Stock Market Analysic and Predicition Sentiment Analysic

SAKTHI HARIRAJAN R[1], SARANARAYANAN R[2], MR ANBUTHIRUVARANGAN K[3], LOGESVARAN SK[4]

[1,2,4] *Department of Computer Science Engineering Sri Manakula Vinayagar Engineering College, Madagadipet Puducherry 605107, India.*

[3]*(Assistant Professor) Department of computer of computer science Engineering Sri Manakula vinayagar Engineering College, Madagadipet Puducheery 605107, India*

Abstract- **Artificial Intelligence (AI) has rapidly evolved into a transformative field, fundamentally altering how data-driven predictions and decision-making processes are executed across various domains. Within financial analytics, Machine Learning (ML) and Deep Learning (DL) techniques have enabled advancements in stock market prediction, achieving accuracy levels in the range of 65–70%. However, inherent market volatility and the influence of numerous factors make predicting stock trends complex and less reliable when relying solely on single models. Decision fusion—an approach that combines predictions from multiple base models—has emerged as a promising solution, effectively mitigating the limitations of individual models by integrating their strengths. Yet, there is a scarcity of systematic studies examining the implementation of decision fusion specifically for stock market prediction. Our research addresses this gap, providing a comprehensive analysis of base learner properties, fusion techniques, and practical recommendations drawn from foundational studies, including our primary source material, to achieve improved prediction precision. The recent advancements and emerging directions within AI-driven financial forecasting. Notably, the integration of decision fusion with sentiment analysis has demonstrated potential in capturing market sentiments reflected in real-time data, such as news headlines and social media feeds. By combining sentiment insights with historical stock data and multi-source decision fusion, our approach offers a more holistic model for stock market prediction. Future directions include refining the model's capabilities to handle various data sources effectively, thus improving its adaptability and predictive power. Ultimately, this study aims to establish an advanced framework for stock market analysis, leveraging the latest in AI to provide more accurate and robust predictions that can support better-informed financial decision making in a constantly changing market landscape.**

Keywords: **Stock price prediction - Deep learning - Time series analysis - Long Short -Term Memory (LSTM) networks - Gated Recurrent Units (GRU)- Convolutional Neural Networks (CNN).**

I. INTRODUCTION

The stock market serves as a critical component of the global economy, facilitating the exchange of capital and providing investors with opportunities for wealth accumulation. However, the inherent volatility and complexity of stock prices pose significant challenges for investors seeking to make informed decisions. This study explores the application of deep learning algorithms in stock market forecasting, emphasizing their potential to enhance predictive accuracy through the integration of various data sources.

1.1 Background

Stock market forecasting involves predicting future price movements based on historical data and various analytical techniques. Traditional methods, such as fundamental analysis and technical analysis, have long been used to assess stock performance. However, these methods often fall short in capturing the intricate patterns and temporal dependencies present in financial data. The advent of machine learning and deep learning technologies has revolutionized this field, enabling more sophisticated models that can analyze vast datasets and identify complex relationships within them. This shift towards data-driven approaches reflects a growing recognition of the limitations of conventional forecasting methods in navigating the dynamic nature of financial markets.

1.2 Importance of Stock Market Forecasting

The significance of accurate stock market forecasting cannot be overstated. Effective predictions allow investors to make informed decisions regarding entry and exit points, thereby maximizing profits and minimizing losses. Forecasting plays a pivotal role in:

- Investment Strategy Development: Investors rely on forecasts to formulate strategies that align with market trends, helping them capitalize on potential gains while mitigating risks.
- Market Timing: Understanding future price movements enables traders to time their trades effectively, ensuring they buy low and sell high.
- Risk Management: Accurate predictions assist in identifying potential downturns or volatility, allowing investors to adjust their portfolios proactively.
- Behavioral Insights: Forecasting can help investors overcome cognitive biases by providing objective data-driven insights rather than relying solely on intuition or emotions.

With advancements in artificial intelligence (AI) and machine learning, particularly deep learning techniques like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), the accuracy and reliability of stock market predictions have significantly improved. These technologies enable the analysis of multifaceted datasets, including historical prices, technical indicators, and macroeconomic factors, resulting in models that are better equipped to navigate market complexities.

1.3 Objectives of the Study

This study aims to achieve the following objectives:

1. Evaluate Deep Learning Models: Assess the effectiveness of various deep learning architectures, particularly LSTM and CNN, in predicting stock price movements compared to traditional forecasting methods.
2. Integrate Diverse Data Sources: Explore how incorporating historical price data, technical indicators, and macroeconomic factors can enhance model performance.
3. Analyze Model Performance: Investigate backtesting strategies to evaluate model accuracy while addressing challenges such as overfitting and market volatility.
4. Provide Practical Insights: Offer actionable insights for investors and financial analysts on leveraging deep learning models for improved decision-making in stock trading.

2.1 Traditional Methods of Stock Market Prediction

Traditional stock market prediction methods primarily fall into three categories: fundamental analysis, technical analysis, and quantitative technical analysis.

- Fundamental Analysis: This method evaluates a company's financial health through its financial statements, such as balance sheets and income statements. Analysts look at economic factors that influence stock prices, such as earnings reports and market conditions. Fundamental analysis is typically used for long-term investment strategies since it relies on slower-moving data.
- Technical Analysis: Technical analysts focus on historical price movements and trading volumes to predict future price trends. They use various indicators derived from past data to identify patterns and trends in stock prices. This approach is largely qualitative, relying on visual chart analysis to gauge market sentiment.
- Quantitative Technical Analysis: Unlike traditional technical analysis, this method employs quantitative models to analyze stock price movements. It utilizes statistical techniques and algorithms to derive predictions from historical data rather than relying solely on visual interpretations.

These traditional methods have been foundational in stock market forecasting but often lack the ability to capture complex patterns present in large datasets.

2.2 Overview of Machine Learning in Finance

The integration of machine learning into finance has transformed stock market prediction by introducing more sophisticated analytical tools. Machine learning models can process vast amounts of data and identify intricate patterns that traditional methods might overlook. Key machine learning techniques applied in finance include

- Regression Models: These models predict future stock prices based on historical data, capturing relationships between variables.
- Decision Trees and Ensemble Methods: Techniques like Random Forests combine multiple decision trees to improve prediction accuracy by reducing overfitting.
- Support Vector Machines (SVM): SVMs classify data points into different categories based on their features, making them useful for predicting stock price movements.

Research has shown that machine learning models can outperform traditional statistical methods like ARIMA (AutoRegressive Integrated Moving Average) in terms of predictive accuracy. However,

while machine learning enhances forecasting capabilities, it still faces challenges related to model interpretability and the need for extensive training data.

2.3 Deep Learning Techniques in Stock Market Analysis

Deep learning represents a significant advancement in the field of stock market prediction, leveraging complex neural network architectures to model non-linear relationships within data. Prominent deep learning techniques include:

- Long Short-Term Memory (LSTM) Networks: LSTMs are particularly effective for time-series forecasting due to their ability to remember long-term dependencies in sequential data. Studies have demonstrated that LSTMs can achieve higher accuracy than traditional models like ARIMA when predicting stock prices
- Convolutional Neural Networks (CNNs): While primarily used for image processing, CNNs have been adapted for financial data analysis by extracting features from time-series data. They are effective at identifying patterns that may not be immediately visible through traditional analytical methods.
- Hybrid Models: Recent research has explored combining LSTM with other techniques, such as Gated

Overall, deep learning techniques have shown great promise in improving the accuracy of stock price predictions by effectively capturing complex relationships within multifaceted datasets. However, challenges remain regarding the interpretability of these models and their ability to adapt to rapidly changing market conditions.

III. LITERATURE SURVEY

There were two important indicators in the literature for stock price forecasting. They are fundamental and technical analysis. Both were used to analyze the stock market [8, 9]

3.1 Prediction Techniques Presented the recent methods for the prediction of stock market and give a comparative analysis [10] of all these Techniques.

Major prediction techniques such as data mining, machine learning and deep learning techniques used to estimate the future stock prices based on these techniques and discussed their advantages and disadvantages. They are, 3.1.1 Holt-Winters 3.1.2 Artificial Neural Network 3.1.3 Hidden Markov Model 3.1.4 ARIMA Model 3.1.5 Time Series Linear Model 3.1.6 Recurrent Neural Networks. Holt-Winters, Artificial Neural Network, Hidden Markov Model are Machine Learning Techniques, ARIMA Model is Time series technique and Time series Linear Model and Recurrent Neural Networks are Deep Learning Techniques. A Survey on Stock Market Prediction Using Machine Learning 925 3.1.1 Holt-Winters Holt-Winters is the appropriate or correct mode when the time series has trend and seasonal factors. The series was divided into three components or parts that are trend, basis and seasonality. Holt-Winters find three trend, level, and seasonal smoothening parameters. It has two variants: Additive Holt Winters Smoothening model and Multiplicative

Holt-Winters model.

The former is used for prediction and the latter is preferred if there are no constant seasonal variations in the series. It is mainly popular for its accuracy and in the field of prediction it has outperformed many other models. In short—term forecasts of economic development trends, Holt-Winters exponential smoothing method with the trend and seasonal fluctuations is usually used. After removing the seasonal trends from the data, the following function is taken as an input and in return, the Holt-Winters makes the pre-calculations necessary for the purpose of forecasting. All parameters required for the forecasting purpose are automatically initialized based on the function data. HWStock1_ng = HoltWinters(ds,gamma = FALSE) predHW = predict(HWStock1_ng,n.ahead = 9)

Artificial Neural Network

An artificial neural network (ANN) is a technique inspired from biological nervous system, such as the human brain [5, 10]. It has a great ability to predict from large databases [11]. On the basis of the back—propagation algorithm, ANN is generally used to forecast the stock market. In the back—propagation algorithm, a neural network of multilayer perceptron (MLP) is used. It consists of an input layer with a set of sensor nodes as input nodes, one or more hidden layers of computation nodes and computation nodes of the output layer. These networks often use raw data and data derived from the previously discussed technical and fundamental analysis [11, 12]. A Multilayer Feed forward Neural Network is a neural

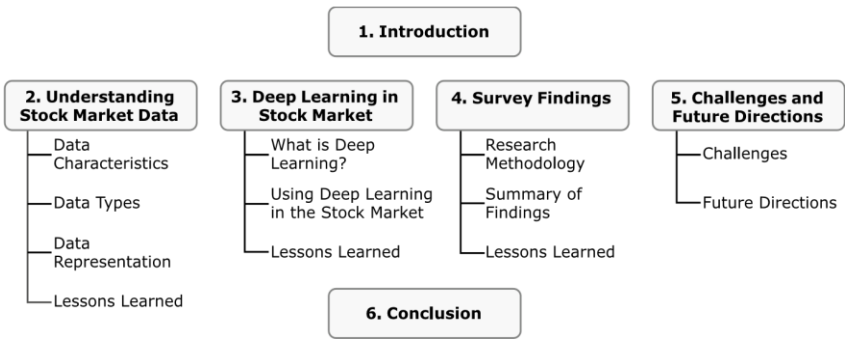
network with an input layer, one or more hidden layers and an output layer. Inputs correspond to each training sample measured attributes. Inputs are passed to input layer simultaneously. The weighted outputs of these units are fed to the next layer of units that make up the hidden layer simultaneously. The weighted outputs of the hidden layers act as an input to another hidden layer, etc. The hidden layers number is an arbitrary design problem. The weighted output of the last the hidden layer acts as inputs to the output layer, which predicts the networks for certain samples. Important parameters of NN are learning rate, momentum and epoch (Fig. 1). Back propagation is a neural network learning algorithm [13]. Use the back propagation algorithm to perform the calculations and compare the predicted output and target output ANNs has been used to solve various problems in financial time series forecasting and can predict the price with approximately 90% accuracy. Disadvantages • Neural Network is suffering from the Blackbox problem; it does not reveal the each variable's significance weight. • The problem of overtraining is another major problem with Neural Networks [9]. The system may lose the ability to generalize if Neural Networks fits the data too well. • Overtraining is a major problem. It usually happens for two main reasons such as Neural Networks have too many nodes or too long training periods.

Hidden Markov Model

In speech recognition Hidden Markov Model was first invented [16] but widely used to predict stock market related data. The stock market trend analysis is based on the Hidden Markov model, taking into account the one-day difference in close value for a given time line. The hidden sequence of states and their corresponding probability values are found for a particular observation sequence. The p probability value gives Fig. 1. Graphical representation of artificial neuron [6] A Survey on Stock Market Prediction Using Machine Learning 927 the stock price trend percentage. In the event of uncertainty, decision-makers make decisions. HMM is a stochastic model assumed to be a Markov Process with hidden states. It has more accuracy when compared to other models. The parameters of the HMM are indicated by A, B and p are found out. Advantages Hidden Markov Model gives better optimization. Disadvantages Hidden Markov Model's main problems are Evaluation, decoding and

S.no.	Techniques	Advantages	Disadvantages	Parameter used
1	Artificial neural network	Better performance compared to regression. Lower prediction error	Prediction gets worse with increased noise variation	Stock closing price
2	Support vector machine for stock prediction	Does not lose much accuracy when applied to a sample from outside the training sample	Exaggerate to minor fluctuations in the training data which decrease the predictive ability	Consumer investment, net revenue, net income, price per earnings ratio of stock, consumer spending,
3	Hidden Markov model	Used for optimization purpose	Evaluation, decoding and learning	Technical indicators
4	ARIMA	Robust and efficient	It is suitable for short term predictions only	Open, high, low, close prices and moving average
5	Time series linear model	Integrate the actual data to the ideal linear model	Traditional and the seasonal trends present in the data	Data and number of months

IV. PROPOSED SYSTEM



Architecture Diagram of Proposed System:

Tick bars Unlike time bars that capture information at regular time intervals, tick bars capture the same information at a regular number of transactions or *ticks*. Ticks are trades in the stock market that can be used to represent the movement of price in trading data (i.e., the *uptick* and *downtick*). Ticks are commonly used for different stages of modeling market data, as in the case of *backtesting*. However, historical stock market data are not as freely accessible in the form of tick bars, especially for academic research purposes. For this purpose, most of the literature reviewed uses time bars, despite its statistical inferiority for predictive purposes.

Volume bars Although tick bars exhibit better statistical properties than time bars (i.e., they are closer to independent distribution), they still feature the shortcoming of uneven distribution and propensity for outliers (de Prado 2018). This can be because a large volume of trade is placed together from accumulated bids in the order book, which gets reported as a single tick, or because orders are equally recorded as a unit, irrespective of size. That is, an order for 10 shares of a security and an order for 10,000 shares are both recorded as a single tick. Volume bars help to mitigate this issue by capturing information at every predefined volume of securities. Although volume bars feature better statistical properties than tick bars (Easley et al. 2012), they are similarly seldom used in academic research.

a. Data collection:

The initial stage involves gathering a dataset of URLs from Kaggle, a popular platform offering diverse open-source data. This dataset contains a wide range of labeled URLs, including both safe and malicious ones, which forms the foundation for training and evaluating the detection model. Kaggle datasets are valuable because they are large, diverse, and frequently updated, capturing various phishing patterns and types. Collecting data from Kaggle ensures that the model is exposed to a broad spectrum of URLs, enhancing its ability to generalize and recognize new phishing tactics effectively.

b. Pre-processing:

In this stage, raw URL data is prepared for analysis. Preprocessing involves cleaning the data by removing duplicates, irrelevant entries, or incomplete URLs that may affect the accuracy of the model. It may also include normalizing the data to ensure consistent formats across different URLs, such as converting all URLs to lowercase.

c. Feature Extraction:

After pre-processing, the next step is to extract critical features from each URL that can indicate malicious activity. Features might include the length of the URL, the presence of certain keywords, the number of dots, special characters, and information on the domain's age. Extracting these features allows the model to identify key patterns that often distinguish malicious URLs from safe ones. This phase is particularly crucial as it creates a refined dataset with essential attributes, helping the model to focus on significant factors and enabling the RNN and autoencoder to learn effectively.

d. Model creation using RNN and Autoencoder:

The core of the detection system is created by combining Recurrent Neural Networks (RNNs) with autoencoders. RNNs are designed to handle sequential data, making them suitable for analyzing URLs, which are character sequences with specific patterns. The autoencoder, on the other hand, reduces dimensionality, focusing on relevant data patterns while filtering out noise. This combination allows the model to capture complex relationships within URL sequences, providing a more nuanced analysis than simpler models. The RNN identifies and learns sequential patterns, while the autoencoder helps in reconstructing data to improve feature representation, resulting in a robust model capable of distinguishing between safe and malicious URLs.

e. Test data:

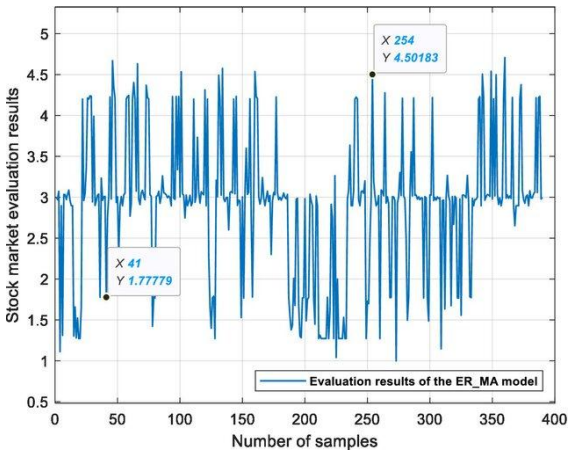
Once the model is trained, it is tested on a separate subset of data that the model has not seen before. This test dataset, also sourced from Kaggle, includes both safe and malicious URLs and serves as a benchmark to evaluate the model's performance. Testing involves feeding URLs into the model to check its classification accuracy. By analyzing results like accuracy, precision, and recall on this test data, researchers can assess how well the model generalizes and whether any adjustments are needed for improvement. Effective testing ensures the model's reliability in real-world applications.

f. Prediction:

In the final stage, the trained model uses its learned patterns to predict whether new URLs are "Safe" or "Malicious." Each URL input goes through the same feature extraction process and is then classified by the

RNN-autoencoder model. This prediction phase is critical, as the model needs to apply its training to unseen data in real-time. The ultimate goal is for the system to accurately flag malicious URLs quickly and reliably, providing proactive security measures against phishing attempts and minimizing risks for users.

V. RESULT AND DISCUSSION



Artificial neurons

A biological neuron primarily comprises a *nucleus* (or *soma*) in a *cell body* and *neurites* (*axons* and *dendrites*) (Wikipedia 2020b). The axons send output signals to other neurons, and the dendrites receive input signals from other neurons. The sending and receiving of signals take place at the *synapses*, where the sending (or *presynaptic*) neuron contacts the receiving (or *postsynaptic*) neuron. The synaptic junction can be at either the cell body or

Machine learning algorithms use evaluation metrics such as accuracy and precision. This is because we are trying to measure the algorithm's predictive ability. Although the same remains relevant for ML algorithms for financial market purposes, what is ultimately measured is the algorithm's performance with respect to returns or volatility. The works reviewed include various performance metrics that are commonly used to evaluate an algorithm's performance in the financial market context.

Recall that in Sect. 3.2.1 emphasized the importance of avoiding overfitting when backtesting. It is crucial to be consistent with backtesting different periods and to be able to demonstrate consistency across different financial evaluations of models and strategies. *Returns* represents the most common financial evaluation metric for obvious reasons. Namely, it measures the profitability of a model or strategy (Kenton 2020). It is commonly measured in terms of rate during a specific window of time, such as day, month, or year. It is also common to see returns annualized over various years, which is known as *Compound Annual Growth Rate (CAGR)*. When evaluating different models across different time windows, higher returns indicate a better model performance.

However, it is also important to consider *Volatility* because returns alone do not relay the full story regarding a model's performance. Volatility measures the variance or how much the price of an asset can increase or decrease within a given timeframe (Investopedia 2016). Similar to returns, it is common to report on daily, monthly, or yearly volatility. However, contrary to returns, lower volatility indicates a better model performance. The The Volatility Index (VIX), a real-time index from the Chicago Board Options Exchange (CBOE), is commonly used to estimate the volatility of the US financial market at any given point in time (Chow et al. 2021).

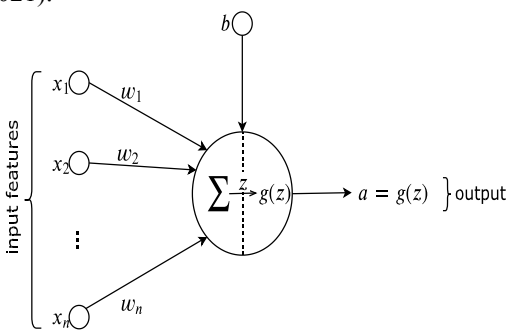


Fig. 4 Model of a typical neuron (Castro) 2006

As each input connects to the neuron, it is individually multiplied by the synaptic weight at each of the connections, which are aggregated in the *summing junction*. The summing junction adds the product of all of the weighted inputs with the neuron's bias value, i.e., $z = \sum + b$. The images essentially represent this. The *activation function* (also referred to as the *squashing function*) is represented as $g(z)$ and has the primary role of limiting the permissible value of the summation to some finite value. It determines a neuron's output relative to its net input, representing the summing junction's output. Thus, the neuron's consequent output, also known as the *activation* (a), becomes:

$$a\left(\sum_{j=1}^{n} w_j x_j + b\right) = g(z) = g$$

CONCLUSION

In conclusion, the proposed deep learning-based approach that integrates Recurrent Neural Networks (RNNs) and autoencoders significantly enhances the detection of phishing URLs by addressing the limitations of traditional blacklisting methods. By leveraging RNNs' ability to analyze sequential data, the system effectively identifies intricate patterns in URL structures, thereby improving predictive accuracy and real-time responsiveness to emerging threats. The incorporation of autoencoders further refines the process by performing efficient feature extraction, allowing the model to focus on the most relevant characteristics of the data. Together, these technologies provide a proactive and robust solution capable of adapting to the ever-evolving tactics employed by phishing attackers, ultimately offering enhanced protection for users against cyber threats. This innovative approach not only improves detection rates but also lays a solid foundation for future developments in cybersecurity, highlighting the critical need for continuous advancements to combat sophisticated phishing techniques.

FUTURE WORK

The scope for work Future directions include refining the model's capabilities to handle various data sources effectively, thus improving its adaptability and predictive power. Ultimately, this study aims to establish an advanced framework for stock market analysis, leveraging the latest in AI to provide more accurate and robust predictions that can support better-informed financial decision making in a constantly changing market landscape Research methodology

This research work set out to investigate applications of DL in the stock market context by answering three overarching research questions:

Question 1 What current research methods based on deep learning are used in the stock market context?

Question 2 Are the research methods consistent with real-world applications, i.e., have they been backtested?

Question 3 Is this research easily reproducible?

Although many research works have used stock market data with DL in some form, we quickly discovered that many are not easily applicable in practice due to how the research has been conducted. Although we retrieved over 10,000 works[1], by not being directly applicable, most of the experiments are not formulated to provide insight for financial purposes, with the most common formulation being as a traditional ML problem that assumes that it is sufficient to break the data into training and test sets. Recall that we categorized learning techniques by data availability in Sect. 3.1.2. When the complete data are available to train the algorithm, it is defined as *offline* or *batch* learning. When that is not the case, and it is necessary to process the data in smaller, sequential phases, as in streaming scenarios or due to changes in data characteristics, we categorize the learning technique as *online*. Although ML applications in the stock market context are better classified as online learning problems, surprisingly, very few research papers approach the problem accordingly, instead mostly approaching it as an offline learning problem, a flawed approach (de Prado 2018).

To apply this approach to financial ML research for the benefit of market practitioners, the provided insight must be consistent with established domain norms. One generally accepted approach to achieving this is backtesting the algorithm or strategy using historical data, preferably across different periods (Bergmeir and Benítez 2012; Institute 2020). Although Sect. 3.2.1 discussed backtesting, we should re-iterate that backtesting does not constitute a “silver bullet” or a method of evaluating results. However, it does assist evaluation of the performance of an algorithm across different periods. Financial time-series data are not IID, meaning the data distribution differs across different independent sets. This also means that there is no expectation that results across a particular period will produce similar performances in different periods, no matter the quality of the presented result. Meanwhile, the relevant performance evaluation criteria are those that are financially specific, as discussed in Sect. 3.2.2. To this end, we ensured that the papers reviewed provide some indication of consideration of backtesting. An ordinary reference sufficed, even if the backtested results are not presented.

We used Google Scholar (Google 2020) as the search engine to find papers matching our research criteria. The ability to search across different publications and the sophistication of the query syntax (Ahrefs 2020) was invaluable to this process. While we also conducted spot searches of different publications and websites to validate that nothing was missed by our chosen approach, the query results from Google

Scholar proved sufficient, notably even identifying articles that were missing from the results of direct searches on publication websites. We used the following query to conduct our searches:
"deep learning" AND "stock market" AND ("backtest" OR "back test" OR "backtest")
This query searches for publications including the phrases "deep learning",
"stock market", and any one of "backtest", "back test" or "back-test".
We observed these three different spellings of "backtest" in different publications, suggesting the importance of catching all of these alternatives. This produced 185 results [2], which include several irrelevant papers. For validation, we searched using Semantic Scholar (Scholar 2020), obtaining approximately the same number of journal and conference publications. We chose to proceed with Google Scholar because Semantic Scholar does not feature such algebraic query syntax, requiring that we search for the different combinations of "backtest" individually with the rest of the search query.
The search query construct provided us with the starting point for answering research questions (1) and (2). Then, we evaluated the relevance to the research objective of the 185 publications and considered how each study answered question (3). We objectively reviewed all query responses without forming an opinion on the rest of their experimental procedure with the rationale that addressing the basic concerns of a typical financial analyst represents a good starting point. Consequently, we identified only 35 papers as relevant to the research objective. Table 8 quantifies the papers reviewed by publication and year of publication. It is interesting to observe the non-linear change in the number of publications over the last 3 years as researchers have become more conscious of some of these considerations

REFERENCES

[1] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al (2016) Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp 265–283

[2] Aceto G, Ciuonzo D, Montieri A, Pescape A (2019) Mobile encrypted traffic classification using deep learning: experimental evaluation, lessons learned, and challenges. IEEE eTrans Netw Serv Manag 16(2):445–458

[3] Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160

[4] Adosoglou G, Lombardo G, Pardalos PM (2020) Neural network embeddings on corporate annual filings for portfolio selection. Expert Syst Appl. https:// doi. org/ 10. 1016/j. eswa. 2020. 114053

[5] Ahrefs (2020) Google Search Operators: the complete list (42 Advanced Operators). https:// ahrefs. com/ blog/ google- advan ced- search-opera tors/

[6] Amel-Zadeh A, Calliess JP, Kaiser D, Roberts S (2020) Machine learning-based financial statement. Analy-
sis. https:// doi. org/ 10. 2139/ ssrn. 35206 84

[7] Arimond A, Borth D, Hoepner AGF, Klawunn M, Weisheit S (2020) Neural Networks and Value at risk.
https:// doi. org/ 10. 2139/ ssrn. 35919 96,

[8] Arnott RD, Harvey CR, Markowitz H (2018) A backtesting protocol in the era of machine learning. SSRN Electron J. https:// doi. org/ 10. 2139/ ssrn. 32756 54

[9] Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213

[10] Boedihardjo H, Geng X, Lyons T, Yang D (2016) The signature of a rough path: uniqueness. Adv Math 293:720–737. https:// doi. org/ 10. 1016/j. aim. 2016. 02. 011

[11] Buehler H, Horvath B, Lyons T, Perez Arribas I, Wood B (2020). A data-driven market simulator for small data environments. https:// doi. org/ 10. 2139/ ssrn. 36324 31

[12] Castro LNd (2006) Fundamentals of natural computing (Chapman & Hall/Crc Computer and Information Sciences). Chapman & Hall/CRC, Boca Raton

[13] Chakole J, Kurhekar M (2020) Trend following deep Q-Learning strategy for stock trading. Expert Syst 37:e12514. https:// doi. org/ 10. 1111/ exsy. 12514

[14] Chalvatzis C, Hristu-Varsakelis D (2020) High-performance stock index trading via neural networks and trees. Appl Soft Comput 96:106567. https:// doi. org/ 10. 1016/j. asoc. 2020. 106567

[15] Chollet F et al (2015) Keras. https:// keras. io Chong E, Han C, Park FC (2017) Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. Expert Syst Appl 83:187–205. https:// doi. org/ 10. 1016/j. eswa. 2017. 04. 030

[16] Chow KV, Jiang W, Li J (2021) Does vix truly measure return volatility? In: Handbook of financial econometrics, mathematics, statistics, and machine learning. World Scientific, pp 1533–1559

[17] Christina Majaski (2020) Fundamentals. https:// www. inves toped ia. com/ terms/f/ funda menta ls. asp

[18] Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. In: EMNLP 2017— conference on empirical methods in natural language processing, proceedings, https:// doi. org/ 10. 18653/ v1/ d17- 1070, arXiv: 1705. 02364

de Prado ML (2018) Advances in financial machine learning, 1st edn. Wiley, New York

[19] Day MY, Lee CC (2016) Deep learning for financial sentiment analysis on finance news providers. In: Proceedings of the 2016 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2016, Institute of Electrical and Electronics Engineers Inc., pp 1127–1134, https:// doi. org/ 10. 1109/ ASONAM. 2016. 77523 81

[20] Derivative (2020) List of electronic trading protocols. https:// www. inves toped ia. com/ terms/d/ deriv ative. asp

[21] Easley D, López de Prado MM, O'Hara M (2012) The volume clock: insights into the high-frequency paradigm. J Portfolio Manag 39(1):19–29. https:// doi. org/ 10. 3905/ jpm. 2012. 39.1. 019

[22] Fabozzi FJ, De Prado ML (2018) Being honest in backtest reporting: a template for disclosing multiple tests. J Portfolio Manag 45(1):141–147. https:// doi. org/ 10. 3905/ jpm. 2018. 45.1. 141

[23] Fang Y, Chen J, Xue Z (2019) Research on quantitative investment strategies based on deep learning. Algorithms 12(2):35. https:// doi. org/ 10. 3390/ a1202 0035

[24] Ferguson R, Green A (2018) Deeply learning derivatives. arXiv: org/ abs/ 1809. 02233

[25] François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J (2018) An introduction to deep reinforcement learning. Found Trends Mach Learn 11(3–4):219–354. https:// doi. org/ 10. 1561/ 22000 00071

[26] Ganesh P, Rakheja P (2018) VLSTM: very long short-term memory networks for high-frequency trading. Papers arXiv: abs/ 1809. 01506, https:// ideas. repec. org/p/ arx/ papers/ 1809. 01506. html

[27] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge

[28] Google (2020) Google Scholar. https:// schol ar. google. ca/

[29] Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. Elsevier Inc., Amsterdam. https:// doi. org/ 10. 1016/ C2009- 0- 61819-5

[30] Hargrave M (2019) Sharpe ratio definition. https:// www. inves toped ia. com/ terms/s/ sharp eratio. asp

[31] Harper D (2016) An introduction to value at risk (VAR). Investopedia pp 1–7, http:// www. inves toped ia. com/ artic les/ 04/ 092904. asp

[32] Hayes A (2020) Maximum Drawdown (MDD) Definition. https:// www. inves toped ia. com/ terms/m/ maxim um- drawd own- mdd. asp

[33] Hinton G (2017) Boltzmann machines. In: Encyclopedia of machine learning and data mining. https:// doi. org/ 10. 1007/ 978-1- 4899- 7687-1_ 31