrees of market efficiency. A multitude of studies (with varying results) have examined the efficiency of different areas of the financial markets. The efficiency of large parts of the financial markets corresponds most closely to the semi-strong form (Stasiulis, 2009).

FX rates are enormously correlated to the macroeconomic features of countries. While company-specific factors have a major influence on other asset classes such as stocks or corporate bonds, for FX it is only macroeconomic factors such as inflation, unemployment, monetary and fiscal policy, and the balance of payment that have a significant influence on prices (Mussa, 2019). Thus, valuation and forecasting models used for other asset classes cannot or only hardly be adapted to FX rates. The absence of a fair market valuation model results in three classic FX strategies: value, carry, and momentum (Gyntelberg & Schrimpf, 2011). Value can be best compared with the attempt to determine an intrinsic

[*] Corresponding author at: Chair of Information Systems, Humboldt University of Berlin, Spandauer Str. 1, 10178 Berlin.
E-mail addresses: tizian.fischer@hu-berlin.de (T. Fischer), sterling@tu-berlin.de (M. Sterling), stefan.lessmann@hu-berlin.de (S. Lessmann).

https://doi.org/10.1016/j.eswa.2024.123538
Received 21 August 2023; Received in revised form 2 January 2024; Accepted 18 February 2024
Available online 13 March 2024

value and is based on the purchasing power parity (PPP) theory. PPP states that the exchange rate must always be such that an identical good (absolute PPP) or an identical basket of goods (relative PPP) costs the same in each currency. If this is not the case, arbitrage can be achieved, which leads to the exchange rate tending towards the PPP exchange rate. While this is true in the long term, the exchange rate can deviate sharply from the fair PPP rate in the short and medium term (Rogoff, 1996). Carry is based on the fact that the uncovered interest rate parity (UIP) theory is not bound by arbitrage. UIP says that the change in the exchange rate over a period of time should correspond exactly to the interest rate difference between the two currencies and thus the forward rate is an unbiased predictor of the future spot rate (Bekaert, Wei & Xing, 2007). Empirical studies show that this is not the case and that the carry strategy can generate positive returns.

(Mikhaylov, Sokolinskaya & Nyangarika, 2018). The momentum strategy is based on technical analysis and empirical studies which show that currency pairs which have achieved a positive return in the previous time step have a higher probability of achieving a positive return in the next time step (Barroso & Santa-Clara, 2015).

These fundamental, macroeconomic drivers are published at relatively coarse intervals (e.g., weekly, monthly, or quarterly) while FX rates are traded continuously (Almeida, Goodhart & Payne, 1998). Empirical studies show that with new information it is not the absolute but the relative deviation from the analyst consensus that explains the subsequent price fluctuations (Abarbanell, 1991). A lot of information about the forecast development is therefore already priced into asset prices. This is one reason for the predictive power of other asset classes for explaining and forecasting FX rates. Furthermore, prices of other assets such as commodities or government bonds also have a direct influence on FX rates by influencing flows on the goods and financial markets in different currencies (Mussa, 2019).

The last peculiarity of FX forecasting that is highlighted in this paper is the stronger focus on intraday developments. This is due to two characteristics of the FX market that were already mentioned. First, the market is open 24 h a day resulting in more intraday flow and activity. Second, the investment focus in other asset classes, such as equity and bonds, is more often on medium to long-term asset allocation, while in the FX area, it is mostly about hedging short-term currency risks (Bartram, 2008).

For a long time, the tools used for predictions stemmed mostly from the fields of financial mathematics, statistics, and probability calculations (Hastie et al., 2009). Thanks to great advances in computing power, the availability of more and higher quality data and new machine learning (ML) architectures, techniques, and theories, ML methods have achieved groundbreaking results in many areas (Alpaydin, 2021). Many papers have shown that ML methods can model non-linear and non-stationary returns of financial time series adaptively (Rundo et al., 2019).

The focus of ML literature on deep learning (DL) methods can be described as a recent development in the field of financial time series forecasting. DL is suitable for various ML tasks and offers the advantage that the model learns to recognize important features and ignore irrelevant features independently (LeCun, Bengio & Hinton, 2015). In DL, specific architectures have proven to be suitable for certain tasks, e.g., convolutional neural networks (CNNs) are particularly suitable for image recognition, while recurrent neural networks (RNNs) are particularly suitable for sequential data such as natural language processing (NLP) and time series (Bishop, 1994). While classic RNNs have been used for forecasting financial markets, more complex architectures such as long short-term memory (LSTM) and gated recurrent units (GRU) have shown great successes in recent years and are considered state-of-the-art with respect to the classification or regression tasks of financial time series (Irie et al., 2016). A breakthrough in NLP translations was published by Vaswani et al. (2017) when the transformer architecture was introduced. This can be viewed as a further development of LSTM, in which the attention mechanism becomes the main part of the architecture and can thus recognize complex patterns in the data even better. The transformer architecture is well established for specific ML tasks and notable models were developed, e.g., BERT (Devlin et al., 2019), and GPT-2/3/4 (OpenAI, 2023a) in NLP; Unispeech (Wang et al., 2021), and Wav2Vec2 (Baevski et al., 2020) in speech synthesis; Image GPT (Chen et al., 2020), and Vision Transformer (Dosovitskiy et al., 2020) in computer vision; ChatGPT (OpenAI, 2023b), and CLIP (Radford et al., 2021) for multi-task and multimodal learning tasks. Besides these established use cases, transformers are used in time series prediction and classification, e.g., for financial time series, see Table 1, and in a variety of fields, for example, for protein structure prediction (Jumper et al., 2021), biomedical information retrieval (Hall et al., 2023) or molecular property prediction (Irwin et al., 2022). The new DL architecture of the transformer has been - to our knowledge - not yet been researched for predictions in the FX area, although it seems promising.

The utilization of advanced ML and DL techniques, such as the transformer, allows for the recognition of patterns in multivariate data. Previous studies, however, have primarily employed univariate or low-dimensional data as inputs for their models, including stock data (Fischer & Krauss, 2018) and FX data (Ni et al., 2019; Dautel et al., 2020). The majority of studies analyzed daily data, specifically stock returns, with limited use of intraday data (Si, Li, Ding, & Rao, 2017; Lachiheb & Gouider, 2018). Only a few studies have focused on FX data (Ni et al., 2019; Dautel et al., 2020) and evaluated model performance through trading strategies (Shen, Tan, Zhang, Zeng, & Xu, 2018; Liu et al., 2019). There is a scarcity of research utilizing multivariate intraday FX data, with only a few studies employing RNN and evaluating FX prediction results directly (Qi, Khushi, & Poon, 2020; Zeng & Khushi, 2020), with no evaluation of trading strategies. In summary, there is a limited body of research in this area.

This paper primarily focuses on a newer DL architecture - the transformer model - and explores its potential in an FX rate forecasting task. Three contributions to the existing literature are made in this research work.

The first goal of the paper is to contribute to the empirical literature on FX forecasting by introducing and assessing the performance of a transformer with time embeddings. In particular, the paper reports original results from a comparative analysis of transformer based neural networks versus LSTM and ARIMA. Considering three FX spot rates, the different models are assessed regarding financial and statistical metrics.

Second, this paper aims to contribute by systematically comparing the prediction performance between univariate and multivariate input data for the transformer and the selected benchmark models. This aims to show whether the hypothesis is confirmed that multivariate technical and fundamental features improve forecast performance and whether this is particularly true for the transformer.

Third, the focus on intraday predictions from a 10-minute observation period was identified as a research gap. The training of neural network-based methods requires many data points to prevent overfitting and to facilitate generalization. As a result, a reasonable amount of financial time series data can often only be collected for intraday data (Arnott et al., 2019). Additionally, market makers' pricing, hedging and risk management necessitate high-frequency forecasts of FX rates.

The remainder of this paper is organized as follows; section 2 briefly covers the related work with the identified research gaps in the existing literature. Section 3 introduces the transformer architecture with time embeddings. Section 4 provides an in-depth discussion of the methodology in this analysis, including the data sample and software packages. Section 5 presents the empirical results and discusses the most relevant findings. Finally, section 6 concludes the work.

2. Background and related work

There is a large amount of literature regarding forecasting financial time series. In addition, an increasing number of recent financial forecast studies employ DL methods to improve the field's state-of-the-art.

Table 1
Prior work on predictions of financial markets with DL methods. *.

						Time -Step			Asset Class			Input			
Reference	Year	Model	Performance Criteria	Trading Strategy	Period	<1d	1d	>1d	FX	Stock /Index	Other	Type	Numbers	Text	Multivariate
Yümlü, Gürgen and Okay (2005)	2005	LSTM	MAPE		2012–2017		X			X		R	X		
Chen, He and Tso (2017)	2017	LSTM	MSE		2007–2017		X				X	R	X		
Di Persio and Honchar (2017)	2017	RNN,LSTM,GRU	Accuracy		2012–2016		X			X		C	X		
Si, Li, Ding, & Rao (2017)	2017	LSTM	Profit,SR	X	2016–2017	X				X		R			X
Hansson (2017)	2017	LSTM	MSE,Accuracy		2009–2017		X			X		R	X		
Elliot and Hsu (2017)	2017	LSTM,RNN	MAE,RMSE,R2		2000–2017		X			X		R	X		
Liu, Zhang and Ma (2017)	2017	CNN,LSTM	Annualized Return,Mxm Retracement	X	2007–2017		X			X		R	X		
Yong, Abdul Rahim and Abdullah (2017)	2017	DNN	RMSE,MAPE,Profit,SR	X	2010–2017		X			X		R	X		
Lee and Soo (2017)	2017	CNN + LSTM	RMSE,Profit	X	2001–2017		X			X		R		X	
Widegren (2017)	2017	DNN,RNN	Accuracy,p -value		1993–2017		X			X		R	X		X
Ausmees, Milovanovic, Wrede and Zafari (2017)	2017	DBN	MAE		2000–2017		X			X		R	X		
dos Santos Pinheiro and Dras (2017)	2017	LSTM	Accuracy		2006–2013		X			X		C		X	
Fischer and Krauss (2018)	2018	LSTM	RMSE,R2,Adj.R2		1989–2015		X			X		R	X		
Hollis et al. (2018)	2018	LSTM + Attention	Accuracy		2007–2017		X			X		C	X	X	X
Siami-Namini and Namin (2018)	2018	LSTM	RMSE		1985–2018		X			X		R	X		
Baek and Kim (2018)	2018	LSTM	ME,MAPE,MAE		2000–2017		X			X		R	X		
Althelaya, El-Alfy and Mohammed (2018)	2018	LSTM	MAE,RMSE,R2		2010–2017		X			X		R	X		
Zhao, Rao, Tu and Shi (2017)	2018	LSTM	Accuracy		2008–2017		X			X		C	X		
Chen et al. (2018)	2018	RNN	Accuracy,MAE,MAPE,RMSE		2015–2017		X			X		R	X	X	X
Chen, Wu and Bu (2018)	2018	LSTM + Attention	MSE,MAE		2004–2018		X			X		R	X		X
Das, Behera and Rath (2018)	2018	RNN	Correlation		2016–2017		X			X		R	X	X	X
Wang, Sun, Liu, Cao and Wang (2018)	2018	CNN	Accuracy,F1,Return,Sharpe Ratio	X	2010–2017		X			X		C	X		X
Iwasaki, Chen, Du and Tu (2018)	2018	LSTM,CNN	Accuracy,R2		2016–2018		X			X		R		X	
Shen, Tan, Zhang, Zeng and Xu (2018)	2018	GRU	Daily return	X	1991–2017		X		X			C	X		
Feng, Polson and Xu (2018)	2018	Fama -French n -factor model DL	R2,RMSE		1975–2017		X			X		R	X		X
Lachiheb and Gouider (2018)	2018	DNN	Accuracy,MSE		2013–2017	X				X		R	X		
Yuan, Zhang and Shao (2018)	2018	DWNN	MSE		2000–2017		X			X		R	X		
Ni et al. (2019)	2019	RNN	MSE		2008–2018		X		X			R	X		
Nikou et al. (2019)	2019	LSTM	MAE,MSE,RMSE		2015–2018		X			X		R	X		
Chen et al. (2019)	2019	LSTM	Accuracy		2016–2017		X			X		C		X	
Li et al. (2019)	2019	RNN	Accuracy,Precision,Recall,F1		2015–2017		X			X		C	X		X
Achkar et al. (2018)	2019	LSTM	MSE		2012		X			X		R	X		
Naik and Mohan (2019)	2019	LSTM	MAE,RMSE		2009–2018		X			X		R	X		
Lai et al. (2019)	2019	LSTM	MAPE,RMSE,MSE		2009–2018		X			X		R	X		
Zhou (2019)	2019	LSTM + MLP	Monthly return,SR	X	1993–2017		X			X		R	X		X
Zhou, Han, Xu, Jiang and Zhang (2019)	2019	LSTM	MSE,MAPE		2006–2017		X				X	R	X		
Xu and Keselj (2019)	2019	LSTM	Accuracy,Matthews Correlation Coefficient		2017–2018		X			X		C	X		X
Nguyen et al. (2019)	2019	LSTM	MAE,MSE,RMSE,MAPE		2009–2019		X			X		R	X		
Lakshminarayanan and McCrae (2019)	2019	LSTM	MAE,MSE,RMSE,MAPE		2014–2018		X			X		R	X		
Rana et al. (2020)	2019	LSTM	RMSE		2008–2018		X			X		R	X		
Fazeli and Houghten (2019)	2019	LSTM	MSE,ROI	X	2014–2019		X			X		R	X		

(*continued on next page*)

Table 1 (*continued*)

Reference	Year	Model	Performance Criteria	Trading Strategy	Period	Time -Step			Asset Class			Input			
						<1d	1d	>1d	FX	Stock /Index	Other	Type	Numbers	Text	Multivariate
Jeong and Kim (2019)	2019	DNN	Total Profit,Correlation	X	1987–2017		X			X		R	X		
Liu et al. (2019)	2019	Transformer	Accuracy,Return	X	2017–2017		X			X		C		X	
Dautel et al. (2020)	2019	LSTM	Accuracy,AUC,Returns,SD,SR	X	1971–2017		X		X			C	X		
Zeng and Khushi (2020)	2020	RNN	RMSE		2019	X			X			R	X		X
Qi et al. (2020)	2020	LSTM	RME,RMSE,MAE,MAPE		2005–2020	X			X			R	X		X
Wu et al. (2020)	2020	Transformer	Correlation,RMSE		2010–2018			X			X	R	X		X
Pang et al. (2020)	2020	LSTM	MSE,MDA		2006–2016		X			X		R	X		X
Feng et al. (2019)	2020	LSTM	MSE,Mean Reciprocal Rank, Investment Return ratio	X	2013–2017		X			X		R	X		
Jin et al. (2020)	2020	LSTM	MAE,RMSE,MAPE,R2		2013–2018		X			X		R		X	X
Long et al. (2020)	2020	LSTM	Accuracy,AUC		2012–2018		X			X		C	X		X
Shen and Shafiq (2020)	2020	DNN	Accuracy		2018–2020		X			X		R	X		X
Nabipour et al. (2020)	2020	LSTM	Accuracy,F1		2009–2019		X			X		C			
Nti et al. (2020)	2020	LSTM	Accuracy		2017–2020		X			X		C	X	X	X
Mehta, Pandya and Kotecha (2021)	2021	LSTM	Accuracy		2014–2020		X			X		C	X	X	X

*Legend: Time-Steps'<1d' represent intraday data, here mostly in the 1–15 min step,'1d' daily data, and'>1d' more distant data, e.g., weekly data. The Type column represents the ML task, e.g.,' R' for regression and' C' for classification.

Surprisingly, the largest market in terms of volume, FX, receives relatively little attention compared to stocks, commodities, derivatives, and cryptocurrencies. Therefore, our literature analysis includes related papers that also use other asset classes. The basis of the analysis is the literature summary in Table 1.

The selected columns are described in detail below and why they were selected for the literature analysis. The Model column describes which category the core model of the paper can be assigned to and thus shows which models are used in state-of-the-art research. Performance Criteria capture which metrics are used to evaluate forecasting performance. This was chosen to align with best practices in performance evaluation. The Environment column was chosen to get a better understanding of the technical implementation. The entry in the trading strategy classifies whether a trading strategy is used and whether, apart from the statistical performance, it is also based on an economic perspective which is particularly related to the research in this paper. Finally, further columns describe the data used, such as the time period, time step, and type of data. The systematic literature analysis aims to identify similarities and differences between the latest studies focusing on the subject. For this purpose, Table 1 shows various model properties such as input, output and general model characteristics for multiple related papers to illuminate the research gap in this section.

Table 1 shows that RNNs, especially LSTM models, were used most frequently for the forecasting tasks in the latest papers. While suitable for the prediction stocks, interpreting the results of LSTM often requires supporting analytics and visualization (Chang et al., 2020). One aspect are temporal dynamics that are hard to capture and understand for RNN, though can be captured by adapted transformer models, e.g., temporal fusion transformers (Lim et al., 2021). Insensitivity to data scale, prediction fragmentation and requiring large amount of input data for training are issues of transformer models. Liu et al. (2022) introduced transformers with memory cells to capture the temporal dynamics and dependencies of multivariate time series, e.g., to obtain longer prediction horizons with shorter input length. Other noteworthy DL methods for predicting trends in the financial market includes reinforcement learning, especially for multistep trend prediction (Li et al., 2020), or self-supervised learning methods (Sun et al., 2022).

The performance metrics utilized for evaluation in the examined papers differ depending on the task. Accuracy was often used for classification tasks, and RMSE, MSE, and MAE for regression tasks. Regression was more often chosen as a task. A trading strategy to evaluate the financial performance of a model was rarely performed. Almost all papers examined have an intraday focus, most often on daily returns of closing prices. The most frequently examined asset class is equity, which includes stocks and stock indices.

The most common input is numerical data, which are mostly complete and univariate, in some cases low-dimensional. Some papers focused on the use of textual input data, e.g., the semantic positivity of analyst reports (Iwasaki, Chen, Du, & Tu, 2018) or social media posts (Liu et al., 2019). So far, multivariate data from different asset classes has rarely been considered model input.

The table thus visually shows the research gap described above and the contribution of this research: there is little literature dealing with the use of the new transformer architecture for the prediction of financial time series, the data used are mostly univariate, and there is no systematic comparison between uni- & multivariate input data and the focus is almost exclusively on intraday predictions. In addition, a consistent trading model was rarely used, thus evaluating the effective economic performance of the model.

In the pursuit of a focused and impactful study, our literature review intentionally concentrates on the underexplored application of the transformer architecture in predicting financial time series. While acknowledging the vast landscape of financial forecasting literature, we have chosen to emphasize the distinctiveness of our research, contributing to a targeted exploration of a niche that has garnered limited attention. This strategic focus allows us to delve deeply into a specific research gap, offering a unique perspective that adds value to the current state of the field.

3. Transformer with time embeddings

The transformer architecture is a DL method and was first presented by (Vaswani et al., 2017). It is based on a classic encoder-decoder structure which has been expanded to include various innovative components. In comparison to other sequential data handling DL methods like RNN and LSTM, the novelty of transformer architecture was the introduction of attention without recursion, thus the ability to focus on certain parts of the input with high resolution while the rest is viewed in low resolution. This also enables bidirectional dependency modeling..

Only the encoder part is relevant for supervised learning problems as forecasting in this paper. Compared to RNN, the transformer model was proven superior in sequence-to-sequence learning tasks and much more parallelizable. Due to that, this architecture can be trained with higher efficiency, generally enabling faster training of large data sets and more timely execution. Two central components of the transformer architecture - the position encoding and self-attention mechanism - are explained and discussed in more detail in this section. The decoder is ignored.

A transformer block is a parameterized function $f_\theta : \mathbb{R}^{n\times d} \rightarrow \mathbb{R}^{n\times d}$ with input data $x^{'} \in \mathbb{R}^{n\times d}$ which has n lagged time steps and d features which contain the concatenated embeddings, which is explained in detail in the following sections. A transformer encoder is a composition of multiple transformer blocks, e.g., $f_{\theta_L}{}^{\circ}\cdots{}^{\circ}f_{\theta_1}(x^{'}) \in \mathbb{R}^{n\times d}$ with L transformer blocks and the set of parameters and weights $\theta_i, i = 1, \cdots, L$ for each transformer block.

4. Time embeddings

Without positional encoding, attention mechanisms have no concept of order and only give an indication of the presence of specific features in the input domain. Nevertheless, it is required to encode the notion of time into the model, since time is an essential feature of the time series representing the data sequence. Otherwise, an FX spot rate from several years ago would have the same influence on the predictions as one a few minutes ago. The positional encoding provides an encoded input sequence to the model by representing an input value and its position in the input sequence.

Time2Vec, a technique inspired by Kazemi et al. (2019) and akin to the Word2Vec embedding function in NLP (Mikolov et al., 2013), is employed for embedding time series data. The vector representation of time generated by Time2Vec captures periodic patterns, making it comparable to other embedding layers. In our transformer architecture, Time2Vec complements the original positional encoding, specifically addressing temporal aspects. This combined approach enhances the model's ability to capture intricate temporal dependencies and is robust against time rescaling, effectively recognizing periodic and non-periodic patterns. The mathematical representation of Time2Vec $t2v(\tau)$ of time τ is as follows:

$$t2v(\tau)[i] = \begin{cases} \omega_i\tau + \varphi_i, if i = 0, \\ F(\omega_i\tau + \varphi_i), if 1 \leq i \leq k. \end{cases} \tag{1}$$

Where $t2v(\tau)[i]$ is the i-th element of $t2v(\tau)$. The time vector representation $t2v$ has two components, $\omega_i\tau + \varphi_i$ represents the linear and non-periodic part and $\mathscr{F}(\omega_i\tau + \varphi_i)$ the periodic feature of the time vector, with the periodic activation function $\mathscr{F}$, e.g., a sinus function. ω_i s and φ_i s are learnable parameters. The input of the transformer encoder $x^{'}$ is than the concatenated input data and time embedding produced by $t2v$.

To provide a more comprehensive understanding of the model's processes, we acknowledge the importance of elucidating the objective function and the collaborative dynamics among its components. The objective function serves as the guiding principle for the model's learning process, and detailing its formulation enhances transparency and aids in the interpretation of our proposed approach.

Furthermore, to better illustrate the synergistic work among the components, we will delve into the interplay between the attention mechanisms, time embeddings, and the overall transformer architecture. This will shed light on how these elements collectively contribute to the model's predictive capabilities in the context of intraday FX-spot predictions.

5. Transformer architecture

A transformer is a neural network architecture that uses a self-attention mechanism that allows the model to focus on the perceived relevant parts of the time series in order to perform predictions, Fig. 1 shows transformer block layer architecture. The self-attention mechanism consists of an attention layer on one head and attention on several heads. Furthermore, the self-attention mechanism can link the sequence of the time series with one another simultaneously, creating long-term dependency understandings. Ultimately, all these processes are parallelized within the transformer architecture, enabling the learning process to be accelerated, which is beneficial for predicting intraday FX rates. As discussed before, the FX market is enormous and major FX pairs are traded in high frequency, thus enabling DL methods to predict high-frequency intraday FX rates which is necessary for pricing, hedging, and risk management for market makers.

5.1. Single-Head attention

Attention is the key part of the model for sequential data with variable length. This is achieved by having an all-to-all comparison, called sequence-to-sequence. For every output of a layer, every possible input from the previous layer is considered $V^{(h)}(x^{'}) = W^T_{V,h}x^{'}$ is a weighted sum of every input with the weighting as the learned function, for one attention head $h = 1, \cdots, H$. Relevance scores are calculated using query $Q^{(h)}(x^{'}) = W^T_{Q,h}x^{'}$ and key vectors $K^{(h)}(x^{'}) = W^T_{K,h}x^{'}$, with projection matrices $W_{V,h}, W_{Q,h}, W_{K,h} \in \mathbb{R}^{d\times d_{TB}}$. For every output position a query and for every input a key with the relevance score as the dot product of the query and key is generated. Therefore, the proposed attention is scaled dot-product attention based on the representation of the input data,

$$Attention\left(Q^{(h)}, K^{(h)}, V^{(h)}\right) = softmax\left(\frac{Q^{(h)}(x^{'})\cdot K^{(h)}(x^{'})^T}{\sqrt{d_k}}\right)\cdot V^{(h)}(x^{'}). \tag{2}$$

5.2. Multi-Head attention

Another innovation is multi-headed attention. The described attention mechanism is used several times. This enables the network to pay attention to different aspects of the input data. A striking example for this is utilizing one attention mechanism for the focus on different aspects of the input data, e.g., momentum, reversal patterns, support and resistance levels. The functionality of a multi-headed attention level is to concatenate the attention weights of n attention levels to a head and then apply a non-linear transformation to compute densities. The output of h single head layers, that.

use projected weight matrices $W_{V,h}, W_{Q,h}, W_{K,h}$ for $h = 1, \cdots, H$ enables the coding of several independent layers of the single head transformation in the model.

$$MultiHead = Concat(Head_1, \cdots, Head_H)W^O \tag{3}$$

Where W^O is the weight matrix to produce final output of the encoder and $Concat$ concatenates all input of the individual heads

$$Head_h = Attention\left(Q^{(h)}, K^{(h)}, V^{(h)}\right), \text{for } h = 1, \cdots, H. \tag{4}$$

The model can therefore concentrate on several time series steps

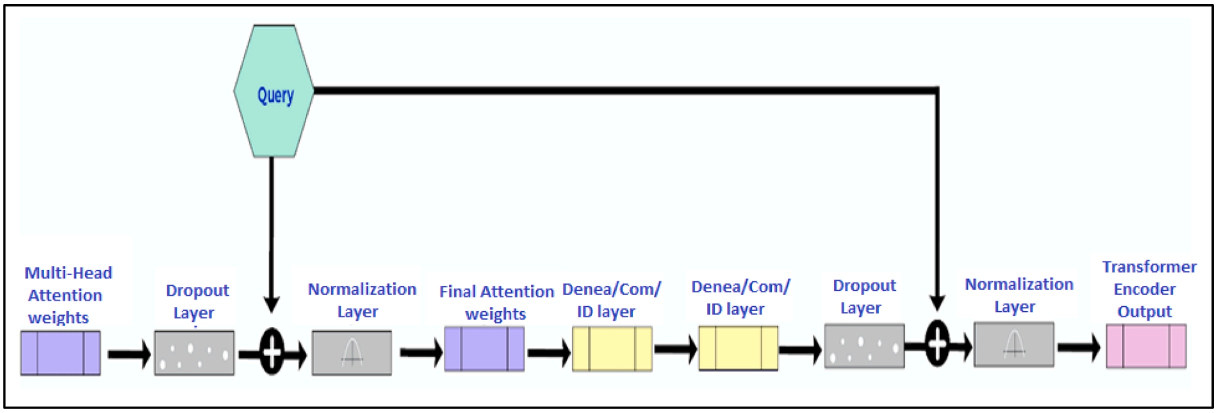

Fig. 1. Transformer Block Layer Architecture.

simultaneously since multi-headed attention allows to comprise different representation subspaces at different previous times and features.

The *MultiHead* attention is regularized with layer normalization and dropout to avoid common deep learning issues and to improve training stability, convergence, and generalization of the trained model. The output of the model is attained by aggregating the output of the transformer encoder using global average pooling and fully connected layers, see Fig. 1.

6. Experimental design

The experimental setup is described in detail below. The code can be found on Github, see github.com/tizianfischer/paper1_fxpred.

6.1. Data

The data used was obtained from Bloomberg and can be broken down into technical and fundamental features. The 433 fundamental features consist of FX (spot rates, volatilities, risk reversals, and swap points for different periods of time, carry returns) and other asset classes (commodities, equity indices, fixed income products). For the other asset classes, such as commodities, equity, and bonds, the futures markets were considered, as these have longer market opening times than, for example, the regular stock exchanges and thus provide complete data. The FX data was obtained for the three most traded FX pairs. The 840 technical features consist of the most essential technical indicators such as Bollinger Bands, Relative Strength Index, Directional Movement Index, etc., which were collected for the FX pairs with the largest trading volume. Thus, the complete data set consists of 1273 technical and fundamental features with a periodicity of ten minutes from November 1st, 2020 to January 31st, 2022. A complete list of the features can be found in the appendix. The target is the closing bid returns of the following currency pairs: EURUSD, USDJPY and GBPUSD.

6.2. Data preprocessing

In the first step, the data was checked for correctness and completeness. Empty values were filled with the last known value to have a value for each feature’s time point. In the next step, the prices were converted into returns (percent changes), increasing the time series’ stationarity. In the last step, the data was normalized to have a mean of zero and a standard deviation of one. Finally, the data was split chronologically into 80 % training, 10 % validation, and 10 % test data, see Fig. 2.

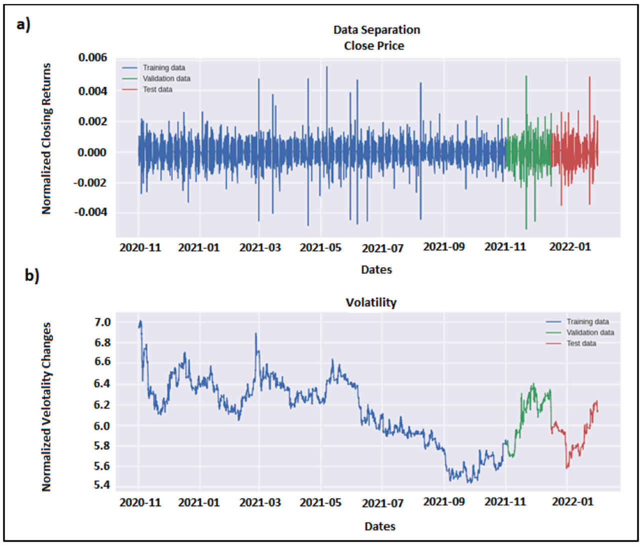

Fig. 2. EURUSD (Top) normalized closing returns and (bottom) volatility for the training, validation and test data.

6.3. Transformer model

During a single training step, the transformer model receives 32 sequences (batch size) of length $n = 128$ with frequency of 10 min and have 1273 features as input, see Fig. 3. Concatenated with the positional and.

time embeddings the input dimension of the transformer encoder is $d = 1275$. The dimensions of K, Q and V are $d_{TB} = 256$ each and the number of heads is $H = 12$. The training duration for each model was defined as 100 epochs with early stopping. A low learning rate of 0.0001 was deliberately defined in the training of the transformer to avoid getting stuck in a local minimum. The output activation function was the identity and the mean squared error (MSE) was minimized as the loss function by the Adam optimizer.

6.4. Benchmark models

In order to validate the performance of the proposed model, it is compared with financial and DL time series forecasting models, e.g., ARIMA and RNN. The ARIMA model parameters are selected by optimization of the EUR/USD exchange rate with regard to the AIC. This resulted in an ARIMA($p = 1, q = 1, d = 0$) model for the training data.

For the DL benchmark model, an LSTM-based architecture (Hochreiter, & Schmidhuber, 1997) was used with five stacked LSTM

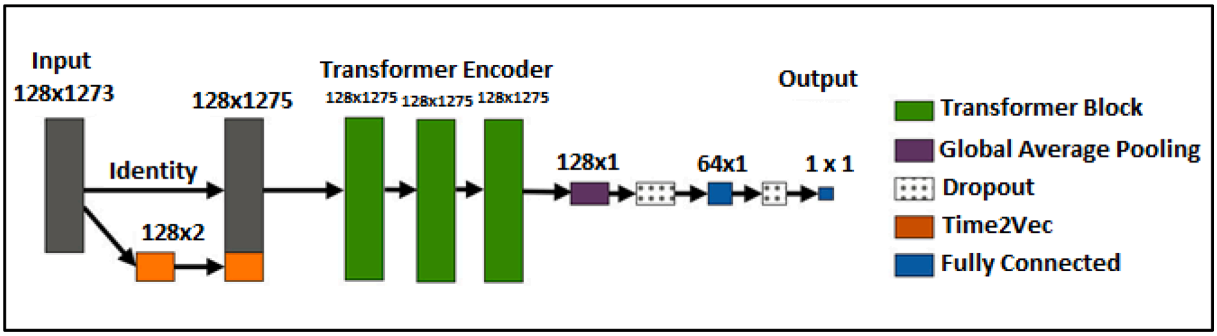

Fig. 3. Architecture of the used transformer model with output dimensions of the layers.

layers, each normalized with Batch Normalization. The LSTM layers used the tanh function as an activation function, with a dropout out of 25 % for univariate and 50 % for multivariate input data models. The number of neurons per LSTM layer was 128 for the univariate and 256 for multivariate input data with 128 for the fifth LSTM layer. The output of the stacked LSTM layers was flattened and downsized to 128 neurons fully connected layer before aggregating the 1-dimensional output.

Since the training of the LSTM models was much more stable a stepwise decaying learning rate is shown in Table 2..

6.5. Performance measurement

An advanced trading model was used to determine economic performance. The predicted return r for the next time interval is used as input for the trading model. The logic of the trading model is as follows:

It starts with a starting capital of 100 monetary units.

The following strategy assignment was used for the predicted return r for each individual time step:

$$Strategy(r) = \begin{cases} Buy, \ if \ r > 0.000001\% \\ Sell, \ if \ r < -0.000001\% \\ Hold, \ else \end{cases} \tag{5}$$

When a Buy prediction is made, all available amount is invested (if not already invested). If Hold is predicted, the previous strategy is kept. If Sell is predicted, then all open positions will be closed. In summary, it can be stated that an investment period starts with a buy prediction and ends with the next sell prediction.

The absolute return, the standard deviation, and the risk-adjusted return (Sharpe ratio) were calculated for the trading strategy and all values were annualized. In order to be able to understand better and explain the differences in the economic metrics, the number of trades was also measured.

6.6. Independent variables

The experimental factors in the experiment are the model, the type of task, and the data used.

Different models (Transformer, LSTM & ARIMA) were used to investigate whether a more complex model can achieve better prediction performance. The transformer will achieve the best prediction performance for the following reasons:

Theoretical derivation: The transformer has the superior capability to handle long-range context dependencies due to its three key characteristics explained in Chapter 3: non-sequential, self-attention, and time embeddings.

Table 2
Stepwise decaying learning rate.

Epoch	0, …, 9	10, …., 19	20, …, 49	50, …, 100
Learning rate	0.001	0.0005	0.0001	0.00001

Table 3
Model performance evaluation of the trading strategy for (a) EURUSD, (b) USDJPY, and (c) GBPUSD.

EUR USD	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	540	9.2	0.036	2.4
Transformer	Classification	Multi	1‘199	7.1	0.066	1.0
LSTM	Regression	Multi	7	−4.8	0.067	−0.8
LSTM	Classification	Multi	20	0.5	0.008	0.3
Transformer	Regression	UNI	1‘095	4.5	0.052	0.8
Transformer	Classification	UNI	627	3.7	0.064	0.5
LSTM	Regression	UNI	1	−0.7	0.067	−0.9
LSTM	Classification	UNI	215	−9.5	0.050	−1.9
ARIMA (1,0,1)	Regression	UNI	1‘128	17.1	0.046	3.6
USD JPY	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	574	22.9	0.041	5.5
Transformer	Classification	Multi	48	12.2	0.029	4.1
LSTM	Regression	Multi	119	−0.3	0.028	−0.2
LSTM	Classification	Multi	30	12.1	0.052	2.3
Transformer	Regression	UNI	1135	5.7	0.034	1.6
Transformer	Classification	UNI	493	3.9	0.041	0.9
LSTM	Regression	UNI	42	11.6	0.019	6.0
LSTM	Classification	UNI	151	11.0	0.051	2.1
ARIMA (1,0,1)	Regression	UNI	916	2.9	0.045	0.6
GBP USD	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	532	24.1	0.045	5.2
Transformer	Classification	Multi	1‘175	14.0	0.062	2.2
LSTM	Regression	Multi	20	8.6	0.062	1.3
LSTM	Classification	Multi	45	9.9	0.061	1.6
Transformer	Regression	UNI	733	8.0	0.031	2.4
Transformer	Classification	UNI	467	7.4	0.048	1.5
LSTM	Regression	UNI	1	8.6	0.062	1.3
LSTM	Classification	UNI	20	10.9	0.062	1.7
ARIMA (1,0,1)	Regression	UNI	687	3.7	0.047	0.7

Empirical derivation: The transformer model has outperformed LSTM models in countless non-financial-time-series tasks. Based on prior successes in non-financial time-series tasks, it is reasonable to assume that transformer models will also perform well in the context of financial and FX prediction.

Data was used to investigate whether higher dimensionality of data leads to better prediction performance. Multivariate input data is expected to improve the prediction performance of DL models and that this is especially true for the transformer models. This is because DL models are able to recognize essential features themselves and accordingly no feature engineering is necessary. It is therefore assumed that the prediction performance either remains the same or is improved by additional input features.

Since regression and classification were used as tasks in the existing literature, the task was included as a variable and conducted experiments for both regression and classification. As a result, any significant

Table 4
Model performance evaluation of prediction for (a) EURUSD, (b) USDJPY, and (c) GBPUSD.

EUR USD	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	540	9.2	0.036	2.4
Transformer	Classification	Multi	1‘199	7.1	0.066	1.0
LSTM	Regression	Multi	7	−4.8	0.067	−0.8
LSTM	Classification	Multi	20	0.5	0.008	0.3
Transformer	Regression	UNI	1‘095	4.5	0.052	0.8
Transformer	Classification	UNI	627	3.7	0.064	0.5
LSTM	Regression	UNI	1	−0.7	0.067	−0.9
LSTM	Classification	UNI	215	−9.5	0.050	−1.9
ARIMA (1,0,1)	Regression	UNI	1‘128	17.1	0.046	3.6
USD JPY	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	574	22.9	0.041	5.5
Transformer	Classification	Multi	48	12.2	0.029	4.1
LSTM	Regression	Multi	119	−0.3	0.028	−0.2
LSTM	Classification	Multi	30	12.1	0.052	2.3
Transformer	Regression	UNI	1135	5.7	0.034	1.6
Transformer	Classification	UNI	493	3.9	0.041	0.9
LSTM	Regression	UNI	42	11.6	0.019	6.0
LSTM	Classification	UNI	151	11.0	0.051	2.1
ARIMA (1,0,1)	Regression	UNI	916	2.9	0.045	0.6
GBP USD	Task	Data	Trades	r (%)	σ	Shape Ratio
Transformer	Regression	Multi	532	24.1	0.045	5.2
Transformer	Classification	Multi	1‘175	14.0	0.062	2.2
LSTM	Regression	Multi	20	8.6	0.062	1.3
LSTM	Classification	Multi	45	9.9	0.061	1.6
Transformer	Regression	UNI	733	8.0	0.031	2.4
Transformer	Classification	UNI	467	7.4	0.048	1.5
LSTM	Regression	UNI	1	8.6	0.062	1.3
LSTM	Classification	UNI	20	10.9	0.062	1.7
ARIMA (1,0,1)	Regression	UNI	687	3.7	0.047	0.7

differences are expected between the tasks.

Thus, the multivariate transformer model is expected to provide the best prediction performance among the compared ones. The closing returns 10 min into the future are predicted for the following three currency pairs: EURUSD, USDJPY, and GBPUSD.

7. Empirical results

In particular, the following research questions are examined in this section:

i. Are the transformer models able to achieve better prediction performance than the benchmark models?
ii. Are the multivariate models able to achieve better prediction performance than the univariate models?

Therefore, the evaluation of the results is twofold: a systematic comparison between transformer and benchmark models and a systematic comparison between multivariate, cross-sectional, and univariate input data, see Table 3 for the results.

To provide a more detailed understanding of the specific differences between the univariate and multivariate models, we acknowledge the importance of explicitly outlining the features that are retained or ablated in each scenario. This includes specifying whether certain fundamental or technical features are excluded or whether features from other asset classes are omitted in the univariate model compared to the multivariate counterpart. Addressing these differences will contribute to the comprehensibility of the experimental results.

First, a comparison between the performance of the transformer architecture and the other approaches along the dimensions of annualized return, annualized standard deviation, and annualized Sharpe ratio prior to transaction costs was made. Irrespective of the currency pair, the best transformer architecture shows favorable characteristics vis-à-vis the other approaches. Specifically, average returns prior to transaction costs are at 18.7 % for the best transformer, compared to 6.8 % for the best LSTM. However, the standard deviation is similar for both models. This results in an excellent average Sharpe ratio of 4.4 for the best transformer compared to 1.2 for the best LSTM. Hypothesis (I) is confirmed and the transformer models achieve better prediction performance than the benchmark models.

Secondly, the multivariate results can be compared to the univariate results. These show that only the multivariate transformer can improve forecasting performance through additional multivariate input data. At the same time, however, one can also see that the multivariate transformer has the best forecasting performance and that none of the univariate models has similarly high returns or Sharpe ratios. Hypothesis (II) is confirmed and the multivariate models achieve better prediction performance than the univariate models. Compared to the fitted ARIMA model predictions, the transformer model yields better results for the currency pairs USDJPY and GBPUSD and worse results for the EURUSD pair. In the latter case, the transformer model's power to contextualize adds complexity that is disadvantageous for this FX pair.

Overall, transformer models are able to achieve good forecasting performance for FX-Spot forecasting in general but also evidence the need for transformer models for multivariate, cross-sectional input data to outperform other state-of-the-art DL networks such as LSTM.

The next step is to identify the sources of the differences in the results. To accomplish this, a comparison between the performance of the transformer and the LSTM models was investigated by analyzing their predictions in detail. A focus on identifying any differences in prediction performance rather than analyzing the models' overall performance was made. An analysis of the USDJPY out-of-sample test data found that the multivariate transformer model outperformed the multivariate LSTM model in 15.5 % of test predictions. The LSTM model outperformed the transformer model in 14.0 % predictions. Looking at the top ten cases where one model performed particularly well compared to the other, it was found that these cases only accounted for a small fraction of the overall difference in performance. That suggests that the differences between the models are due to many relatively unimportant cases rather than a few key predictions. This also highlights the robustness of the results as the better forecasting performance is not based on a few “lucky shots” but is consistently present throughout the model.

After analyzing where the differences between the transformer & LSTM model come from, the next step is to analyze whether there are any similarities and patterns in the cases where the transformer model outperforms. Examples of this would be the realization that the transformer outperforms particularly strongly in particularly volatile market phases or when vital news or economic data appear.

In the first step, linear and non-linear correlation and causality analysis were carried out to analyze whether one or more of the input features is a primary driver for the different prediction performances between the models. However, no feature with statistical significance could be identified.

In order to bundle the information from many individual variables into a few main components and thus check whether one of these main components can explain the differences, a principal component analysis was carried out in the next step. Even the main components identified in this way could not explain the differences in the prediction performance between the transformer & LSTM model statistically significantly across the different currency pairs. This leads to the realization that there are no recognizable commonalities in the cases where the transformer model outperforms and the whole accordingly corresponds to a black box in certain parts. A limitation, however, is that only the data used in the experiments could be used to analyze similarities. Since in example the economic data published at regular intervals (e.g., unemployment or inflation figures) are not part of the data set, this could not be identified as a commonality, even if it were one.

The predictions and prediction performance results observed in the experiments suggest that statistical metrics do not capture these findings. In order to check this assumption, statistical metrics for the predictions are collected and analyzed in the next step. The selected statistical metrics include mean squared error (MSE) and mean absolute error (MAE). The MSE measures the sum of the squared deviations between the prediction and the true value, while the MAE measures the absolute deviation between the prediction and the true value.

Practically no differences can be seen in the statistical metrics, which confirms the assumption made. This is mainly due to the intraday focus, which leads to a considerable amount of data and accordingly to a strong robustness of the experiments, but at the same time also blurs insights into statistical metrics such as the MSE and MAE.

8. Summary, implications and limitations

This paper applies transformer architecture networks to a large-scale financial market prediction task on the EURUSD, USDJPY, and GBPUSD currency pairs, from November 2020 until January 2022. With our work, three key contributions to the literature were made: The first contribution focuses on the introduction and performance assessment of a transformer with time embeddings to the empirical literature on FX forecasting. Specifically, we frame a proper prediction task, derive sensible features, standardize the features during preprocessing to facilitate model training, discuss a suitable transformer architecture and training algorithm, and derive a trading strategy based on the predictions, in line with the existing literature. A comparison between the results of the transformer against an LSTM as well as a simple ARIMA regression was investigated. The transformer, a methodology inherently suitable for this domain, beat the benchmark models by an evident margin. Our findings of statistically and economically significant average returns of 18.7 % per year prior to transaction costs posit a clear challenge to the semi-strong form of market efficiency and show that DL can be effectively deployed in this domain. This is especially true since the transaction costs for the currency pairs examined are extremely low at around 0.002 % per trade. With an average of 549 trades per currency pair, the transaction costs would be 1,098 %. Accordingly, the average return would still be clearly positive even after transaction costs. Also, the conceptual and empirical aspects of transformer networks outlined in this paper go beyond a pure financial market application but are intended as a practical example for other researchers, wishing to deploy this effective methodology to other time series prediction tasks with large amounts of training data.

Second, a contribution by systematically comparing the prediction performance between univariate and multivariate input data for the transformer and the state-of-the-art benchmark models was made. This shows that the hypothesis is confirmed that multivariate technical and fundamental features improve forecast performance, particularly for the transformer architecture.

Third, the focus on intraday predictions from a 10-minute observation period can also be viewed as a contribution compared to most other literature focusing only on daily forecasts.

In this paper, a successful demonstration, was made, proves that a transformer network can effectively extract meaningful information from noisy financial time series data. Moreover, compared to LSTM and ARIMA, it is the method of choice with respect to yearly returns before transaction costs. As it turns out, DL - in the form of transformer networks - hence seems also to constitute an advancement in this domain.

CRediT authorship contribution statement

Tizian Fischer: Conceptualization, Writing – original draft, Visualization, Methodology, Project administration, Software. **Marius Sterling:** Conceptualization, Writing – original draft, Visualization, Methodology, Project administration, Software. **Stefan Lessmann:** Conceptualization, Funding acquisition, Writing – review & editing, Methodology, Supervision.

Declaration of competing interest

Stefan Lessmann acknowledges funding through the project “AI4E-Fin AI for Energy Finance”, contract number CF162/15.11.2022, financed under the Romania’s National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8.

Data availability

Experimental results are available as DataFrame/Excel File upon request.

References

Abarbanell, J. S. (1991). Do analysts’ earnings forecasts incorporate information in prior stock price changes? *Journal of Accounting and Economics, 14*(2), 147–165. https://doi.org/10.1016/0165-4101(91)90003-7

Achkar, R., Elias-Sleiman, F., Ezzidine, H., & Haidar, N. (2018). Comparison of BPA-MLP and LSTM-RNN for Stocks Prediction. 2018 6th International Symposium on Computational and Business Intelligence (ISCBI). DOI: 48–51. https://doi.org/10.1109/ISCBI.2018.00019.

Almeida, A., Goodhart, C., & Payne, R. (1998). The effects of macroeconomic news on high frequency exchange rate behavior. *Journal of Financial and Quantitative Analysis, 33*(3), 383–408. https://doi.org/10.2307/2331101

Alpaydin, E. (2021). Machine learning. *MIT Press.* https://mitpress.mit.edu/9780262529518/machine-learning/.

Althelaya, K. A., El-Alfy, E. S. M., & Mohammed, S. (2018). *April). Evaluation of bidirectional LSTM for short-and long-term stock market prediction* (pp. 151–156). IEEE.

Arnott, R., Harvey, C., & Markowitz, H. (2019). A Backtesting Protocol in the Era of Machine Learning| journal = The Journal of Financial Data Science. *ISSN, 64–74*, 2640–3943. https://doi.org/10.2139/ssrn.3275654

Aslam, F., Aziz, S., Nguyen, D. K., Mughal, K. S., & Khan, M. (2020). On the efficiency of foreign exchange markets in times of the COVID-19 pandemic. *Technological forecasting and social change, 161*, Article 120261. https://doi.org/10.1016/j.techfore.2020.120261

Ausmees, K., Milovanovic, S., Wrede, F., & Zafari, A. (2017). Taming Deep Belief Networks. Retrieved from https://api.semanticscholar.org/CorpusID:43756575.

Baek, Y., & Kim, H. Y. (2018). ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. *Expert Systems with Applications, 113*, 457–480.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems, 33*, 12449–12460.

Barroso, P., & Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics, 116*(1), 111–120. https://doi.org/10.1016/j.jfineco.2014.11.010

Bartram, S. M. (2008). What lies beneath: Foreign exchange rate exposure, hedging and cash flows. *Journal of Banking & Finance, 32*(8), 1508–1521. https://doi.org/10.1016/j.jbankfin.2007.07.013

Bekaert, G., Wei, M., & Xing, Y. (2007). Uncovered interest rate parity and the term structure. *Journal of International Money and Finance, 26*(6), 1038–1069. https://doi.org/10.1016/j.jimonfin.2007.05.004

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments, 65*(6), 1803–1832. https://doi.org/10.1063/1.1144830

Bradshaw, M. T. (2011). Analysts’ forecasts: What do we know after decades of work? *Available at SSRN, 1880339*. https://doi.org/10.2139/ssrn.1880339

Chang, V., Man, X., Xu, Q., & Hsu, C.-H. (2020). Pairs trading on different portfolios based on machine learning. *Expert Systems, 38*(3), e12649.

Chen, Y., He, K., & Tso, G. K. (2017). Forecasting crude oil prices: a deep learning based model. *Procedia computer science, 122*, 300–307.

Chen, L., Qiao, Z., Wang, M., Wang, C., Du, R., & Stanley, H. E. (2018). Which artificial intelligence algorithm better predicts the Chinese stock market? *IEEE Access, 6*, 48625–48633.

Chen, Y., Wu, J., & Bu, H. (2018, July). *Stock market embedding and prediction: A deep learning method* (pp. 1–6). IEEE.

Chen, M.-Y., Liao, C.-H., & Hsieh, R.-P. (2019). Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. *Computers in Human Behavior, 101*, 402–408. https://doi.org/10.1016/j.chb.2019.03.021

Chen, M., Radford, A. Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning (ICML’20), Vol. 119. JMLR.org, Article 158, 1691–1703. https://dl.acm.org/doi/10.5555/3524938.3525096.

Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia computer science, 132*, 956–964.

Dautel, A. J., Härdle, W. K., Lessmann, S., & Seow, H. V. (2020). Forex exchange rate forecasting using deep recurrent neural networks. *Digital Finance, 2*(1), 69–96. https://doi.org/10.1007/s42521-020-00019-x

Davidson, P. (2003). Financial markets, money, and the real world. *Edward Elgar Publishing.* https://www.abebooks.fr/9781843764847/Financial-Markets-Money-Real-World-1843764849/plp.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, MN, USA, June 2–7, 2019). Association for Computational Linguistics, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Di Persio, L., & Honchar, O. (2017). Recurrent neural networks approach to the financial forecast of Google assets. *International journal of Mathematics and Computers in simulation, 11*, 7–13.

dos Santos Pinheiro, L., & Dras, M. (2017, December). Stock market prediction with deep learning: A character-based neural language model for event-based trading. In Proceedings of the Australasian Language Technology Association Workshop 2017 (pp. 6-15).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Elliot, A., & Hsu, C. H. (2017). Time series prediction: Predicting stock price. arXiv preprint arXiv:1710.05751.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance, 25*(2), 383–417. https://doi.org/10.2307/2325486

Fazeli, A., & Houghten, S. (2019). Deep Learning for the Prediction of Stock Market Trends. *IEEE International Conference on Big Data (Big Data), 2019*, 5513–5521. https://doi.org/10.1109/BigData47090.2019.9005523

Feng, G., Polson, N. G., & Xu, J. (2018). Deep learning factor alpha. arXiv preprint arXiv: 1805.01104, 2326-2377.

Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T.-S. (2019). Temporal Relational Ranking for Stock Prediction. *ACM Transactions on Information Systems, 37*(2). https://doi.org/10.48550/arXiv.1809.09441, 27:1–27:30.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Frankel, J. A., & Froot, K. A. (1990). Chartists, fundamentalists, and trading in the foreign exchange market. *The American Economic Review, 80*(2), 181–185. http://www.jstor.org/stable/2006566.

Gyntelberg, J., & Schrimpf, A. (2011). FX strategies in periods of distress. BIS Quarterly Review, December. https://ssrn.com/abstract=1971145.

Hall, K., Jayne, C., & Chang, V. (2023). A Transformer-Based Framework for Biomedical Information Retrieval Systems. In *International Conference on Artificial Neural Networks* (pp. 317–331). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44223-0_26.

Hansson, M. (2017). On stock return prediction with LSTM networks.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2,, 1–758.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hollis, T., Viscardi, A., & Yi, S. (2018). A Comparison of LSTMs and Attention Mechanisms for Forecasting Financial Time Series. *ArXiv.* https://arxiv.org/pdf/1812.07699.pdf.

Irie, K., Tüske, Z., Alkhouli, T., Schlüter, R., & Ney, H. (2016, September). LSTM, GRU, highway and a bit of attention: an empirical overview for language modeling in speech recognition. In Interspeech (pp. 3519-3523). https://doi.org/10.21437/Interspeech.2016-491.

Irwin, R., Dimitriadis, S., He, J., & Bjerrum, E. J. (2022). Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology, 3*(1), Article 015022.

Iwasaki, H., Chen, Y., Du, Q., & Tu, J. (2018). Topic Sentiment Asset Pricing with DNN Supervised Learning. Social Science Research Network. Retrieved from https://doi.org/10.2139/ssrn.3228485.

Jeong, G., & Kim, H. Y. (2019). Improving financial trading decisions using deep Q-learning: Predicting the number of shares, action strategies, and transfer learning. *Expert Systems with Applications, 117*, 125–138.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications, 32*(13), 9713–9729. https://doi.org/10.1007/s00521-019-04504-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., ... & Brubaker, M. (2019). Time2vec: Learning a vector representation of time. arXiv preprint arXiv: 1907.05321.

Lachiheb, O., & Gouider, M. S. (2018). A hierarchical deep neural network design for stock returns prediction. *Procedia Computer Science, 126*, 264–272.

Lai, C. Y., Chen, R.-C., & Caraka, R. E. (2019). Prediction Stock Price Based on Different Index Factors Using LSTM. 2019 International Conference on Machine Learning and Cybernetics (ICMLC), 1–6. https://doi.org /10.1109/ICMLC48188.2019.894916.

Lakshminarayanan, S., & McCrae, J. P. (2019). A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction. *Irish Conference on Artificial Intelligence and Cognitive Science.* https://www.semanticscholar.org/paper/A-Comparative-Study-of-SVM-and-LSTM-Deep-Learning-Lakshminarayanan-McCrae/fa0cdb970e41ed319a39cf959f40fb3f148e6d6e.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444. https://doi.org /10.1038/nature14539.

Lee, C. Y., & Soo, V. W. (2017). December). Predict stock price with financial news based on recurrent convolutional neural networks. In *2017 conference on technologies and applications of artificial intelligence (TAAI)* (pp. 160–165). IEEE.

Li, C., Song, D., & Tao, D. (2019). Multi-task Recurrent Neural Networks and Higher-order Markov Random Fields for Stock Price Movement Prediction: Multi-task RNN and Higer-order MRFs for Stock Price Classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1141–1151). https://doi.org/10.1145/3292500.3330983

Li, Y., Ni, P., & Chang, V. (2020). Application of deep reinforcement learning in stock trading strategies and stock forecasting. *Computing, 102*, 1305–1322. https://doi.org/10.1007/s00607-019-00773-w

Lim, B., Arık, S., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, S., Zhang, C., & Ma, J. (2017). *CNN-LSTM neural network model for quantitative strategy analysis in stock markets, 24* pp. 198–206). Springer International Publishing.

Liu, J., Lin, H., Liu, X., Xu, B., Ren, Y., Diao, Y., & Yang, L. (2019, August). Transformer-based capsule network for stock movement prediction. In Proceedings of the first workshop on financial technology and natural language processing (pp. 66-73).

Liu, Y., Wang, Z., Yu, X., Chen, X., & Sun, M. (2022). Memory-based Transformer with shorter window and longer horizon for multivariate time series forecasting. *Pattern Recognition Letters, 160*, 26–33. https://doi.org/10.1016/j.patrec.2022.05.010

Long, J., Chen, Z., He, W., Wu, T., & Ren, J. (2020). An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. *Applied Soft Computing, 91*, Article 106205. https://doi.org/10.1016/j.asoc.2020.106205

Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science, 7*, e476.

Mikhaylov, A., Sokolinskaya, N., & Nyangarika, A. (2018). Optimal carry trade strategy based on currencies of energy and developed economies. *Journal of Reviews on Global Economics, 7*, 582–592. https://doi.org/10.6000/1929-7092.2018.07.54

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781.

Mussa, M. (2019). *The exchange rate, the balance of payments and monetary and fiscal policy under a regime of controlled floating* (pp. 97–116). Routledge.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & Shahab, S. (2020). Deep learning for stock market prediction. *Entropy, 22*(8), 840. https://doi.org/10.3390/e22080840

Naik, N., & Mohan, B. R. (2019). Study of Stock Return Predictions Using Recurrent Neural Networks with LSTM. In J. Macintyre, L. Iliadis, I. Maglogiannis, & C. Jayne (Eds.), *Engineering Applications of Neural Networks* (pp. 453–459). Springer International Publishing. https://doi.org/10.1007/978-3-030-20257-6_39.

Nguyen, D. H. D., Tran, L. P., & Nguyen, V. (2019). Predicting Stock Prices Using Dynamic LSTM Models. In H. Florez, M. Leon, J. M. Diaz-Nafria, & S. Belli (Eds.), *Applied Informatics* (pp. 199–212). Springer International Publishing. https://doi.org/10.1007/978-3-030-32475-9_15.

Ni, L., Li, Y., Wang, X., Zhang, J., Yu, J., & Qi, C. (2019). Forecasting of forex time series data based on deep learning. *Procedia Comput. Sci., 147*, 647–652. https://doi.org/10.1016/j.procs.2019.01.189

Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management, 26*(4), 164–174. https://doi.org/10.1002/isaf.1459

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review, 53*(4), 3007–3057. https://doi.org/10.1007/s10462-019-09754-z

OpenAI (2023a). GPT-4 Technical Report, 2023. arXiv:2303.08774 (2023). arxiv.org/abs/2303.08774.

OpenAI. (2023b). ChatGPT. https://chat.openai.com.

Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2020). An innovative neural network approach for stock market prediction. *The Journal of Supercomputing, 76*(3), 2098–2118. https://doi.org/10.1007/s11227-017-2228-y

Qi, L., Khushi, M., & Poon, J. (2020). Event-Driven LSTM For Forex Price Prediction. *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020*, 1–6. https://doi.org/10.1109/CSDE50874.2020.9411540

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR. https://arxiv.org/abs/2103.00020.

Rana, M., Uddin, M. M., & Hoque, M. M. (2020). Effects of Activation Functions and Optimizers on Stock Price Prediction using LSTM Recurrent Networks. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence* (pp. 354–358). https://doi.org/10.1145/3374587.3374622

Rogoff, K. (1996). The purchasing power parity puzzle. *Journal of Economic literature, 34*(2), 647–668.

Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences, 9*(24), 5574. https://doi.org/10.3390/app9245574

Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of big Data, 7*(1), 1–33.

Shen, G., Tan, Q., Zhang, H., Zeng, P., & Xu, J. (2018). Deep learning with gated recurrent unit networks for financial sequence predictions. *Procedia computer science, 131*, 895–903.

Si, W., Li, J., Ding, P., & Rao, R. (2017). December). A multi-objective deep reinforcement learning approach for stock index future's intraday trading. In , *2. 2017 10th International symposium on computational intelligence and design (ISCID)* (pp. 431–436). IEEE.

Siami-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: ARIMA vs. LSTM. arXiv preprint arXiv:1803.06386.

Stasiulis, D. (2009). *Semi-strong form efficiency in the CEE stock markets*. Riga: Rigas Ekonomikas Augstskola Stockholm School of Economics.

Sun, J., Qing, Y., Liu, C., & Lin, J. (2022). In *Self-FTS: A Self-Supervised Learning Method for Financial Time Series Representation in Stock Intraday Trading* (pp. 501–506). IEEE. https://doi.org/10.1109/INDIN51773.2022.9976077.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., … & Huang, X. (2021, July). Unispeech: Unified speech representation learning with labeled and unlabeled data. In International Conference on Machine Learning (pp. 10937-10947). PMLR.

Wang, J., Sun, T., Liu, B., Cao, Y., & Wang, D. (2018). December). Financial markets prediction with deep learning. In *17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 97–104). IEEE.

Widegren, P. (2017). Deep learning-based forecasting of financial assets (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-208308.

Wu, N., Green, B., Ben, X., & O'Banion, S. (2020). Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case (arXiv:2001.08317). arXiv. http://arxiv.org/abs/2001.08317.

Xu, Y., & Keselj, V. (2019). December). Stock prediction using deep learning and sentiment analysis. In *2019 IEEE international conference on big data (big data)* (pp. 5573–5580). IEEE.

Yong, B. X., Abdul Rahim, M. R., & Abdullah, A. S. (2017). A stock market trading system using deep neural network. In Modeling, Design and Simulation of Systems: 17th Asia Simulation Conference, AsiaSim 2017, Melaka, Malaysia, August 27–29, 2017, Proceedings, Part I 17 (pp. 356-364). Springer Singapore.

Yuan, Z., Zhang, R., & Shao, X. (2018, May). Deep and wide neural networks on multiple sets of temporal data with correlation. In Proceedings of the 2018 International Conference on Computing and Data Engineering (pp. 39-43).

Yümlü, S., Gürgen, F. S., & Okay, N. (2005). A comparison of global, recurrent and smoothed-piecewise neural models for Istanbul stock exchange (ISE) prediction. *Pattern Recognition Letters, 26*(13), 2093–2103.

Zeng, Z., & Khushi, M. (2020, July). *Wavelet denoising and attention-based RNN-ARIMA model to predict forex price* (pp. 1–7). IEEE.

Zhao, Z., Rao, R., Tu, S., & Shi, J. (2017). November). Time-weighted LSTM model with redefined labeling for stock trend prediction. In *2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI)* (pp. 1210–1217). IEEE.

Zhou, Y. L., Han, R. J., Xu, Q., Jiang, Q. J., & Zhang, W. K. (2019). Long short-term memory networks for CSI300 volatility prediction with Baidu search volume. Concurrency and Computation. *Practice and Experience, 31*(10), e4721.

Zhou, B. (2019). Deep learning and the cross-section of stock returns: Neural networks combining price and fundamental information. Available at SSRN 3179281.