Expert Systems With Applications 197 (2022) 116659

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

Machine learning techniques and data for stock market forecasting: A literature review

Mahinda Mailagaha Kumbure [a,*], Christoph Lohrmann [a], Pasi Luukka [a], Jari Porras [b]

[a] School of Business and Management, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland
[b] School of Engineering Science, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland

ARTICLE INFO

Keywords:
Classification
Data mining
Financial market
Predictive performance
Regression
Stock market prediction

ABSTRACT

In this literature review, we investigate machine learning techniques that are applied for stock market prediction. A focus area in this literature review is the stock markets investigated in the literature as well as the types of variables used as input in the machine learning techniques used for predicting these markets. We examined 138 journal articles published between 2000 and 2019. The main contributions of this review are: (1) an extensive examination of the data, in particular, the markets and stock indices covered in the predictions, as well as the 2173 unique variables used for stock market predictions, including technical indicators, macroeconomic variables, and fundamental indicators, and (2) an in-depth review of the machine learning techniques and their variants deployed for the predictions. In addition, we provide a bibliometric analysis of these journal articles, highlighting the most influential works and articles.

1. Introduction

The average person's interest in the stock market has experienced an exponential growth over the last few decades (Badolia, 2016). Hence, it is unsurprising that assets worth billions of dollars are traded on stock exchanges every day (Hoseinzade & Haratizadeh, 2019), with investors acting on the market with the desire to achieve a profit over their investment horizon. If a market participant such as a private or institutional investor could forecast the behavior of the market accurately, this would enable them to consistently earn higher risk-adjusted returns than the market. This motivates the use of machine learning and computational intelligence methods to create accurate models for the prediction of the stock market. Indeed, a large number of published studies has attempted to forecast stock markets accurately by developing sophisticated forecasting models/systems (Sedighi et al., 2019; Song et al., 2019) and some studies reported that their models could generate profits (Armano et al., 2005; Atsalakis & Valavanis, 2009a; Weng et al., 2017). In general, stock market prediction is recognized as one of the most relevant but highly challenging tasks (Chen & Hao, 2017) in financial research. However, the ability of an investor to consistently achieve a higher risk-adjusted return than the market can be in violation of the so-called efficient market hypothesis.

Fama (1970) established the efficient market hypothesis (EMH), which assumes that the market price follows a random walk, i.e., future changes in the market's price cannot be predicted using existing information. In particular, the EMH distinguishes three forms of market efficiency: weak-form, semi-strong form, and strong-form efficiency (Atsalakis & Valavanis, 2009a; Fama, 1970).

Weak-form market efficiency assumes that information contained in past prices of a time series is already reflected in the current stock price and does not help in predicting future price movements (Fama, 1970). Therefore, in the weak form of EMH, technical analysis cannot outperform a buy-and-hold strategy in terms of expected return (Fama, 1965; Leigh et al., 2002). The second form of the EMH is the semi-strong market efficiency, which states that stock prices fully reflect all publicly available information (Fama, 1965). All publicly available information also includes information about past prices, which means that technical analysis may also not lead to consistently higher expected returns. Moreover, all publicly available information encompasses, for instance, fundamental information about economic conditions, political events, interest rates, and company-specific information, which is available to the public and affects stock prices (Wang et al., 2011). Notwithstanding, in the semi-strong form of market efficiency, publicly available information, including fundamental data, does not enable an investor to consistently outperform the market. This implies that active management that uses all publicly available information will not consistently yield higher risk-adjusted returns than passive management (e.g., buy-and-hold a stock market index). In contrast to the semi-strong form, the strong form of the EMH states that all information, including insider information, is reflected in stock prices. This precludes any investor, even those with insider information, from consistently achieving higher expected returns than the market (Fama, 1965, 1970; Leigh et al., 2002). Therefore, the EMH in its strongest form effectively

* Corresponding author.
E-mail addresses: mahinda.mailagaha.kumbure@lut.fi (M.M. Kumbure), christoph.lohrmann@lut.fi (C. Lohrmann), pasi.luukka@lut.fi (P. Luukka), jari.porras@lut.fi (J. Porras).

https://doi.org/10.1016/j.eswa.2022.116659
Received 28 April 2021; Received in revised form 9 December 2021; Accepted 5 February 2022
Available online 19 February 2022

states that returns in the stock market are unforecastable (Timmermann & Granger, 2004). The strong form of the EMH is rather extreme, and even Eugene (Fama, 1970) himself stated that one would not expect that insider information (e.g., of company officers) cannot be used to generate higher expected profits.

Over time, there has been an increasing number of challenges of the efficient market hypothesis and the fact that securities are priced rationally (Borovkova & Tsiamas, 2019; Daniel et al., 1998). There have been several market anomalies (Malkiel & Mullainathan, 2005) such as the overreaction of financial markets (Bondt & Thaler, 1985, 1990) and their underreaction, the existence of short-term momentum, long-term reversal, and the high volatility of asset prices (Daniel et al., 1998) which represent support against the efficient market hypothesis (especially in its weak-form). Some researchers discussed explanations for such anomalies that are in line with the EMH such as that over- and under-reactions happen randomly and are equally frequent (Fama, 1998) and the possibility of institutional investors being able to offset the anomalies created by less sophisticated investors (Shiller, 2003). However, there remained doubt that a model based on investor rationality can accommodate the observed anomalies (Daniel et al., 1998). This led to a shift towards models incorporating human psychology, leading to the emergence of behavioral finance (Bondt & Thaler, 1990; Shiller, 2003), which questions the perfect rationality of investors due to behavioral biases such as loss aversion, overreaction, and overreaction (Lo, 2004). One attempt to reconcile the EMH and behavioral finance was the proposal of the adaptive markets hypothesis (AMH), which acknowledges and explains the existence of anomalies in financial markets (Lo, 2004). For a detailed discussion of the evolution of the efficient market hypothesis, see, Lim and Brooks (2011).

Because of the fact that market anomalies may exist, it is unsurprising that a large number of market participants use information of past market prices, company-specific information such as past earnings and profits, as well as other factors to build their expectation about future stock prices (Patel & Marwala, 2006). Moreover, investors often expect that short-term returns continue since past returns may reflect the investor sentiment (Bustos et al., 2011). Given such expectations and the existence of market anomalies, it seems plausible to use information about the past to forecast the stock market.

In stock market prediction studies, in general, two well-known analytical approaches, fundamental analysis and technical analysis, are deployed (Lam, 2004; Lohrmann & Luukka, 2019; Sedighi et al., 2019). Fundamental analysis focuses on fundamental information. In case a company's stock price or return is forecasted, fundamental information is, for instance, a company's revenues and expenses, yearly growth rate, position in the market, and other information contained in financial statements or reports (Bodie et al., 2009; Murphy, 1999). In case a stock index, which represents a set of numerous company stocks, is forecasted, the same kind of information as well as information on the market environment can be deployed, including national productivity, trade, exchange rates, or interest rates, which will likely have an impact on the operation of the companies contained in the stock index. Conversely, technical analysis is the study of historical stock price and volume data to predict the movements of the stock price (Lohrmann & Luukka, 2019; Turner, 2007; Wei et al., 2011).

Most previous studies have applied statistical time-series methodologies based on historical data to forecast stock prices and returns (Efendi et al., 2018). Among them, the auto-regressive conditional heteroscedasticity (ARCH) model, auto-regressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) models, moving average, Kalman filtering, and exponential smoothing are the most popular techniques (Chen & Chen, 2015; Wei et al., 2011; Yeh et al., 2011). Later, with the introduction of artificial intelligence (AI) and soft computing, these techniques have received increased attention within stock market prediction studies. Unlike traditional time series methods, these techniques can handle the nonlinear, chaotic, noisy, and complex data of the stock market, leading to more effective predictions (Chen & Hao, 2017). Consequently, these methods represent innovative and advantageous alternatives, which makes them attractive to be adopted by researchers for financial market forecasting.

In addition to the forecasting methodology, data are an essential component in stock market forecasting and play a vital role in the prediction process. In particular, all the analysis methods discussed above deploy selected data (variables) for a specific time period for model building. Data related to the stock market often contain time series data in various forms, such as stock index prices, returns, volatility, and interest rates (Enke & Thawornwong, 2005). To our knowledge, there has been very limited research on the variables included during the model building process for stock market prediction. In this paper, we focus on a systematic review of the literature in financial market forecasting with a focus on the data included in the studies and the statistical and machine learning methods used.

Thus, this work aims to broaden the current knowledge in stock market predictions through a systematic literature study. From this perspective, we first established the research questions: (1) What kind of data (variables and type of variables as well as time horizons) are used? and (2) what machine learning and AI techniques were applied in stock market prediction studies? In addition to these primary research tasks, this study includes several other contributions: a discussion on state-of-the-art machine learning-based forecasting models over the last two decades, a bibliometric analysis of the selected studies with the most significant research contexts (e.g., keywords and citation performances), and validation methods, with respect to the selected literature.

Across this research, to present a comprehensive review, we followed the guidelines in Snyder (2019) and methodologies from Ahmad et al. (2018), Ambreen et al. (2018), Kitchenham and Brereton (2013). The search result was 138 articles in total published between 2000 to 2019. In addition, we used predefined exclusion and inclusion criteria.

We organized the rest of the paper as follows: Section 2 briefly presents a general overview of the data (and their properties) used, as well as machine learning methods applied in stock market prediction. In Section 3, related work is presented. Section 4 briefly describes the methodology followed to conduct this literature review. The results of the review study are displayed and discussed in Section 5. Eventually, Section 6 outlines the conclusions made according to the results and evaluations of the review.

2. General overview of the data and machine learning techniques used in stock market prediction

2.1. Data sources

In the literature, variables such as technical indicators, financial variables, and macro-economic variables have been considered the most influential ones affecting stock price movements (Tsai & Hsiao, 2010). However, such studies incorporated different sets of variables as input data for their prediction models, because there is no general consensus about all variables that are relevant for stock market forecasting. In this study, we classified all the variables in the selected literature that were used as input data into four main categories and several subcategories, as illustrated in Fig. 1.

Technical indicators are widely applied input variables in most prediction studies, since they play a vital role in buy and sell signals for stocks (Sedighi et al., 2019). In general, we categorized all technical indicators into "basic technical indicators" and "other technical indicators". A description and analysis of each type and sub-type are provided in Section 5.3, which discusses the variables and stock market indices covered in the literature. Several studies (Chun & Park, 2005; Hadavandi et al., 2010; Hassan et al., 2007) have used basic technical indicators as the input data for forecasting stock markets successfully, while many studies (Chang & Wu, 2015; Fadlalla & Amani, 2014;

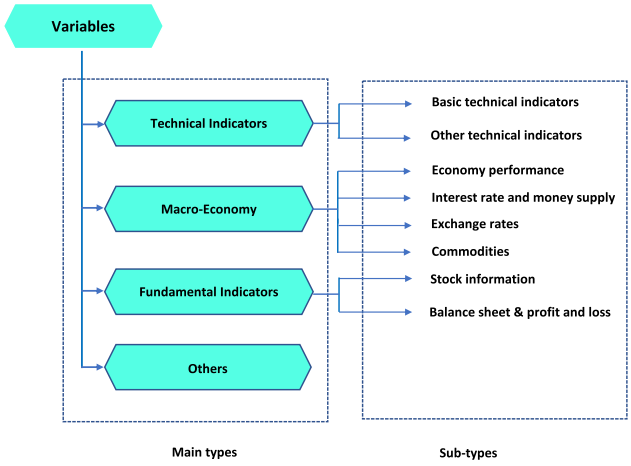

Fig. 1. Variable categories for stock price and return predictions.

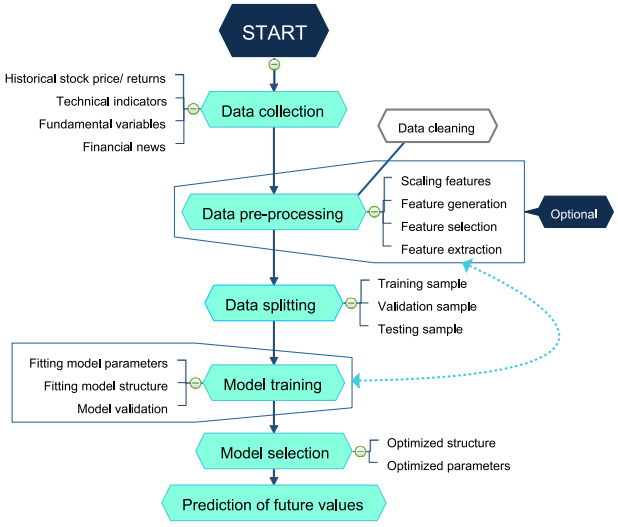

Fig. 2. Workflow of a stock market prediction model with supervised learning.

Khaidem et al., 2016; Laboissiere et al., 2015; Son et al., 2012; Zhang et al., 2018a) adopted various other technical indicators.

Furthermore, macro-economic variables have also received considerable attention in stock market studies (Enke et al., 2011; Leung et al., 2000; Tsai et al., 2011; Zhong & Enke, 2017a, 2017b). We defined four subcategories for this category, namely "exchange rates", "commodities", "economic performance", and "interest rate & money supply".

Fundamental indicators related variables have been another variable type of interest among stock market prediction studies (Barak et al., 2017; Barak & Modarres, 2015; Lai et al., 2009; Lam, 2004; Olson & Mossman, 2003; Tsai et al., 2011). Such variables are also discussed withing two sub-categories: (1) "stock information variables", which are based on or related to a particular company stock traded on a public stock exchange, (2) "balance sheet & profit and loss statement variables", which refer to financial reporting variables.

Lastly, the rest of the variables are classified as "other variables" without any sub-divisions. For example, some studies predicted a specific stock market deploying price data of other indices (Niaki & Hoseinzade, 2013; Zhong & Enke, 2017b, 2019) or variables extracted from financial news (Chen et al., 2017; Lien Minh et al., 2018; Shynkevich et al., 2016), ad hoc announcements (Feuerriegel & Gordon, 2018), email data (Zhou et al., 2018), and tweets (Shi et al., 2019).

2.2. Machine learning

The philosophy behind machine learning is to extract knowledge from data (Kubat, 2017, p. 1). Supervised learning is the most widely used machine learning techniques in stock market prediction. Fig. 2 illustrates a general workflow of a supervised learning-based approach applied for stock market prediction.

The process starts with choosing time-series data (e.g., stock price and/or return) and/or relevant information (e.g., financial news) from a specific time period. If the task is a classification problem, the target class is either known or needs to be predicted.

First, the corresponding data require pre-processing, which initially includes the cleaning and removal of incomplete or obviously irrelevant data (such as identifiers). Next, technical indicators can be calculated based on the underlying time-series data, such as close price information. Once the cleaned data including technical indicators are obtained, the data are pre-processed further through scaling and dimensionality reductions (i.e., feature selection, feature extraction, and feature generation) to obtain relevant variables and to filter out irrelevant ones. Using pre-processed data often leads to effective predictions (Chen et al., 2019). In Fig. 2, however, this step is shown as optional because it usually depends on the selected domain and is also up to the author's choice. Once the input data are ready, the task is to select an existing or novel machine learning technique to predict the target variable. For this purpose, input data are usually divided into training data (to train the model with certain parameters and structure of the model), validation data (to evaluate the performance of all trained models and select the best model structure and parameters), and test data (to evaluate the generalization performance of the final model on observations that it has not encountered before during training and validation). Since feature selection in its simplest form (i.e., filter methods) can be used independently of the learning algorithm, it is listed as a data pre-processing step. However, it may be connected to model training by using the learning algorithm's performance to perform feature selection (wrapper method) or may be integrated in the model construction itself (embedded method) (Guyon & Elisseeff, 2003; Lohrmann et al., 2018). To account for this potential link between feature selection and model training, the training step is connected with the pre-processing step via a dashed line in the flow chart. Lastly, prediction is performed using the trained classification model or in regression using the trained regression model.

In the literature, several variants of machine learning techniques have been developed for application to stock market predictions. Among them, artificial neural networks (ANNs) (Nermend & Alsakaa, 2017; O'Connor & Madden, 2006), support vector machines (SVMs) (Cao & Tay, 2001; Huang et al., 2005), and their variants (Ebrahimpour et al., 2011; Enke & Thawornwong, 2005; Pan et al., 2017) are the most frequently applied methods since they have shown promising results in prediction. With the introduction of intelligent systems in fuzzy theory (Zadeh, 1965) that deal with uncertainty in data, a considerable amount of literature has been published discussing stock market forecasting models based on fuzzy theory. These models include fuzzy time-series (Cagcag Yolcu & Alpaslan, 2018; Chu et al., 2009), adaptive network-based fuzzy inference systems (Wei et al., 2011), Takagi–Sugeno–Kang (TSK) type fuzzy systems (Chang & Fan, 2008; Chang & Liu, 2008) and other variants (Lai et al., 2009; Pal & Kar, 2019).

In addition, according to the literature, other machine learning techniques such as random forests (Khaidem et al., 2016; Lohrmann & Luukka, 2019), decision trees (Tsai & Hsiao, 2010), k-nearest neighbor (KNN) classifiers (Zhang et al., 2017), and Bayesian networks (Malagrino et al., 2018) have been often applied as well. However, because most of the mentioned techniques have their own merits and limitations, some researchers tended to enhance the forecasting accuracy

of those methods. On account of this, combinations of several methods such as KNN + SVM (Cao et al., 2019; Chen & Hao, 2017), ANN + SVM (Lu & Wu, 2011; Weng et al., 2017), and others have been investigated for predicting stock prices or returns. In addition, the forecast accuracy of the techniques mentioned above have been improved using feature selection methods (Barak & Modarres, 2015; Zhang et al., 2014), feature extraction methods such as principal component analysis (PCA) (Chen & Hao, 2018; Wang & Wang, 2015), evolutionary algorithms such as genetic algorithms (GA) (Ye et al., 2016), Wavelet transforms (Chiang et al., 2016), and particle swarm optimizations (Chai et al., 2015), to name a few. Moreover, in contrast to the previously discussed supervised learning techniques, the ability of clustering as an unsupervised method was also examined for forecasting stock prices (e.g., Vilela et al., 2019).

Recent developments in the field of stock market prediction has led to renewed interest in deep learning methods. Deep learning is a branch of machine learning (Lien Minh et al., 2018) that can identify hidden nonlinear relationships and extract relevant features from complex and noisy data without relying on human expertise and economical assumptions (Chong et al., 2017). Accordingly, deep neural networks (Singh & Srivastava, 2017), convolutional neural networks (Cao & Wang, 2019; Gunduz et al., 2017), and long short-term memory networks (LSTM) (Fischer & Krauss, 2018) have also been applied comprehensively in stock market price and return forecasting.

3. Related work

In recent years, there has been a growing interest in review studies conducted in many different fields, including finance and business (Henrique et al., 2019), computer science (Ahmad et al., 2018), social science (Ahmed et al., 2019), medicine (Klemm et al., 2003), and engineering (Ambreen et al., 2018), to name a few. Concerning market forecasting models in finance, literature review studies are comparably rare and only few examples exist. One survey of stock market forecasting techniques with more than 100 relevant articles published until 2009 was presented by Atsalakis and Valavanis (2009b). This survey is primarily concerned with studies that apply neural networks and neuro-fuzzy systems. Similarly, Rather et al. (2017) surveyed portfolio selection and stock market prediction models across studies that used traditional mathematical models as well as AI. Farias Nazário et al. (2017) presented a review on technical analysis indicators as forecasting methods for stock markets. They provided not only an overview of technical analysis through the existing literature but also recommendations and directions for new studies in this subject.

A recent review on machine learning techniques in financial market prediction can be found in (Henrique et al., 2019). This review provides a bibliographic analysis over 57 selected studies while commenting on the most cited articles, authors, and co-citation frequencies. Similarly, Gandhmal and Kumar (2019) performed a systematic analysis and review study using over 50 research articles related to stock market prediction. They essentially classified the selected studies with respect to the applied prediction methods with a detailed discussion of them, the year published, the performance metrics, and the software tools. Shah et al. (2019) also provided a concise review and taxonomy of stock market forecasting models.

Some recent reviews attempted to cover a broader range of studies in the field. For example, Bustos and Pomares-Quimbaya (2020) presented a systematic review of the prediction methods used in the stock market, covering 52 studies published from 2014 to 2018. This review focused on different types of machine learning techniques, including deep learning, text mining, and ensemble techniques. Moreover, a study by Jiang (2021) surveyed deep learning models applied for stock market predictions in the last three years. It also provided a brief overview of the data used and the data processing methods used and pointed out some future research directions based on existing research. By examining 30 journal and conference articles, Kumar et al. (2021) provide a complete overview of various aspects adopted in stock market prediction studies, including machine learning algorithms, performance measures, datasets, and journals.

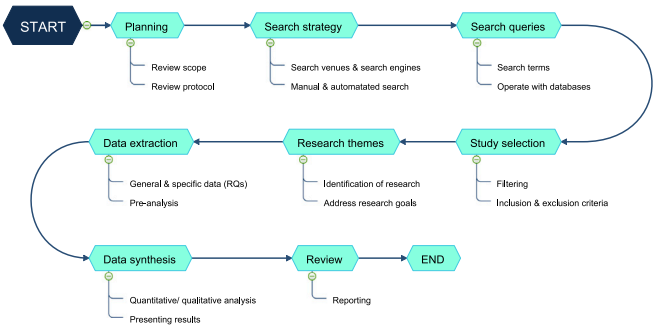

Fig. 3. The literature review process.

However, most of these reviews did not analyze the forecasting approaches themselves in detail and only provide an overview of them in the form of tabled information. Rather than considering an in-depth review of the applied methods, the deployed input variables, and the data used in the studies, they mostly focus on a bibliographic analysis together with a classification of the research elements of the studies. According to Weng et al. (2017), the main components in stock market prediction are the tracking of relevant information about data and predictor variables, and the selection of AI techniques that are effective for prediction and analysis. For this purpose, and due to the found research gap, in our study, we provide an in-depth review of the machine learning models used in forecasting. In addition to the main methods and their variants, we also discuss other sub-techniques used in conjunction with machine learning to improve the predictive power of these methods. Furthermore, we emphasize the specific variables and variable types that are used to conduct the forecast as well as an analysis of the stock markets involved.

4. Research methodology

4.1. Planning

The scope of this literature review was defined based on our objectives and research questions. We concentrated on research works in the time period from 2000 to 2019 and limited ourselves to articles using machine learning methods to predict stock markets. To conduct a systematic literature review, defining the review protocol (i.e., the complete plan) is essential for getting primary studies and reducing the bias (e.g., publication bias) in our research. Therefore, in this review study, we applied the review protocol introduced by Kitchenham (2004). It includes steps concerning the planning and review phase of a systematic literature review. In conjunction with that, the plan was created for our systematic literature review study, and it was implemented with the process presented in Fig. 3.

The steps of this review process are discussed in detail within the next subsections.

4.2. Search strategy

The purpose of the search strategy was to find an appropriate and effective set of studies to answer the research questions. The search process of this review study consisted of two stages for searching the literature. In the first stage, we performed a “manual search” by selecting the pilot set of papers through defined search venues. Then, using this initial set of articles, snowballing was conducted following the strategy introduced by Wohlin (2014). In the next stage, “automated search” was conducted using the technique proposed in Kitchenham

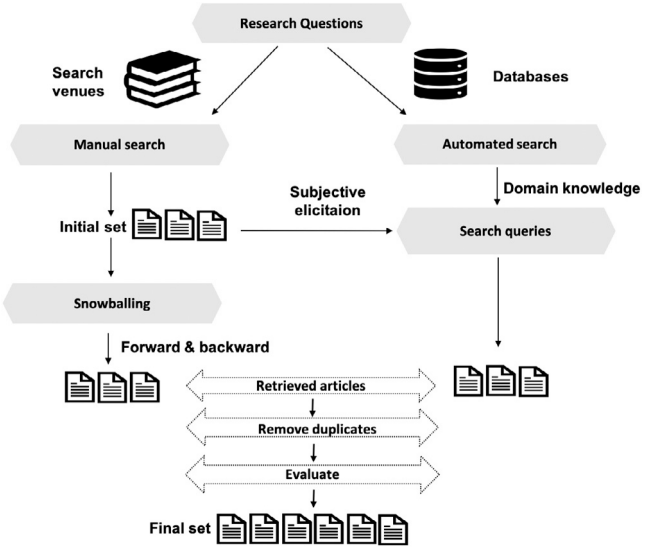

Fig. 4. The mechanism involved in the search process.

and Brereton (2013) to find more relevant articles. In this strategy, we first defined the scientific databases to be used in the search and then a list of primary search terms to be applies in the search queries for each database. In summary, the mechanism used in the search process is illustrated in Fig. 4.

4.2.1. Manual search

Our initial set of primary studies included nine journal articles (Atsalakis & Valavanis, 2009a; Chang & Liu, 2008; Chong et al., 2017; Enke & Thawornwong, 2005; Huang et al., 2005; Kara et al., 2011; Lohrmann & Luukka, 2019; Patel et al., 2015; Tsai et al., 2011). We found these articles in the manual search for literature on stock market predictions, and we decided to use them as the initial set after examining the full text of each of these articles. Moreover, this set of articles was part of the final analysis. Next, backward and forward snowballing were performed using the initial set, and relevant articles were selected within several rounds. In backward snowballing, the first step was to scan through the reference list and include the relevant studies that fulfill the inclusion and exclusion criteria. In this step, the titles of the articles were examined. Next, the articles that were found earlier through backward snowballing or from the initial set were removed from this list. After repeated articles were excluded, abstracts (and other parts if needed) were read to find the relevant articles in the final set.

In forward snowballing, the search was based on finding articles that cited a single article (or multiple articles) of the initial set. This citation analysis for each research article was performed using Web of Science. The filtering options available in this database provided an additional advantage to find only relevant studies based on the inclusion and exclusion criteria. Once the set of relevant articles was identified, the duplicate articles that were also contained in any previous set were excluded. Next, the abstract in each article was studied, and if the abstract was not deemed sufficient to judge the relevance of the article for the potential inclusion in our study, other parts of the article were examined to select the final set of articles.

4.2.2. Automated search

In the search process, our objective was to find a broad set of relevant studies. We considered having a comprehensive set of relevant studies as an essential factor in making our study distinguishable from existing review studies, helping in minimizing the bias, and delivering more reliable evaluations of results. However, we managed to get only 103 related articles in total in the final set with manual search. As we aimed to use a broad sample of relevant studies, we next extended the search process using automated search to get more articles into the final sample. We also realized that it is possible to miss more relevant articles during manual search since the area of stock market predictions is more widespread.

Accordingly, we applied an automated search strategy to extend the literature search and find more relevant articles that might have been missed during the manual search. We used four databases, "Scopus", "IEEE *Xplore*", "Web of Science", and "Science Direct" across the automated search to find relevant articles. This process was started by defining search queries, and then relevant studies were identified after using the inclusion and exclusion criteria. Next, duplicate articles to those identified either within the manual search or the automated search were removed. After evaluating the abstracts (and other parts if needed) of the remaining articles, we included the articles into the final set.

Table 1
Search query for each database.

Database	Search query
Scopus	(("Prediction" OR "Forecasting") AND ("Stock price" OR "Stock return") AND ("AI" OR "Machine Learning"))
IEEE *Xplore*	(('Prediction' OR 'Forecasting') AND (('Stock' AND 'Price') OR ('Stock' AND 'Return')) AND (('Machine' AND 'Learning') OR ('AI'))
Web of Science	("Prediction" OR "Forecasting") AND ("Stock price" OR "Stock return") AND ("Machine learning" OR "AI")
Science Direct	(Prediction OR Forecasting) AND ("Stock price" OR "Stock return") AND (AI OR Machine Learning)

4.3. Search queries

The search queries were defined as they emphasize the purpose of our research. To define them, we used our experience and domain knowledge as well as the subjective elicitation from the titles, keywords, and abstracts of the articles used in the manual search. The terms "Prediction" and "Forecasting" were used separately for each data source, since we noticed that relevant articles had used at least one of them in their definitions. Prediction or forecasting was expected for either "Stock price" or "Stock return" by using them in the same search query. The term "Machine Learning" was used as the closed term of the defined query since the expected articles are supposed to be based on the forecasting models that applied machine learning methods. However, it became apparent that using "AI" separately with "Machine Learning" resulted in a larger set of relevant articles. Table 1 summarizes the queries used in each data source in the automated search.

4.4. Inclusion and exclusion criteria, and study selection

Inclusion and exclusion criteria were applied to select the set of the most relevant articles for our study. All of those options are presented in Table 2. We decided to exclude an article if it was out of the year range between 2000 and 2019, a conference paper, an unpublished article, a report publication, a thesis work, or a book publication. We wanted to access full texts of the article, so we considered the full-text availability of it. Articles that were written in other languages except English were also excluded.

No articles prior to 2000 were included because (i) these were published more than 20 years ago and, thus, cannot be considered contemporary, (ii) there were very few relevant publications on this topic around 2000 and prior to it (see Fig. 6), and (iii) the search for 2000 to 2019 already yielded a suitable number of articles to include in our study.

Table 2
Applied filtering options during the search process.

Filtering options	Specification
Published year	2000–2019
Document type	Journal only
Language	English
Published stage	Final
Text availability	Full texts available

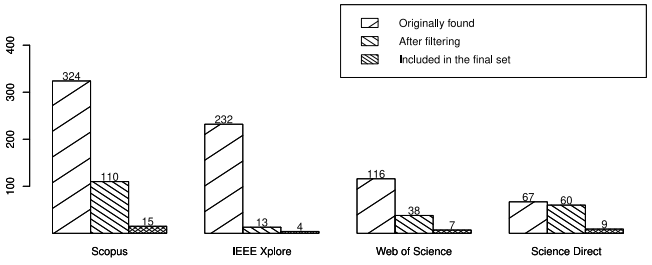

Fig. 5. Databases and the corresponding number of articles in each stage.

Table 3 displays the results obtained from the steps utilized in snowballing. As the table shows, backward snowballing resulted in a total of 427 references, most of them being quite old (published before 2000). Thus, after the removal of irrelevant papers, 91 articles were selected for further investigation. These selected articles were compared with previously identified articles to remove the duplicates. Next, the retained articles were read and examined, which led to the inclusion of a total of 32 articles into the final set of articles. With forward snowballing, the initial result of the search was 1191 articles, but the inclusion and exclusion criteria led to the identification of 238 relevant articles. After excluding duplicate articles and studying the content of the rest of the selected articles, a total of 62 articles were selected for the final collection. In summary, 103 (9 + 32 + 62) articles were selected for the final set from the manual search within several rounds.

We also applied the automated search in the search process since we wanted to include more articles in this review and decided the number of articles from the manual search was not sufficient. Concerning the inclusion and exclusion criteria, we used the available filtering options for each data source during the search process to select the most relevant studies during the automated search.

The search results in each stage within the selected database are visually presented in Fig. 5. A total of 739 articles (Scopus: 324, IEEE *Xplore*: 232, Web of Science: 116, and Science Direct: 67) were the first result in the database search. Next, applying the filtering criteria resulted in a total of 221 articles (Scopus: 110, IEEE *Xplore*: 13, Web of Science: 38, and Science Direct: 60). It is noteworthy that we obtained a low number of relevant articles from IEEE *Xplore*. Most of the articles resulting from search queries within this source were conference papers. As the final result, a total of 35 (Scopus: 15, IEEE *Xplore*: 4, Web of Science: 7, and Science Direct: 9) articles were selected to be included in the final set after removing duplicates and with a brief analysis of the content of each article. After merging the final set of articles from manual and automated search, a total of 138 articles were selected to be used in the data extraction process. Both the first and second author have been involved in this search strategy.

4.5. Research themes

The primary goals of this research, as previously noted, were to explore the characteristics of the data used and to identify the machine learning approaches and their variants applied to forecast the stock market. Moreover, through a short bibliographic analysis, this study also aimed to gain an understanding of the current state of the published researches focused on financial market forecasting. Furthermore, we also addressed the data orientation process, including pre-processing methods, evaluation metrics, and performances. Accordingly, we defined four research themes based on the research objectives, and they are presented in Table 4.

4.6. Data extraction

Before starting data extraction, we determined which types of information/attributes need to be summarized to achieve the objectives of this literature study. The relevant information was then identified and extracted through the defined attributes via reading the entire text of each article. That information was then stored in a Microsoft Excel sheet for analysis. We primarily collected data through the attributes listed in Table 5. In the table, the "Study" column refers to the IDs of the research themes mentioned in Table 4, and "Subject" refers to the category of the group of several attributes. In addition, attribute names, together with a short description, are presented in the last two columns. It is noteworthy that some attributes were divided further (using more columns) to include all relevant information. For example, if there were more than one model in a study, then it was reported as Model 1, Model 2, etc. Other relevant attributes were also changed as desired. Including all of these attributes, there were 79 unique attributes in total.

Besides the information presented in the data extraction step, as described in Table 5, there was another part of the data extraction that was specifically relevant to RT_2. Most studies on stock market prediction used a specific number of features from the data as input variables. As we already know, there are various types of variables under the categories of technical indicators, macro-economic variables, fundamental indicator variables, and others. Therefore, we collected information on variables included in the data that were deployed in the literature for stock market predictions. We extracted this information by assigning "1" and "0" to each specific feature (predefined, as specific as possible, e.g., "simple moving average (5 periods)") concerning each study so that if the particular feature was used, then a value of 1 was placed and otherwise 0. If a feature was used in a study in multiple configurations but no information was provided to identify and differentiate these modifications, then the assigned value was the number of configurations instead. The purpose of collecting this information was to present a comprehensive analysis of features and their patterns used for forecasting stock market values.

4.7. Data synthesis

Data synthesis is aimed at analyzing and summarizing information observed from the selected articles to answer the research questions posed in this study. We used a thematic synthesis (Braun, 2006) based on qualitative data, as well as descriptive analysis to find answers to the research questions based on quantitative data. Here, we performed a qualitative and quantitative analysis of the evidence from the data.

5. Results of the literature review

In this section, we present the results obtained from the analysis of the review connected with the research questions defined earlier. The analysis was performed along with the data extracted from the 138 selected primary studies. Moreover, the reviewed results of (1) bibliographic information, (2) data and their properties, and (3) applied and developed machine learning methods are presented and discussed under three subtopics.

Table 3
Search results in the snowballing process.

Articles in the initial set	Backward snowballing			Forward snowballing		
	References	Relevant articles	Final set	Cited studies	Relevant articles	Final set
Lohrmann and Luukka (2019)	64	17	15	1	0	0
Kara et al. (2011)	43	14	9	184	48	16
Patel et al. (2015)	26	10	0	136	39	7
Enke and Thawornwong (2005)	46	1	0	165	23	11
Huang et al. (2005)	30	5	1	325	50	8
Tsai et al. (2011)	33	6	1	54	9	0
Atsalakis and Valavanis (2009a)	84	8	0	120	28	5
Chong et al. (2017)	77	24	6	62	13	6
Chang and Liu (2008)	24	6	0	144	28	9
Total	427	91	32	1191	238	62

Table 4
Research themes.

ID	Research theme	Objectives
RT_1	Bibliographic information	To find the yearly distribution of the selected articles To find the most relevant journal in the related area To find the distribution of keywords To find the most relevant studies in the related area
RT_2	Data and data orientations	To explore the characteristics of data used To find most frequently used variables in the data To find the data pre-processing methods applied
RT_3	Machine learning/AI methods	To discover the most common machine learning and AI techniques used To find the sub-techniques that were deployed with the main approaches To find recent trends with techniques in the related area To obtain the state-of-the art relevant models
RT_4	Performance metrics	To find the most frequently used performance metrics To find the performance of each machine learning model applied

RT_i: Research theme i, for $i = 1, 2, 3, 4$.

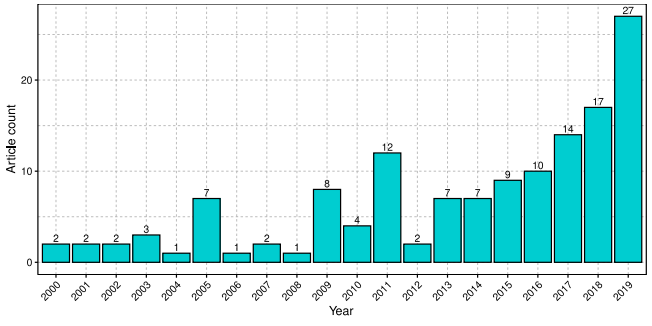

Fig. 6. Number of included articles by year.

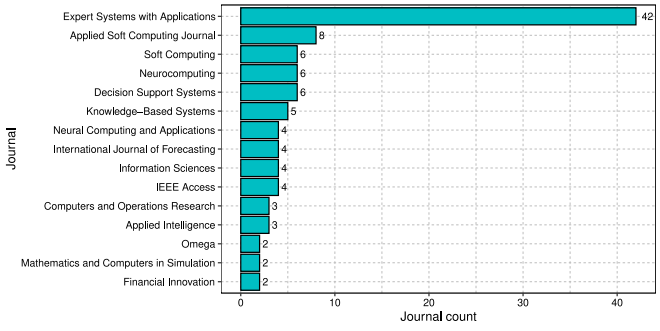

Fig. 7. Article counts by most relevant journals.

5.1. A review of bibliographic information

Each of the following subsections present one of the highlighted cases of the bibliographic analysis that we designed for answering questions in the first phase of our research. Accordingly, we analyze the selected articles by year, journal, frequently used keywords, and most cited articles. First, we present an overview of the selected studies used in this research.

5.1.1. Overview of extracted data

The characteristics of the selected studies in this research are presented in Table 6. This review study used 138 primary articles in total, which were published in 52 peer-reviewed journals in the last two decades.

5.1.2. Publications by year

The number of articles published by year is presented in Fig. 6. Most studies found in this research were up-to-date as they were recently published. More than 50% of articles in the data were published between 2015 and 2019, while the remaining articles are from the 15-years period prior to 2015. In general, there appears to be an exponential growth in the number of studies that focused on stock market predictions using machine learning.

5.1.3. Publications by journal

Fig. 7 shows the most relevant journals of the selected studies. There are 52 unique journals in which the selected articles were published. In the figure, we present only those 16 journals that contained more than one article from our set of 138 scientific articles.

In this classification, it is visible that the journal *Expert Systems with Applications* was the top journal in which most articles related to stock market forecasting were published. It accounts for approximately 31% of the papers selected. Other journals such as Applied *Soft Computing, Soft Computing, Neurocomputing, Decision Support Systems*, and *Knowledge-Base Systems* account together for approximately 22%. These results indicate which journals were the main publishing sources of research on stock market predictions using machine learning.

Table 5
Attributes in data extraction.

Study	Subject	Data attribute name	Description
RT_1	Bibliographic information	Authors	The authors of the article
		Year	Year of publication
		Journal	Where the paper is published
		Keywords	The keywords listed in the article
		Title	The title of the article
		Objective	The purpose the research
		Findings	A short description of the key findings of the research
RT_3	Theoretical concepts and experimental procedure	Approach	Main approach used in the methodology (e.g., regression)
		Model (Category)	Novel AI/machine learning model OR main method
		Model (Sub-category)	Other models used in the model (e.g., PCA, Clustering)
		Specifications	Final model specifications/ parameters
		Programming Language	The software/language used in the study
		Benchmark	Existing methods used for comparison purposes
		Training Iterations	Number of iterations in the training model
		Prediction Type	Stock price or return or movement of the price or return
RT_2	Data descriptions	Stock Index	Name of the stock index
		Stock Trading Volume	Number of shares that changed hands during a given day
		Individual Stocks	Number of individual stocks
		Individual Stocks Name	Names of the individual stocks (of companies)
		Frequency	Time window of the used data (e.g., daily/weekly/monthly)
		Observations	Number of observations in the data
		Beginning Period	The starting date of the data
		End Period	The end date of the data
		Stock Markets	Number of stock markets in the used data sets
RT_2	Input features in the data	Technical Indicators	Number of technical indicators in the data
		Economic /Fundamental	Number of economic/ fundamental variables in the data
		Other Features	Number of other variables in the data
RT_3	Data pre-processing	Feature Selection	Is feature selection applied? Yes or No
		Feature Extraction	Is feature extraction applied? Yes or No
		Feature Construction	Is feature generated? Yes or No
		Feature Normalization	Is feature normalized? Yes or No
		Analysis of Features	Is an analysis of features included? Yes or No
		Number of Classes	Number of categorizations of the prediction
		Validation Method	Validation method in the developing process
		Data Split	The ratio of the data split
RT_3	Analysis of trading strategies	Trading simulation	Is trading simulation applied? Yes or No
		Best Strategy (Daily)	Is best daily strategy/benchmark found? Yes or No
		Best Strategy (Weekly)	Is the best weekly strategy/benchmark found? Yes or No
		Best Strategy (Monthly)	Is the best monthly strategy/benchmark found? Yes or No
RT_4	Evaluation metrics and performances	Performance Measure	Name of the performance measures (PMs)
		Value of PMs	The value of the performance measure, e.g., accuracy
RT_3	Market efficiency	EMH investigated	Is the EMH investigated? Yes or No
		Transaction Cost	Is transaction cost considered? Yes or No
		Tax	Is tax considered? Yes or No
		Slippage	Is slippage considered? Yes or No
		Bid–Ask Spread	Is bid–ask spread considered? Yes or No

Table 6
An overview of the studies in the data collection.

Description	Result
Number of articles	138
Number of journals	52
Period	2000–2019
Number of authors	346
Number of author keywords	351
Single-authored documents	11
Multi-authored documents	127

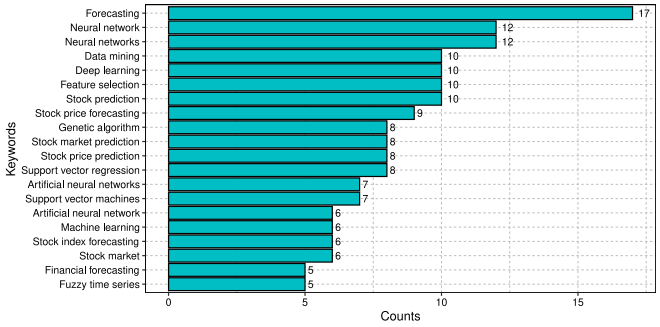

Fig. 8. Article counts by most frequently used author keywords.

5.1.4. Variation of keywords

The list of keywords is one of the most important settings of a research article, making the article searchable and guaranteeing that it is easily accessible by other researchers. Hence, the 20 most frequently used author keywords are shown with their respective counts in Fig. 8. It can be easily seen from the figure that the keywords *forecasting*, *neural network*, *data mining*, and *stock price prediction/ forecasting* are ranked highest on the list of keywords. It is also visible that the top keywords include synonyms and singular and plural forms of some keywords (e.g., "neural network" and "neural networks"), revealing that taking these distinctions into account might be worthwhile in the search for publications.

In addition to presenting the main author keywords, we also investigated the dynamics of these keywords year by year. Fig. 8 highlights the top 20 keywords with their respective overall count. More details about the keywords dynamics and citation performance can be found in Appendix.

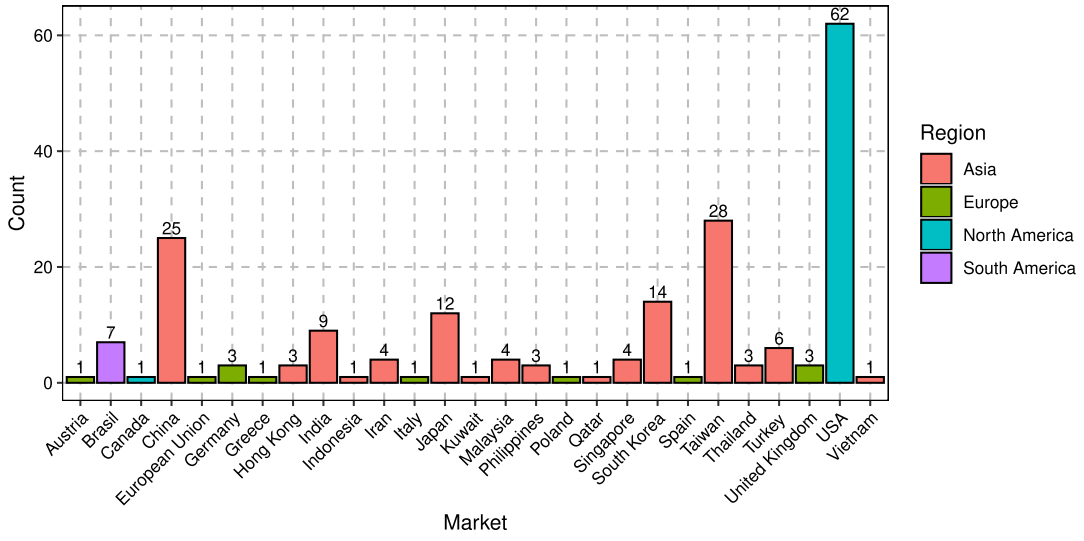

Fig. 9. Frequency by market.

5.2. Markets, indices, and stocks

The focus of this paper is on the prediction of stock markets, i.e., stocks traded on stock exchanges and the indices that represent the performance of a set of specific stocks. An example of a stock is "Apple Inc.", which is a well-known "Information Technology" company in the USA, and an example of a stock index is the S&P 500 index, which tracks the performance of the 500 largest companies in the USA, including "Apple Inc". The research articles considered in this study contain 26 different markets as well as one index for the European Union as a whole. Fig. 9 illustrates the number of publications in which the stock market indices and/ or specific stocks of a market were discussed.[1]

It is apparent that the indices and stocks in the USA are the most investigated ones (62), followed by Taiwan (28) and China (25). Overall, only four out of seven global regions are present in our study: Asia, Europe, North America, and South America. Neither Africa and Australia are covered, nor is, obviously, Antarctica. The region North America is essentially limited to the USA, with Canada only being featured once in the literature. Furthermore, South America is only represented by Brazil. Conversely, Asia contains 16 individual markets, which is the largest set of markets in this study. Europe is represented by seven individual markets as well as the European Union itself. It is remarkable that even though the USA is by far the single most covered market, North America as a whole with 63 coverages in the literature clearly only ranks second among the regions, i.e., behind Asia with 119 instances. Europe with 12 mentions and South America with 7 mentions appear to play a minor role in the stock market prediction literature.

Fig. 10 displays the frequency of exchanges/indices mentioned in our study and indicates the respective market of each of them. As the figure illustrates, 45 unique exchanges/indices are present in the investigated literature (for a complete list with full names and markets, please see Appendix Table C.1).[2,3] The market that is Hong Kong Special Administrative Region (China) is abbreviated in Figs. 9 and 10 with Hong Kong.

It is apparent that the S&P 500, which is an index that covers the performance of the 500 largest companies in the USA, is with 50 mentions; this is the index that is most commonly referred to in our study. The second place is taken by the TAIEX [Taiwan] with 24 references, and the third place is shared by the NASDAQ Composite Index [USA] and the SSE Composite Index [China] with 12 references. Once again, the USA takes the first place in terms of references in the literature and contains three indices in the top five most mentioned indices. However, in terms of region, Asia is once again taking the leading role with 108 references to Asian indices, with North America only ranking second with 89. European indices are with 13 counts, a distant third, and South American indices with seven are an even further distant fourth. A different picture emerges if the actual number of indices is considered. Out of the 46 unique indices, Asia represents 24 of them, which is a majority share of 52.2%, whereas North America only contains 10 indices (21.7%). Europe is with 10 indices (21.7%), on an equal footing with North America even though the gap in terms of number of mentions of these indices compared with the North American ones is obvious. For South America, only one index (2.2%), namely the Brazilian IBOVESPA, is contained in this study.

Fig. 11 displays the frequencies and the market capitalization of indices with at least three mentions in the literature as well as selected other European and Asian indices of varied size and importance (see Appendix Table D.1).

From the overall 22 indices contained in this comparison, the four indices from the USA are the largest ones considered in the research papers. These indices range from 7.252 Trillion USD (Dow Jones Industrial Average Index) to 30.213 Trillion USD (NYSE Composite Index) (see Appendix Table D.1). All European and Asian markets in this study are below the lower limit of this range. In particular, the largest Asian market is the Shanghai Stock Exchange (SSE) with a market capitalization of 4.724 Trillion USD, and the largest European market is the

[1] Since each article can refer to stocks and indices of one or more market, the sum of all counts is larger than the number of research papers in this study. For instance, Chu et al. (2009) cover both, the NASDAQ Composite Index [USA] and the TAIEX [Taiwan], so this is accounted for once in the count of the "USA" and in the count of "Taiwan".

[2] Since each article can refer to one or more indices/exchanges, the sum of all count is larger than the number of research papers in this study and does not have to correspond to the market count. For instance, Cao et al. (2019) cover the NASDAQ Composite Index [USA], the S&P 500 [USA] and the Dow Jones Industrial Average Index [USA], so this is accounted for once in the count of the "USA" in Fig. 9 and overall three times in Fig. 10 (for each of these indices once).

[3] In the few cases where a certain exchange was supposed to be covered, no specific index was mentioned and no composite index of the market as whole exists, this study assumes that the main index of that exchange was used. For instance, Gocken et al. (2016) state that they cover the Turkish market (Istanbul Stock exchange), which we assume to refer to the BIST 100 index containing 100 large companies traded on the Borsa Istanbul.

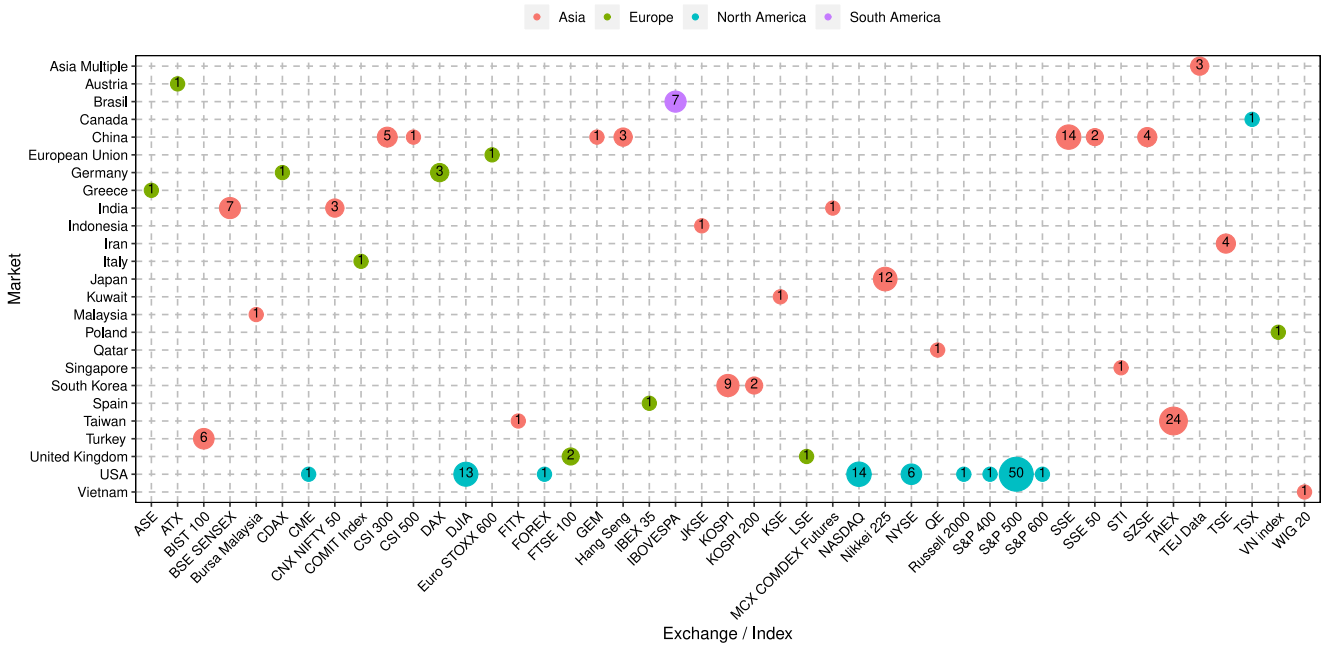

Fig. 10. Frequency of exchanges/indices (by market).

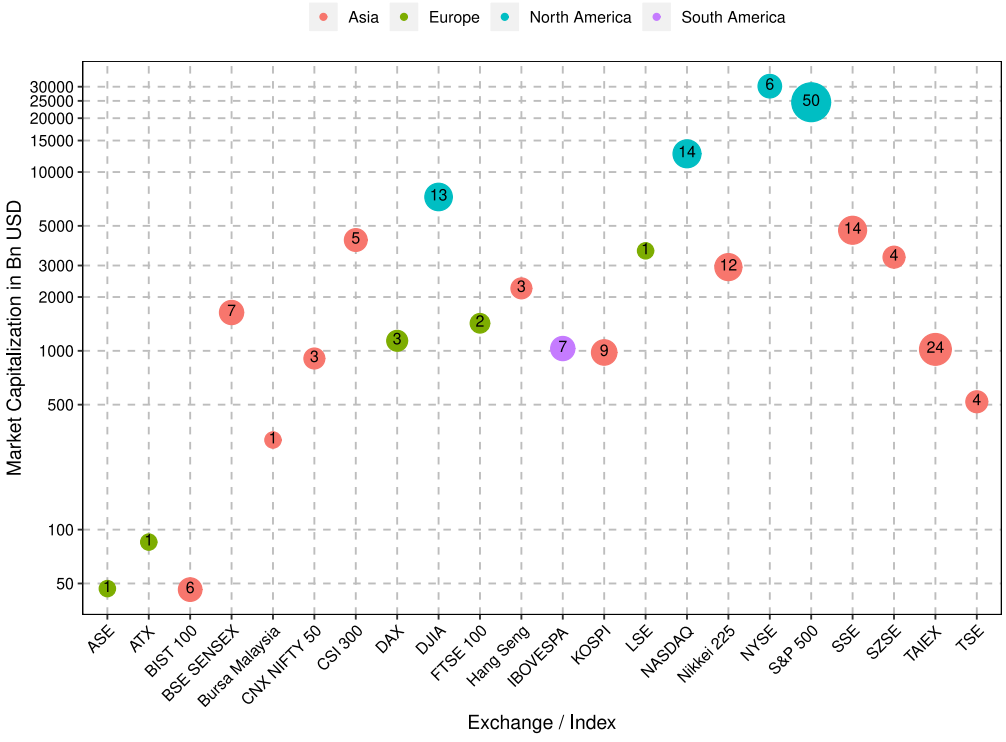

Fig. 11. Frequency of indices by market capitalization.

main equity market of the London Stock Exchange (LSE) encompassing 3.625 trillion USD. The correlation between the mentions of markets in the literature and their market capitalization is measured. The Pearson correlation of 0.545 and the Spearman rank correlation of 0.503 both indicate a moderate positive correlation between the frequency at which these indices/exchanges are mentioned in scientific literature and their respective market capitalization. Hence, there appears to be a tendency that larger stock markets (in terms of their market capitalization) are covered more often in the scientific literature than smaller ones. However, not all markets follow this trend. An example is the BIST 100 index from the Istanbul stock exchange, which is one of the comparably smaller markets with 46.2 billion USD but is referred to only six times. Conversely, the main equity market of the London Stock Exchange is only referred to once, but it belongs, with 3.625 trillion USD, to the larger stock markets considered in this study. It is also apparent that the largest stock market of each region is usually not the most investigated one. For North America, the largest index is the NYSE Composite Index [USA], but the S&P 500 [USA] is referred to 8 times

more.[4] For Asia, the largest index is the SSE Composite Index [China], but the TAIEX [Taiwan] is referred to almost twice as often. The same picture can be observed in Europe, where the LSE main market [United Kingdom] is the largest market in this region, but the DAX [Germany] is referenced 3 times more often.

In the following paragraphs, all stocks that were explicitly covered in our study – either as input to a stock market prediction model or as the target variable itself – are analyzed in more detail.

Overall, the largest share of stocks belongs to North American indices, closely followed by Asia. European and South American stocks were also considered in the scientific literature. In comparison with the indices presented in Fig. 10, it is apparent that the stocks covered in the scientific literature are more commonly North American (50.8%, excl. "Ceased to Exist") versus Asian stocks (42.2%), whereas North American stock markets (42.0%) were less frequently covered than Asian stock markets (50.2%).

It is noteworthy that the category "Ceased to Exist" represents stocks of companies that existed around the time a scientific article was written but went bankrupt/ceased to exist up to the time this study was conducted. This does not include stocks that were objects of a merger or an acquisition, since these stocks in this study are simply referred to by the company name after the merger or acquisition. This category is highlighted as "North America" since three out of four of the stocks contained in this group are from North America and only one is European.

This study includes 187 stocks that were mentioned at least once in one of the research papers considered in this study. However, of these 187 stocks, only 41 are covered in more than one research paper, which means that approximately 78% of stocks are only mentioned once. Those 41 stocks, which are subsequently termed "Top 41", and their frequency of being referenced is presented in Fig. 12.

At first glance, it is apparent that only stocks from North America and Asia were covered multiple times in the articles reviewed in this study. The stock that is most often referred to is the technology stock "Apple Inc". with 13 references to it and belonging to the GICS (Global Industry Classification Standard)[5] sector "Information Technology". The two companies "Microsoft" and "General Electric" belonging to the GICS sectors "Information Technology" and "Industrials" are on the second and third place, respectively. It is apparent and unsurprising that the stocks that are commonly referenced in the scientific literature are well-known, large cap stocks, i.e., companies with a large market capitalization, from North America. Since the regions of the Top 41 stocks are not representative of the regions of all stocks (187 positions), this warrants an investigation of the differences between the more frequently mentioned stocks (Top 41) and all stocks in this study. First, the distribution of the regions for both groups of stocks is highlighted in Fig. 13.

The comparison of all 187 stocks with the Top 41 stocks that were at least mentioned twice in the scientific articles, shows that the categorization into regions differs considerably between these two groups. It is apparent that the Top 41 stocks neither contain any positions from Europe nor from South America, which previously at least accounted for 2.1% and 3.2% of all stocks, respectively. In addition, the balance between stocks from North America and Asia is tipped in favor of North American stocks, as indicated in Fig. 12. Previously, Asia accounted for 42.2% of all stocks and North America for a slightly higher share of 52.4%. Conversely, the Top 41 contains only 12.2% (5) Asian stocks opposed to 87.8% (36) for North America. This trend of having an overweight of North American stocks only strengthens when higher frequencies are selected. For instance, when the Top 12 stocks are selected (threshold of 4 occurrences), this subset exclusively contains stocks from North America and no other region. This finding suggests that research is conducted on a set of stocks that is reasonably diverse in terms of regions, but, at the same time, is very centered on North American stocks.

Fig. 14 displays the frequency of the GICS sectors of all stocks compared with the set of the Top 41 stocks.

According to the GICS sector classification, there are overall 11 sectors that a company can be attributed to. For this study, we added one additional sector termed "Various (Conglomerate)" to account for companies that engage in multiple GICS sectors and/or have subsidiaries that do so. In our study, there were three stocks that belong to this sector, namely Mitsu & Co. [Nikkei 225], Sony Corp [Nikkei 225], and Reliance Industries Limited [CNX NIFTY 50]. Overall, the research articles contained in our study covered all 11 GICS sectors. The largest sectors for all stocks are "Information Technology" with a share of 18.2% (34), "Health Care" with 15.5% (29), and "Financials" with 13.9% (26). Conversely, the three largest sectors in the Top 41 do not contain "Financials" anymore and list "Health Care" with 24.4% (10) at the top, followed by "Information Technology" with 19.5% (8) and "Consumer Discretionary" with 14.6% (6).

Given that Apple Inc. and Microsoft are the top two stocks (see Fig. 12), it is unsurprising to find a large share of "Information Technology" companies among the most referred to stocks. However, the considerable share of "Health Care" appears, at first glance, surprising. However, most "Health Care" stocks within the Top 41 are only referred to twice in the literature. Notably, the considerable number of "Health Care" stocks is largely accounted for by two articles by Shynkevich et al. (2016) and Sedighi et al. (2019). The article of the former authors exclusively contains 28 "Health Care" stocks, whereas the paper of the latter authors simply covers the largest 50 American stocks, which is a large proportion of the S&P 500 overall. By themselves, these two papers together mention seven out of the 10 "Health Care" stocks (or their acquisitions) in the Top 41 at least twice and the remaining three out of 10 stocks once. Overall, it is apparent that technology stocks embody a large share of the stocks that are very often referenced in the scientific literature on stock market predictions. It is also noteworthy that no stocks from the sectors "Materials", "Utilities", and "Real Estate" are in the Top 41 stocks. Fig. 15 illustrates the comparison of all stocks to the Top 41 stocks in terms of Index, GICS sector, and region.

At first glance, the set of all stocks for each of the indices/exchanges appears rather diversified in terms of the sectors, especially for the S&P 500 [USA], the Nikkei 225 [Japan], and the KOSPI [South Korea]. However, it is noteworthy that for the TAIEX [Taiwan], which is the fourth most common index in this study, the GICS sectors only contain "Information Technology" (9) and "Industrials" (4). As highlighted previously, the Top 41 stocks only contain stocks from "North America" and "Asia". In particular, of the original 16 indices, only five are linked to the Top 41 stocks. These include the S&P 500, which is the most common index in our study, as well as the KOSPI and Nikkei 225, which are the second and third most common indices, respectively. In addition, one stock of the CNX NIFTY 50 is contained in the subset of the Top 41, which is surprising since this index and/or stocks from it are covered in only a few publications. This stock is "Reliance Industries Limited" [India], which is covered in the article by Patel et al. (2015) as well as in the one by Selvamuthu et al. (2019). It is

[4] It should be remarked that most of the companies contained in the S&P 500 are traded on the NYSE and NASDAQ exchanges and, hence, are also contained in the NYSE and NASDAQ Composite Indices. In case an index was covered in a research paper, the authors of this paper counted the specific index (e.g., S&P 500). In case a number of stocks were considered, and these were contained in multiple indices, each stock is associated with the narrowest of the indices (of the list of indices covered by the remaining scientific articles in this study) it is contained in. For instance, the S&P 500 covers 500 of the largest American companies and is narrower than the NASDAQ Composite Index with more than 2000 securities covered. In case of multiple stocks with the narrowest indices being in the same market, a majority vote was used to only keep the most common index for the stocks in the same market.

[5] https://www.msci.com/documents/1296102/11185224/GICS+Sector+definitions+Sept+2018.pdf/afc87e7b-bbfe-c492-82af-69400ee19e4f

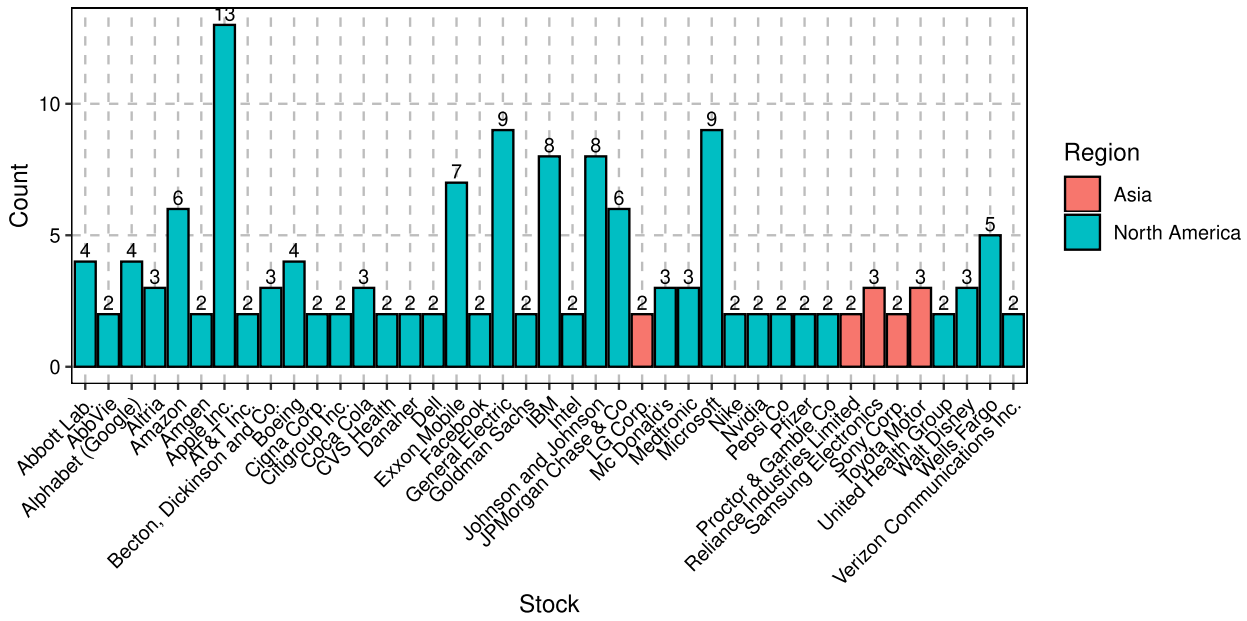

Fig. 12. Frequency of the Top 41 stocks.

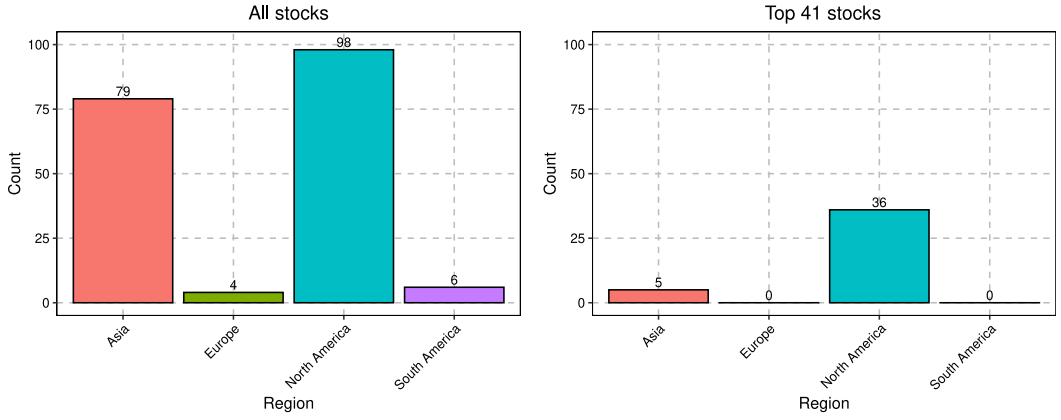

Fig. 13. All stocks vs the Top 41 stocks by region.

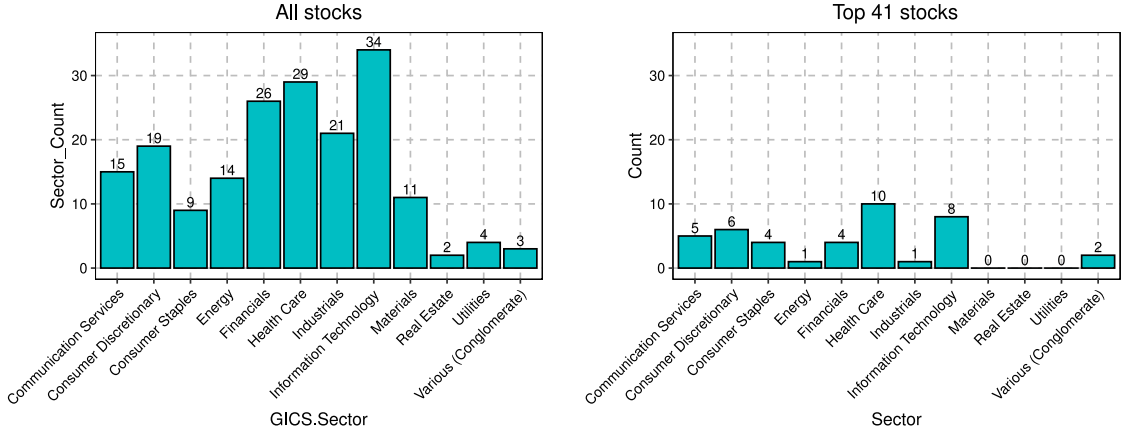

Fig. 14. All stocks vs the Top 41 stocks by GICS sector.

apparent that the S&P 500 still covers the same eight GICS sectors in the Top 41 as it does in the set of all stocks. The four most common combinations in the Top 41 are stocks from the S&P 500 in "Health Care" (10), "Information Technology" (5), "Communication Services" (5), and "Consumer Discretionary" (5). It is worth mentioning that the number of companies for a certain sector may not be representative of a sector's weight in the stock market or in a specific stock market index. For instance, for a market capitalization-weighted index such as the

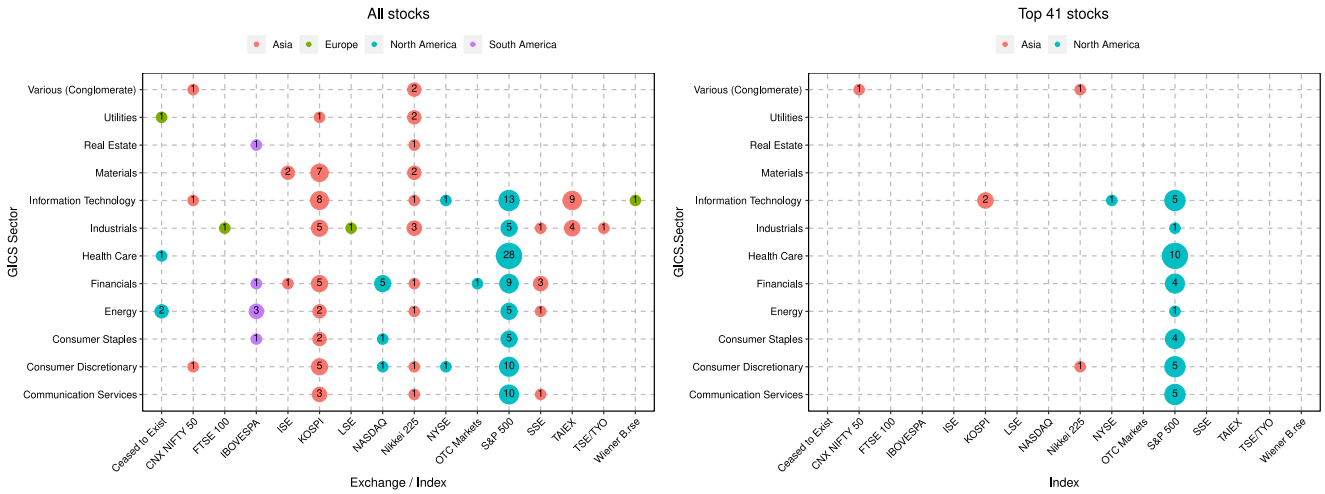

Fig. 15. All stocks vs the Top 41 stocks by index/exchange, GICS sector and region.

S&P 500, the size (= market capitalization) of the company determines its weight on the index, whereas for a price-weighted index such as the DJIA the magnitude of the price determines its impact on the index level. Thus, depending on the index, the actual weight of a sector may be linked only indirectly to the number of stocks in that sector but directly to the companies' capitalizations or prices.

Overall, the analysis of the markets, indices, and stocks indicated that the stock market prediction literature is focused on stocks from "North America" and "Asia". In addition, the S&P 500 is the most investigated stock market index, and the most common stocks under investigation are "Apple Inc.", "Microsoft", and "General Electric", which are three constituents of the S&P 500.

5.3. Variables for stock and index prediction

This literature review encompasses 138 scientific articles containing 2173 unique variables. Each variable was assigned to one of four types: "Technical Indicator", "Macro-Economy", "Fundamental Indicator", and "Other". The category "Technical Indicator" concerns variables used in finance as part of technical analysis of time-series data, most commonly, on company stocks. In this study, this type is further subdivided into the sub-types "Basic Technical Indicator" and "Other Technical Indicator", where the former focuses on essential indicators such as closing, open, high and low prices, as well as volume, and the latter includes other, on average, more sophisticated technical indicators such as the relative strength index (RSI), moving averages, stochastic oscillators, etc. The variable type "Macro-Economy" is segmented the most. It contains, as sub-types, "Economic Performance", which focuses on performance indicators such as GDP and industrial production, "Interest Rate & Money Supply", which contain, e.g., treasury bill rates and different variables indicating the circulation of money, "Exchange Rate", which contains exchange rates such as USD/EUR, and "Commodity" which contain commodities such as precious metals (e.g., gold) or crude oil. The type "Other" has no sub-group division and encompasses all other variables, including financial model-related ones (beta in the capital asset pricing model (CAPM), excess return, etc.) and machine learning ones (linear regression line, bag of words, etc.). Fig. 16 displays the number of variables by sub-type.

It is apparent that the largest type by the number of features in our study is "Technical Indicator" with 1348 variables (62.0%), in particular "Other Technical Indicators". The "Macro-Economy" type is with 279 variables (12.8%), ranked third, and these variables are distributed well among the four sub-types. For the type "Fundamental Indicator" with overall 157 variables (7.2%), the majority of variables come from the financial reporting of companies, specifically from the balance sheet and profit and loss statements.

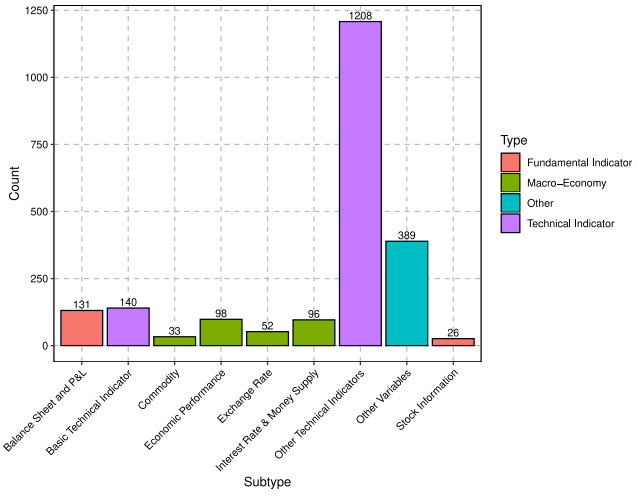

Fig. 16. Number of variables by sub-type.

The average number of features used in an article contained in this study (excluding bag of words variables[6]) is approximately 29.1. The minimum number of variables is one, which commonly concerns time-series papers such as the one by Pai and Lin (2005), which only uses the closing price to construct an ARIMA model. The maximum number of variables is 780 in the paper of Sun et al. (2019) when papers with bag of words variables are excluded. Fig. 17 illustrates how the (average) proportion of the type, including the bag of words variables, differs with the number of variables used in a research paper.

It is apparent that the proportions of the types used in the papers change as the overall number of variables in a paper increases. Articles that incorporate less than 10 variables rely extensively on "Technical Indicator" variables and only have a small share of "Macro-Economy" variables. Towards the 100 variables used, the share of technical variables continuously decreases, whereas "Macro-Economy"

[6] The four research works of Zhang, Zhang et al. (2019), Shi et al. (2019), Zhou et al. (2018) as well as (Feuerriegel & Gordon, 2018) each contain several thousand articles/tweets / news which are often converted into a bag of words (text mining approach), meaning counts for the main terms in the text. Since the number of the most important terms (usually several thousands and more) exceeds the number of features in each of the remaining papers by far and can be considered outliers, the average does not include them.

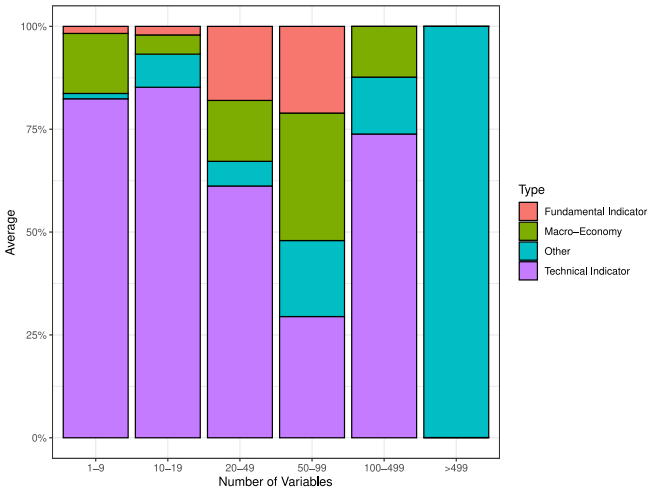

Fig. 17. Type by number of variables.

and "Fundamental Indicator" type variables become more common. For "100–499" variables, the trend appears to shift again to a higher share of technical indicators. However, there are only three papers belonging to this range of variables used, which is substantially less than at least 10 papers for all the previous categories and makes this trend less reliable. For the last category exceeding 499 variables in a study, there are six papers. At first glance, it appears that only "Other" variables are used. This is not entirely true since only five out of the six of these papers are almost exclusively using the bag of words/text mining approach to generate a comparably large set of features, often exceeding hundreds or thousands of key terms. Hence, even though at least the paper by Sun et al. (2019) is having a minority share of "Other" variables, such as the bag of words among the 780 variables in the paper, the large number of bag of words variables from the remaining papers concedes only a negligible share to the other three variable types. Even though the number of articles in this last category is comparably low, it is plausible that especially approaches that can lead to a large set of variables, such as text mining, will take a large proportion of the variables in papers with large variable sets. In the subsequent sections, all variable types and sub-types will be presented and depicted in more detail.

5.3.1. Technical indicator variables

Basic technical indicator variables

The basic technical indicators contain 140 variables that express essential price and volume information from stock markets and were directly used as input variables. Such variables can be the close price of the past day, i.e., the last recorded price at which a stock was traded on that trading day, the high and low prices, which mark the maximum and minimum price during that time period, the open price and the volume, as well as variations in these variables. All these basic technical indicator variables can be assigned to seven categories: "Close Price" with 87 counts (24.4%) but only 20 variables, "Volume" with 76 counts (21.3%) and 28 variables, "Range" with 61 counts (17.1%) but also 61 variables, "Open Price" with 51 counts (14.3%) and 11 variables, "High Price" with 41 counts (11.5%) and 9 variables, "Low Price" with 39 counts (10.9%) and also 9 variables, as well as the "Bid/Ask Price" minority category with only 2 counts (0.6%) for 2 variables. Fig. 18 shows all basic technical indicator variables that are mentioned at least twice in the literature.

It is apparent that five basic variables are used considerably more often than all the others. These are the close price with lag 1, i.e., from the previous day/period, the high price with lag 1, the low price with lag 1, the open price with lag 1, as well as the volume of a single period. Since these represent the most recent information for a forecast of the next period, these variables are more commonly used than those with higher lag values, longer time periods, or less commonly available information such as the number of trades.

Other technical indicator variables

This section contains all technical indicators that can, on average, be considered more sophisticated than basic technical indicators, and often require some form of processing of the basic information such as the closing prices or volume. It encompasses all financial technical indicators that were not part of the basic technical indicators analyzed in the previous section and embodies the largest single group with 1208 variables. These indicators can be grouped into four main categories: The "Momentum Indicator" (46.4%), which measures the speed with which a price is changing over time, the "Trend Indicator" (25.7%), which focuses on the direction and strength of a change, the "Volatility Indicator" (18.0%), which measures how much a variable, such as price, is changing and fluctuating during a certain time period, and the "Volume Indicator" (9.5%), which focuses on the volume that is associated with changes. Since this category contains a large number of variables but most of these have been used in the literature only once, the subsequent analysis focuses only on variables that were at least deployed twice. This is true for only 187 out of the 1208 variables (15.5%). Fig. 19 displays the 31 variable subcategories that these 187 technical indicators can be assigned to.

The subcategory that is by far the largest is the "Return" category, which contains variables such as return over the last period (14 counts) and the return over the last 10 periods (12 counts). The second largest category is the "Simple Moving Average" (SMA), for which the 10-period SMA (26 counts), and the 5-period SMA (22 counts) of the original prices are the most common representatives. Ranked third is the "Relative Strength Index" (RSI) with 14 periods (30 counts), 6 periods (11 counts) and 5 periods (11 counts) being the most commonly applied time windows. Overall, it is apparent that the "Momentum" and "Trend" indicators constitute the majority of the 187 technical indicators. Fig. 20 displays the Top 30 variables by count of all the variables in this section.

Unsurprisingly, all of the variables mentioned before can also be found in the Top 30 variables. In particular, the RSI of 14 periods ranks first among all technical indicators.[7] The second, third, and fourth ranked variables are all "Trend" indicators, including the SMA with 10 periods and with 5 periods, as well as the MACD based on the difference between a 12-period EMA and a 26-period EMA using a 9-period EMA signal line.[8] Besides that, there are many return variables with different time frames (including 1, 5, 10, 15, and 20 periods), different Stochastic %K and Stochastic %D indicators with different time frames, Williams %R and exponential moving averages (EMA) with periods of different lengths. It is apparent that the most deployed technical indicators focus on the "Momentum" of prices as well as the "Trend" that a time series is following. Only very few "Volume" and "Volatility" indicators are included among the Top 30 variables.

In addition to the presented results of technical indicators in our review, it is worth mentioning additional studies (Ahmadi et al., 2018; Chavarnakul & Enke, 2018; Picasso et al., 2019) that used technical analysis to examine stock market behaviors.

[7] It should be remarked that the RSI with 14-period is assumed as the standard RSI. In a few cases when the RSI was used in a publication, but the time period was unspecified, it was assigned to the RSI (14-period).

[8] It should be remarked that the MACD (12-EMA – 26-EMA, 9-EMA) is assumed as the standard MACD. In a few cases when the MACD was used in a publication, but the EMA periods were unspecified, it was assigned to the MACD (12-EMA - 26-EMA, 9-EMA).

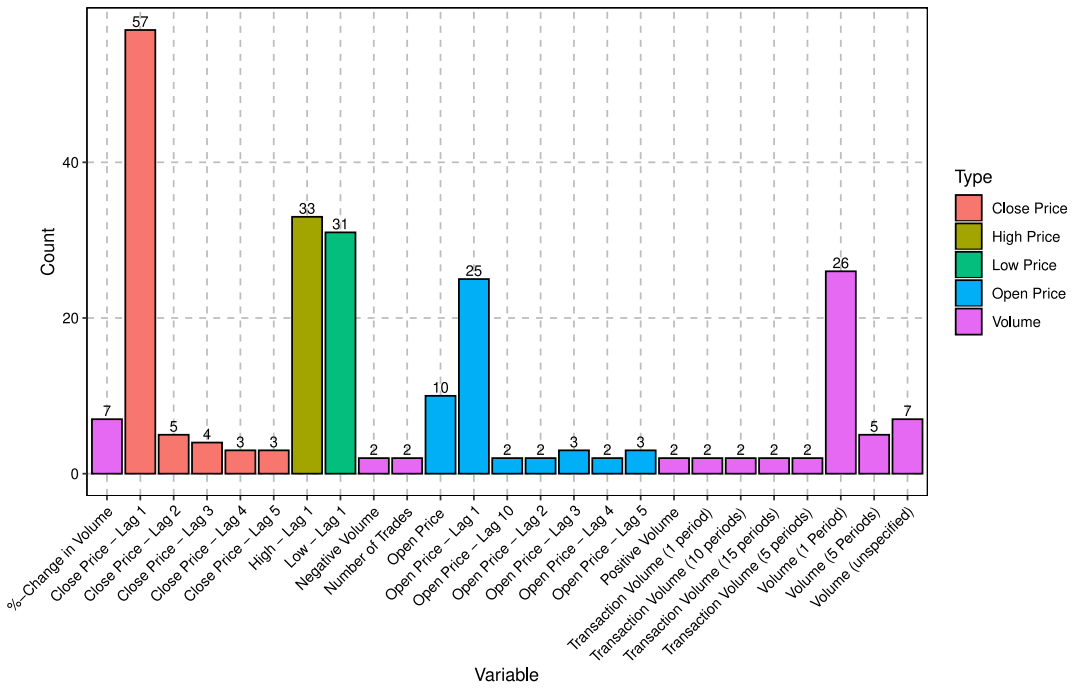

Fig. 18. Basic technical indicators.

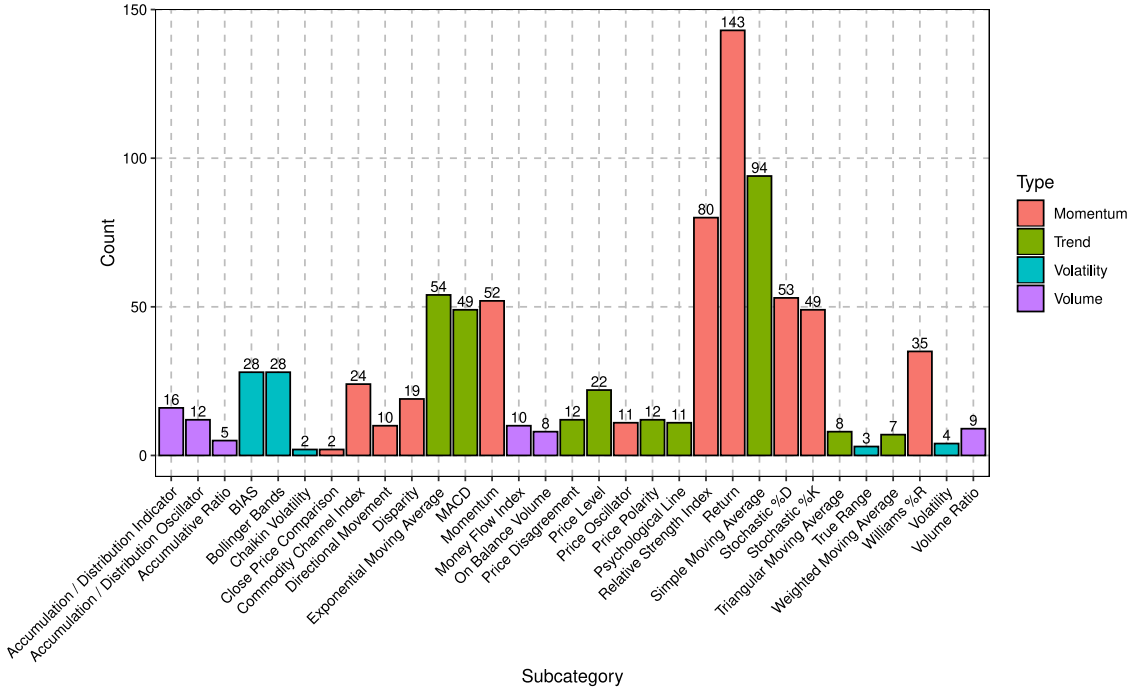

Fig. 19. Subcategories of technical indicators.

5.3.2. Macro-economic variables

Exchange rates

The articles analyzed contained overall 53 unique variables related to exchange rates. These exchange rates include 12 currencies: 11 single currencies (such as the US Dollar) and one basket of currencies. This basket of currencies contains six currencies for the US Dollar Index, which compares the US Dollar with six major currencies (Euro, Japanese Yen, British Pound Sterling, Swiss Franc, Canadian Dollar, and Swedish Krona). The most common currency considered for exchange rates is the US Dollar, which was contained in 51 out of 53 exchange rate-based variables (96.2%), followed by the Japanese Yen with 14 (26.4%) and the Chinese Yuan with 10 occurrences (18.9%). Fig. 21 displays the specific exchange rates covered and the regions are involved.

First, it is apparent that the US Dollar to Japanese Yen is the most frequently used exchange rate, followed by the US Dollar to the Chinese Yuan and the US Dollar to the Canadian Dollar. On a regional level, the exchange rates were mainly focused on North America to Asia with 24 counts (45.3%), followed by North America to Europe with 14 counts (26.4%). Looking further onto the variable level, the most common variables included in stock market prediction are the %-change in the

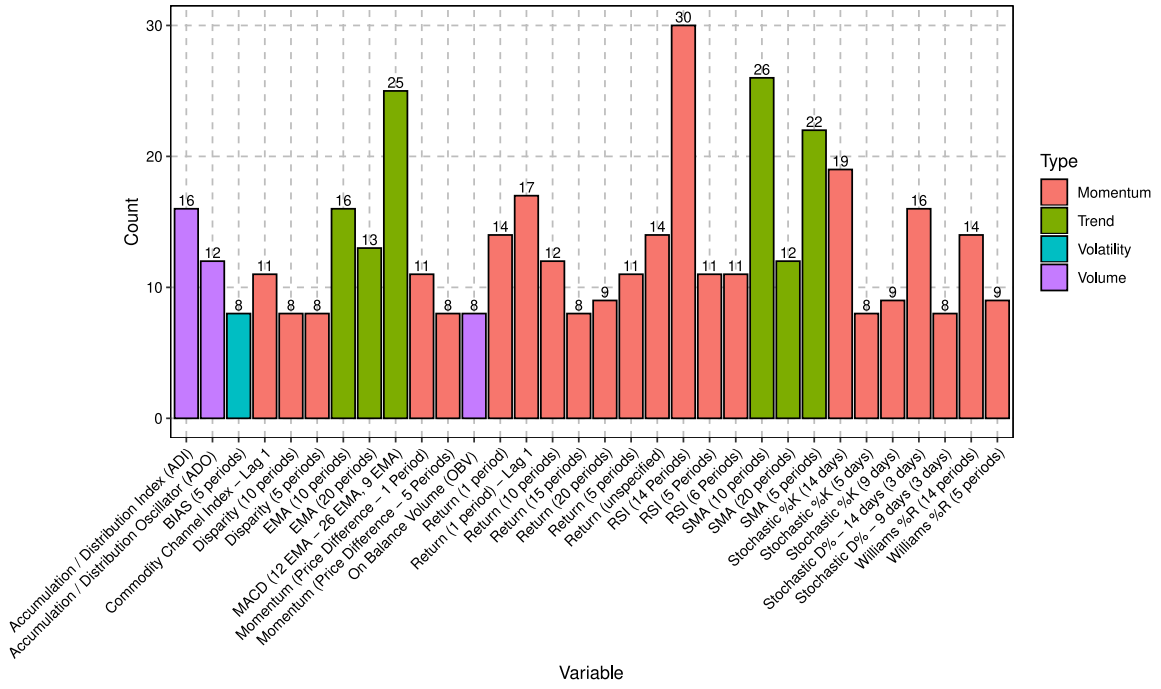

Fig. 20. Top 30 technical indicators.

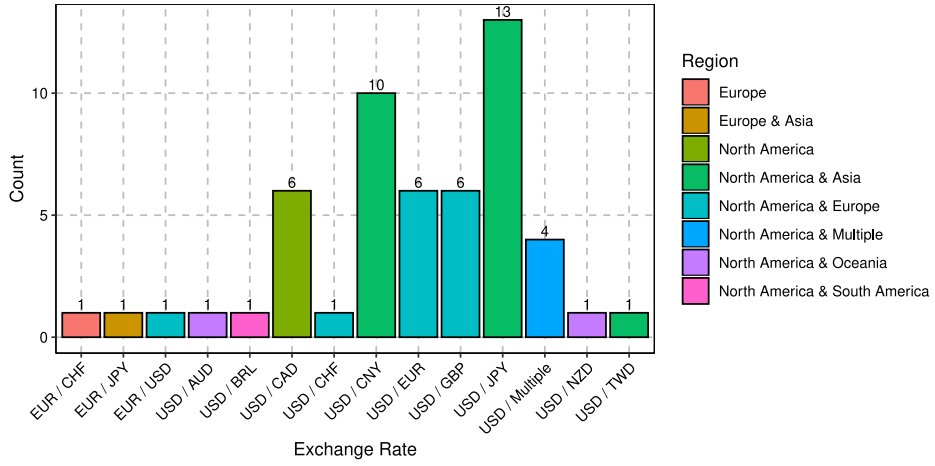

Fig. 21. Exchange rates.

exchange rate: USD/JPY, USD/GBP, and USD/CAD with 5 counts each, as well as the exchange rate USD/JPY with lag 1, also with 5 counts.

Commodities

There are overall 33 unique variables covering commodities in this study, which encompass precious metals and energy-related commodities such as crude oil and gas. The largest share of these (57.6%) are represented by indices (such as the United States Commodity Index) or via exchange traded funds (ETFs). Using the commodity prices for a single commodity is the second most common way to incorporate these variables (36.4%), and using commodity futures is the least common approach (6.1%). Fig. 22 displays all 33 commodity-based variables and the type of commodity they represent.

Overall, it is apparent that crude oil is with 20 occurrences (42.6%), the commodity that is most often referred to in our study. In addition, gold has, with 14 mentions (29.8%), a considerable share in the stock market prediction literature. The four remaining commodities only account for approximately one quarter of all mentions in the literature. Commodities indices covering a basket of commodities account for 14.9%, silver for 6.4%, gas for 4.3%, and copper for only 2.1%. There are three variables that were most often considered in the stock market prediction literature in our study. These are the %-change in the crude oil price, the gold price, as well as the crude oil price with lag 1, i.e., from the previous period.

Economic performance

In this category of macro-economic variables, there are 98 unique variables relating to roughly five categories. These categories are (1) "National Productivity", (2) "Trade", (3) "Income & Investment", which includes private and governmental income as well as foreign investments and governmental spending, (4) "Labor Market", and the (5) "Sector" performance of selected sectors in an economy. Since

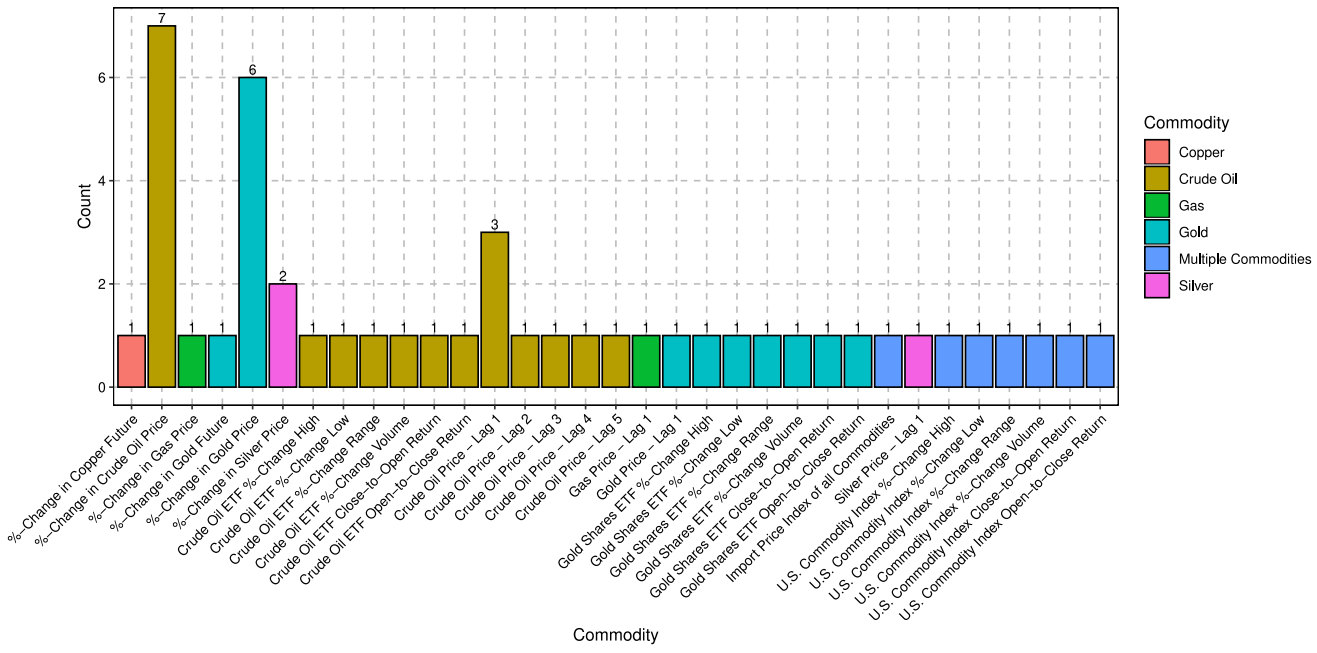

Fig. 22. Commodities.

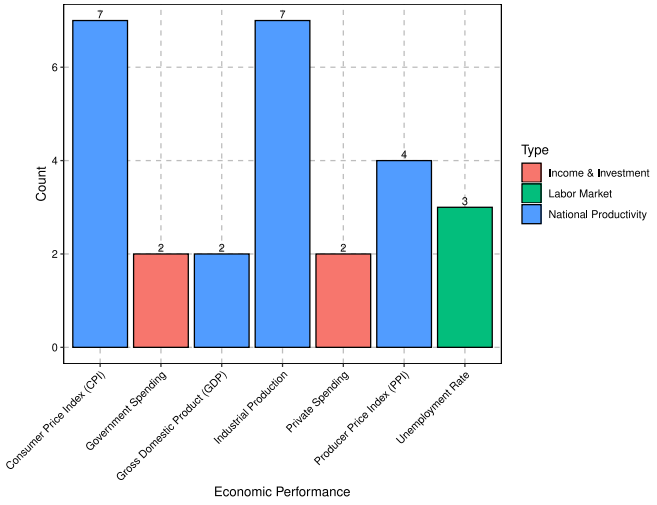

Fig. 23. Top 7 Economic variables.

"National Productivity" is the broadest of these categories, it is unsurprising that it encompasses 35 variables (35.7%), which is largest number of variables. "Trade" is with 28 unique variables (28.6%), ranked second, followed by "Sector" with 18 variables (18.4%) and "Investment & Spending" with a count of 15 (15.3%). The last category "Labor Market" only contains two variables (2.0%) and is by far the smallest category. Fig. 23 depicts the Top 7 variables by count for Economic Performance, which are the variables that were deployed at least twice in the articles covered in our study. Once again, it is apparent that only a small share of variables is used multiple times throughout all journal articles.

Roughly, only 7.1% of the variables in this category are used at least twice in the scientific literature that this study investigated. Of these seven variables, the three most common ones all belong to the category "National Productivity". In particular, "Industrial Production", a measure for the industrial output, and "Consumer Price Index,[9]" a measure for inflation, both accumulate a count of 7 and share the first rank. The "Producer Price Index" as a measure of the change of selling prices that producers receive, is ranked third. It is noteworthy that the national "Unemployment Rate" is the only "Labor Market" variable that is contained in the Top 7. Variables from the categories "Trade" and "Sector" are not contained in the set of variables that is at least deployed twice in the literature.

Since only seven out of 98 variables are mentioned at least twice, the counts for all subcategories of the Economic Performance are illustrated in Fig. 24 to provide a more detailed picture of all the variables.

It is unsurprising that the subcategories "Consumer Price Index" and "Industrial Production" belong to the set of most referred to subcategories. The third largest subcategory within "National Productivity" is the "Gross Domestic and National Product", which includes all variables related to the GDP and gross national product (GNP). This subcategory is comparably large, since it encompasses several variables that were only used once in the literature, such as the GNP deflator, real GDP, as well as changes in GDP and GNP over 3, 6, and 12 months. Considering the counts of all variables, the subcategories belonging to the category "Trade", namely "Export" and "Import", also appear relevant. "Export" represents 14 counts with variables such as real exports, export volume, export amount, and export growth rate. The subcategory "Import" accounts for 10 mentions with similar variables, including real imports, import volume, import amount, and import growth rate. The category "Income & Investment" consists of three subcategories that focus on "Government Spending & Income", in absolute terms or in relation to the GDP; "Private Spending & Income", including the change of private spending over 3, 6, and 12 months; and "Foreign Investment", which consists of the approval of foreign investments as well as related discount rates for investments.

[9] Note that even though it is widely accepted that in the last decades the short-term interest rates respond to inflation, interest rate shocks seem not to significantly impact inflation (Bataa et al., 2019). Moreover, interest rates lag inflation and economic activity (Bataa et al., 2019) and it is not necessarily a one-to-one change, which provides support to consider inflation (in terms of CPI) and interest rates separately

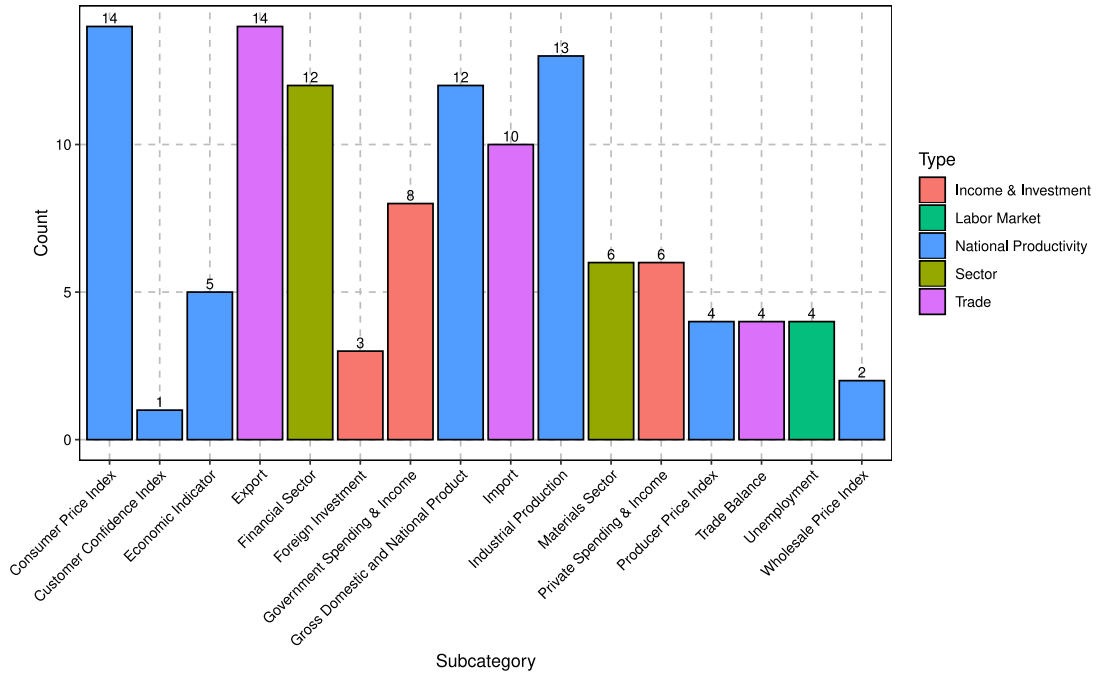

Fig. 24. Variables by sub-category.

The "Labor Market"-related variables can be exclusively grouped into the subcategory "Unemployment" and cover the national and foreign unemployment rates.

Interest rates and money supply

In this category, there are 55 variables used in the literature that were used in overall 96 different forms (e.g. prices, differences, %-changes, yields etc.) in the corresponding studies. The variables belong to seven categories defined by the authors of this study. These categories are "Certificate of Deposit", "Corporate bond", "Default spread", "Money Supply", "Term Spread", and "Treasury". Certificates of deposit, corporate bonds as well as treasuries are financial instruments that enable investors to invest their funds with the objective to either obtain interest/coupon payments and/or capital gains from a bank, corporation, or government, respectively. Term spreads refer to the differences in yields of securities with different time horizon (maturity) whereas default spreads refer to the difference of risk associated with different investments.

Fig. 25 displays the 28 variables that were mentioned at least twice in the articles contained in this study.

It is apparent that only variables from six categories are mentioned at least twice and that the "Loans & Interest Rates" category is missing entirely. The "Treasury" category containing governmental debt obligations, such as treasury bills (maturity of up to 1 year), treasury notes (maturity of 2 to 10 years), and treasury bonds (maturity of more than 10 years), embodies the largest category. In addition, it contains the six largest variables overall, with the 3-month treasury bill ranking first, followed by the 10-year treasury note and the 5-year treasury note. It is apparent that short- and medium-term government securities are the most commonly used, whereas the long-term 30-year treasury bond represents the least discussed security in this category. The second largest category is the "Default Spread", which highlights the differences in the yield between securities with different risk levels. In most cases, the spread is measured between corporate bonds with a rating Baa, which is a relatively low-risk bond compared to governmental securities, which commonly – also in case of the USA – have a better credit rating at Aaa/AAA (or comparable), indicating the highest credit worthiness and the lowest level of risk. Beside the fact that in most cases the spread to a corporate bond of quality Baa is used as reference, there appears to be no clear preference with respect to the maturity of the treasury it is compared with. However, in most cases, a short-term treasury bill with a maturity ranging from one to six months is used. In addition to comparing default spreads, term spreads are apparently a common variable choice for stock market predictions. The "Term Spread" category contains variables that represent the difference in the rates of treasuries with different maturities. The most common one appears to be a medium- vs short-term spread between a 10-year treasury note and a 3-month treasury bill. The remaining variables presented in Fig. 25 mainly highlight other debt instruments such as certificates of deposit (CDs) and corporate bonds, as well as money supply variables such as M1 and M2 money supply, indicating the value of the money available (e.g., in circulation).

5.3.3. Fundamental indicator variables

Stock information variables

This section includes variables that are based on or related to the stock of a company that is traded on a public stock exchange. There are 26 variables in this category that can be grouped according to six categories: "Price Ratio" with 18 counts (30.5%), "Earnings" with 13 counts (22.0%), "Market Value" with 10 counts (16.9%), "Dividend" with 8 counts (13.6%), "Book Value" with 5 counts (8.5%), and "Number of Stocks & Shareholders" with 5 counts (8.5%). The twelve variables that were used at least twice in the articles in this study are presented in Fig. 26.

The category "Price Ratio" has the largest share in the subset of variables used at least twice. In particular, the price/earnings ratio, often termed as the PE Ratio, which sets into relation the price of a share to the earnings per share, is by far the most common variable in this category. The price/sales ratio, the earnings-per-share (EPS), and the market value of the stock, i.e., the market capitalization, are all ranked second after the PE ratio. In addition, it is apparent that a company's earnings play a front role in the stock and stock market prediction literature, given that six out of 12 variables mentioned twice in the literature use some form of the earnings (included as part of a price ratio). Overall, each of these variables is either related to the earnings power of a company, its ability to pay dividends to shareholders, or its value (book and market).

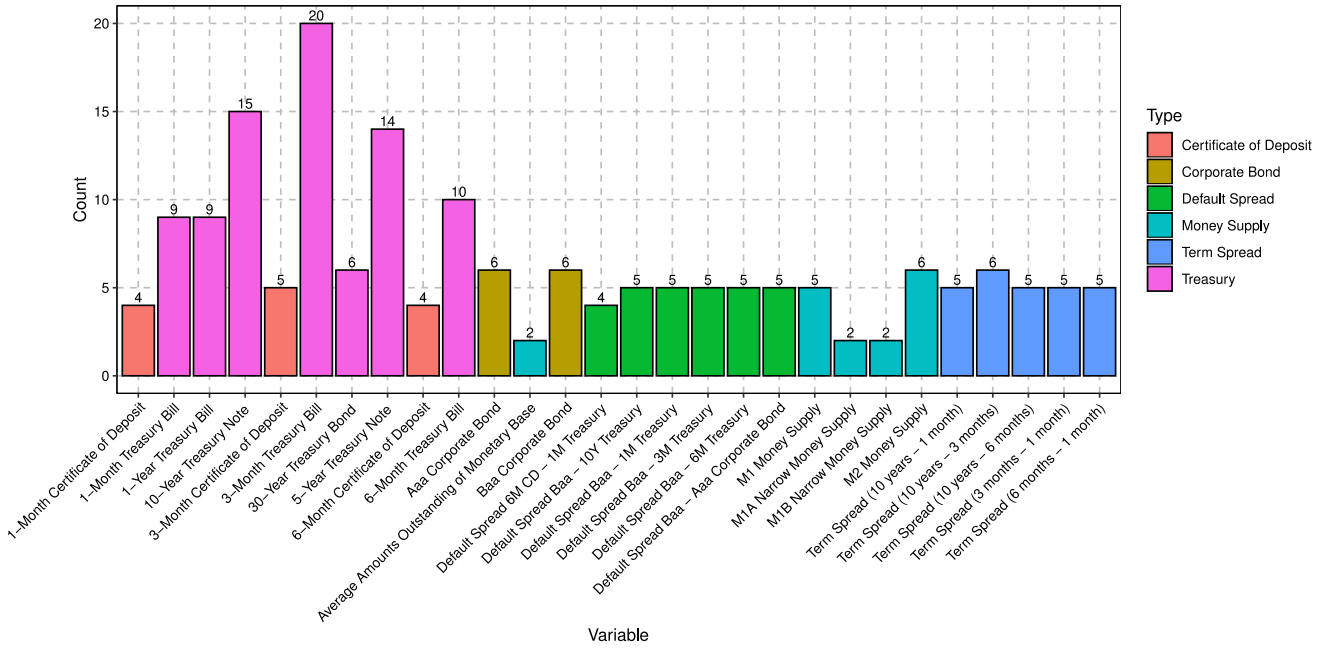

Fig. 25. Interest rates and money supply.

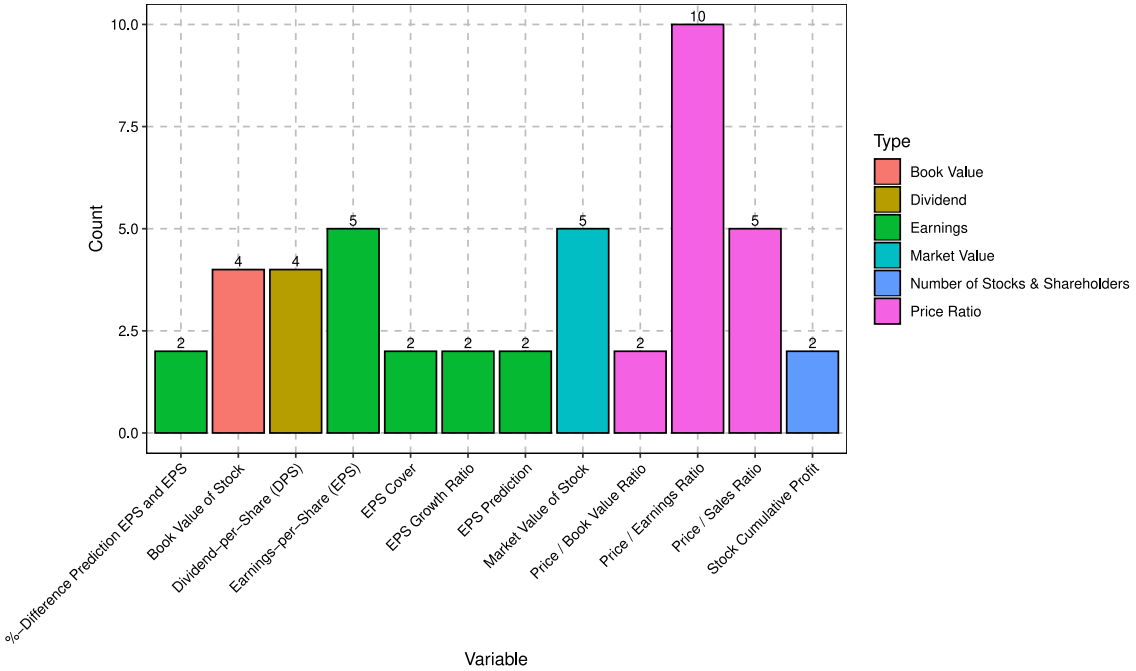

Fig. 26. Stock information variables.

Balance sheet and profit and loss statement variables

This section contains variables related to the financial reporting of companies, in particular, the balance sheet and the profit and loss (P&L) statement. The 131 variables contained in this category were assigned to eight categories. Of these categories, "Profitability Ratio" is with 44 variables (33.6%), the largest, mostly containing some measure of income or profit that is set into relation with some other income or capitalization measure to indicate the company's ability to achieve financial profit/gain. The second and third largest categories are "Capitalization Ratio" with 22 variables (16.8%), indicating how and how well a company is capitalized, and "Activity Ratio" with 19 variables (14.5%), showing how well a company deploys its assets to generate sales or cash. The remaining categories by size are "Liquidity Ratio" (9.9%), "Growth Ratio" (9.9%), "Profitability" (8.4%), which measures the profitability in absolute and not in relative terms (as a ratio), "Investment & Financing", (3.8%), and "Capitalization" (3.1%). Since the vast majority of variables concerning capitalization and profit were ratios (e.g., debt- to-equity ratio, operating income to total assets, etc.), the authors decided to keep them separate from the capitalization and profitability variables that were presented in absolute terms (e.g., total assets and operating profit). Of all 131 variables, only 43 (32.8%) are mentioned at least twice in the scientific articles in this study. These 43 variables and their types are displayed in Fig. 27.

The quick ratio, a "Liquidity Ratio" indicating a company's ability to cover its short-term liabilities with highly liquid assets, ranks first with eight counts. The variables ranking second are the current ratio

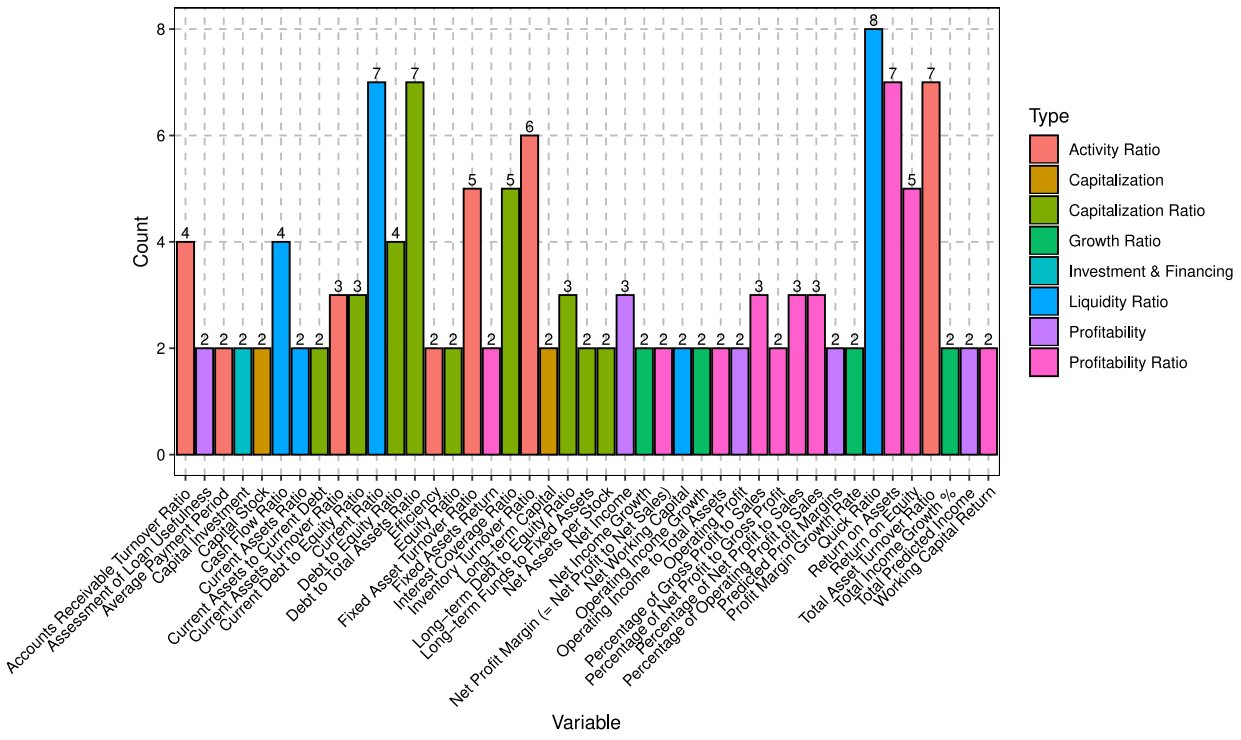

Fig. 27. Balance sheet and profit & loss variables.

("Liquidity Ratio"), debt-to-total assets ratio ("Capitalization Ratio"), the return on assets ("Profitability Ratio"), and the total asset turnover ratio ("Activity Ratio"). It is noteworthy that the three largest categories are on par with each other with a count of 31 for the category "Profitability Ratio", a count of 30 for the "Capitalization Ratio" and a count of 29 for the "Activity Ratio". The "Liquidity Ratio" only encompasses 23 counts, whereas all other categories encompass 11 or less counts. Hence, the relative profitability of a company, how relatively well it is funded, and its relative ability to use its assets appear to be the most relevant variables from financial statements for stock and stock market predictions.

5.3.4. Other variables

There are two sets of variables that were not included in the previous categories presented. The first one contains the stocks and market variables used as input for a stock market prediction model. In their prediction models on the S&P 500 ETF, Zhong & Enke (2017, 2019) used seven major global equity market indices (HSI, SSE, CAC 40, FTSE 100, DAX, DJIA, NASDAQ) as well as eight of the largest stocks contained in the S&P 500 (Apple, Microsoft, Exxon Mobile, General Electric, Johnson & Johnson, Wells Fargo, Amazon, JP Morgan Chase). Niaki and Hoseinzade (2013) also use a combination of major world indices containing six of the same market indices (HSI, CAC 40, FTSE 100, DAX, DJIA, NASDAQ) and most of the same stocks (Microsoft, Exxon Mobile, General Electric, Johnson & Johnson, Proctor & Gamble). A similar approach is pursued by Baek and Kim (2018) for their KOSPI 200 and S&P 500 prediction for ten of the largest constituents of each of these indices.

Conversely, several papers only made use of other market indices without relying on index constituents or other large cap stocks. This approach was followed by Lohrmann and Luukka (2019), who deployed, in their S&P 500 prediction model, basic technical indicators and returns from six major equity markets (DAX, Nikkei 225, HSI, FTSE 100, Euro STOXX 50, Russell 2000) as well as the Vanguard Total World Stock ETF, the iShares MSCI Emerging Markets ETF, and the volatility index VIX. Rosillo et al. (2014) use three variables based on the VIX index to also predict the S&P 500. Hoseinzade and Haratizadeh (2019) predict five major stock market indices using the return information of 10 major stock market indices including those that are predicted. A similar number of major stock market indices is incorporated into the Bayesian Network of Malagrino et al. (2018) to predict the direction of the daily movement of the IBOVESPA index. They deployed a set of up to 12 other stock markets for their predictions. The authors (Na & Sohn, 2011) predict the KOSPI using the directional information of eight global stock market indices (DJIA, Nikkei 225, SSE, TAIEX, HSI, CAC 40, FTSE 100, DAX). Conversely, Huang et al. (2005) include only a single stock market into their predictions. In particular, for the prediction of the Nikkei 225, these authors use the S&P 500 index as one of the model inputs. Essentially the opposite is done by Tsai and Hsiao (2010), who use information on the TAIEX to predict stocks traded on the Taiwan Stock Exchange. Lastly, Zhang, Shao et al. (2019) incorporate into their SZ50 prediction not only the Dow Jones Industrial Average Index but also information on 30 technical indicators that were derived from it.

It is interesting to note that Zhang, Shao et al. (2019) additionally used information from financial news using a bag of words approach. Several other authors also included and followed a bag of words approach for their predictions. Feuerriegel and Gordon (2018) used 80,813 ad-hoc announcements for their prediction of the DAX, CDAX, and Euro STOXX 600. Another noteworthy use of textual information was implemented by Shi et al. (2019), who used 341,310 news articles as well as additional stock-related tweets (from Twitter) for the prediction of stocks in the S&P 500.

In addition to the bag of words approach, other model variables and measures are incorporated in this category. From the comparably few that were referred to at least twice in the literature, the most noteworthy are the Sharpe ratio (with different lags), offering a comparison of the return of an investment above the risk-free rate to the risk associated with that investment, as well as variables related to the capital asset pricing model (CAPM). In particular, the excess return, i.e., the return above the risk-free rate, as well as the beta factor and the expected return of a company according to the CAPM were

used. Finally, other model variables, such as those from the GARCH model, the EGARCH model, and linear regression, can be found in the literature.

The efficacy of social media data for stock market prediction

The rise of micro-blogging websites and social media produces a vast amount of data and information (Carta et al., 2021a), which often contains individuals' opinions. Some studies (see, e.g., Yuan, Liu et al., 2020) have already studied how social media activities of influential people affect market movements. In addition, web- and text-mining-based prediction models in stock market prediction studies have received much attention recently due to the rise of financial activities and discussions in social media and online platforms (Huang & Liu, 2020). Given this, in this sub-section, we specifically focus on studies that used social media data for stock market forecasting.

In Section 5.3.4, we already presented some study examples of stock market forecasting based on financial news (Zhang, Shao et al., 2019), ad-hoc announcements (Feuerriegel & Gordon, 2018), and stock-related news and tweets (Shi et al., 2019). In addition, we further explored some related works, which are summarized next. Even though these additional papers were not part of the reviewed literature in our study, we examine them to provide additional evidence in favor of the recent trend in stock market predictions using social media data.

A study by Li et al. (2017) used Twitter data to forecast stock price movements of 30 companies listed on the NASDAQ or the New York Stock Exchange and confirmed the effectiveness of textual data on stock price changes. Zhang et al. (2018b) extracted sentiments from web news and social media to develop a stock price prediction framework. They collected stock-related events and web news from the Chinese financial discussion board (Guba) to extract stock-related posts. (Carosia et al., 2019) also attempted to examine to effects of social media sentiments (in tweets) on the Brazilian stock market, conducting several experiments with different machine learning techniques.

Similar to news data, social media data in the literature come in various forms. Even though most of the existing research used Twitter data, some studies explored data from other social media platforms such as StockTwits. For example, Carta et al. (2021a) proposed a method to detect highly relevant events in finance using the textual information extracted from traditional news articles and Stocktwits messages. This proposed approach was validated in the experiment conducted on the data from News and Analytics, Dow Jones, StockTwits, and the price time series of the S&P 500 index. Huang and Liu (2020) used chip indicators, social media reviews, and replies as a source of research data to develop an improved stock forecasting model based on sentiment analysis. Based on the findings, they highlight a correlation between sentiment scores and stock price movements.

5.3.5. Granularity of the data used

The time series data used for stock market forecasting can not only concern different time periods in history but also different granularity (5-minute frequency, daily, weekly, monthly). The dot plot in Fig. 28 illustrates, for the selected studies, the combination of granularity and the length of the time horizon for the data contained in the corresponding study. Most studies focused on daily predictions, with some studies using such data for a period of several months (see, e.g., Hsieh et al., 2011; Lai et al., 2009) or even several years, for example, five years (see e.g., Chang & Wu, 2015) or 10 years (see, e.g., Zhong & Enke, 2017b). The second most common frequency of the data was of monthly predictions. For monthly data, each study covered enough monthly observations to span a time horizon of more than 10 years, with a maximum period length of 80 years (from 1926 to 2005) in the study of Yu et al. (2009). Other periodicities were less common in our study. It is noteworthy that three out of the four studies that used 5-minute frequency data utilized deep learning as the forecasting technique.

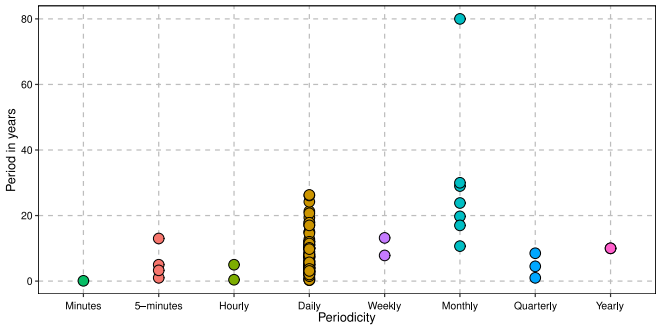

Fig. 28. Granularity of the data used in each study.

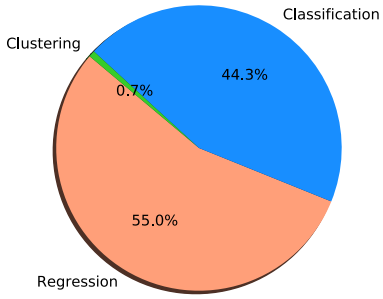

Fig. 29. Classification results of the reviewed articles.

5.4. Review of machine learning techniques

The second main objective of our review study was to provide a thorough presentation of the machine learning-based forecasting models deployed in the literature for stock market forecasting.

We classified all prediction models in our set of reviewed articles into regression (supervised learning), classification (supervised learning), or clustering (unsupervised learning) according to the type of prediction used. Fig. 29 displays a pie chart showing the percentages of each of these model types in the selected articles. As highlighted in the pie chart, regression type studies account for the majority of cases (54.3%). Classification-based studies still account for a large share (44.3%), whereas clustering models only covered approximately 1.4% (two studies only).

Table 7 lists the main machine learning methods deployed in the literature as well as the publications that contained these methods. The references were allocated to the method, which was the main approach applied for the prediction in the corresponding article. It is apparent that neural networks were the most widespread machine learning technique applied for stock market predictions. In 39 of the 138 studies, neural networks constitute the main machine learning method. In addition to simple neural networks (which commonly use only a single hidden layer), there are 15 recently published articles that focus on deep learning (neural networks with several hidden layers) to conduct stock market forecasting. Most of the remaining studies covered either SVM or fuzzy-theory-based methods, accounting for 24 and 23 articles, respectively. The last category, "Others", contains the studies that developed novel machine learning methods or simultaneously applied several approaches to compare the predictive performance.

5.4.1. Neural networks

Early studies in stock market research have used ANN models alone to predict stock prices or their movements (Altay & Satman, 2005; Cao et al., 2005; de Faria et al., 2009; Fadlalla & Amani, 2014; Kara et al., 2011; Maknickiene & Maknickas, 2013; Nermend & Alsakaa, 2017; Niaki & Hoseinzade, 2013; O'Connor & Madden, 2006; Olson & Mossman, 2003; Safer, 2002). Besides, several authors have combined

Table 7
Summary of article counts with machine learning techniques and relevant references.

Main method	Article count	Reference
Neural Network	39	Armano et al. (2005), Cao et al. (2005), Chang et al. (2009), Chen et al. (2003), Chiang et al. (2016), de Faria et al. (2009), Ebadati and Mortazavi (2018), Ebrahimpour et al. (2011), Enke and Thawornwong (2005), Gocken et al. (2016), Hsieh et al. (2011), Hu et al. (2018), Hyup Roh (2007), Kara et al. (2011), Kim and Han (2000), Laboissiere et al. (2015), Lam (2004), Liao and Wang (2010), Lu and Wu (2011), Maknickiene and Maknickas (2013), Mo and Wang (2018), Mostafa (2010), Nermend and Alsakaa (2017), Niaki and Hoseinzade (2013), O'Connor and Madden (2006), Olson and Mossman (2003), Pei et al. (2017), Qiu and Song (2016), Qiu et al. (2016), Ramezanian et al. (2019), Safer (2002), Selvamuthu et al. (2019), Sun (2014), Sun et al. (2019), Ticknor (2013), Wang and Wang (2015), Wang et al. (2011), Zhou et al. (2019)
SVM/SVR	24	Cao and Tay (2001), Chai et al. (2015), Chang and Wu (2015), Chen et al. (2017), Chen and Hao (2018), Das and Padhy (2018), Gowthul Alam and Baulkani (2019), Huang et al. (2005), Kao et al. (2013), Kazem et al. (2013), Kim (2003), Lu (2013), Lu et al. (2009), Pai and Lin (2005), Pan et al. (2017), Rosillo et al. (2014), Rustam and Kintandani (2019), Sedighi et al. (2019), Tay and Cao (2001), Xiong et al. (2014), Yeh et al. (2011), Yu et al. (2009), Zhang, Shao et al. (2019), Zhang, Teng, and Chen (2019)
Fuzzy theory	23	Anbalagan and Maheswari (2015), Atsalakis and Valavanis (2009a), Cagcag Yolcu and Alpaslan (2018), Chang and Fan (2008), Chang and Liu (2008), Chang et al. (2016), Chen and Chen (2015), Chen et al. (2014), Chu et al. (2009), Efendi et al. (2018), Hadavandi et al. (2010), Javedani Sadaei and Lee (2014), Kaur et al. (2016), Lai et al. (2009), Pal and Kar (2019), Rajab and Sharma (2019), Rubio et al. (2017), Wang (2002), Wei et al. (2011), Yang et al. (2011), Ye et al. (2016), Zhang, Zhang et al. (2019)
Deep Learning	15	Baek and Kim (2018), Borovkova and Tsiamas (2019), Cao and Wang (2019), Chen et al. (2019), Chong et al. (2017), Chung and Shin (2018), Fischer and Krauss (2018), Gunduz et al. (2017), Hoseinzade and Haratizadeh (2019), Lien Minh et al. (2018), Shi et al. (2019), Singh and Srivastava (2017), Song et al. (2019), Wen et al. (2019), Zhong and Enke (2019)
Feature selection	6	Barak and Modarres (2015), Huang and Tsai (2009), Lohrmann and Luukka (2019), Tsai and Hsiao (2010), Weng et al. (2017), Zhang et al. (2014)
Classifier ensembles	5	Barak et al. (2017), Basak et al. (2019), Chun and Park (2005), Khaidem et al. (2016), Tsai et al. (2011)
Text Mining	2	Feuerriegel and Gordon (2018), Shynkevich et al. (2016)
Bayesian Network	2	Malagrino et al. (2018), Zuo and Kita (2012)
KNN	2	Cao et al. (2019), Zhang et al. (2017)
Clustering	2	Vilela et al. (2019), Zhong and Enke (2017a)
PCA	1	Zhong and Enke (2017b)
Others	17	Anish and Majhi (2016), Araújo (2011), Araújo et al. (2015), Bisoi et al. (2019), Chen and Hao (2017), Chou and Nguyen (2018), Enke et al. (2011), Göçken et al. (2019), Gorenc Novak and Velušček (2016), Hassan et al. (2007), Leung et al. (2000), Na and Sohn (2011), Patel et al. (2015), Son et al. (2012), Yang et al. (2019), Zhang et al. (2018a), Zhou et al. (2018)

neural networks with other soft computing techniques to improve their prediction model. Such research studies are illustrated in more detail in Fig. 30, together with the auxiliary methods they incorporated. Studies that proposed a novel approach were highlighted in the figure as well.

Kim and Han (2000) proposed an ANN approach combined with a genetic algorithm (GA) for the prediction of stock prices. The GA was deployed here not only to optimize the model parameters but also to discretize the feature space. Qiu and Song (2016) also optimized an ANN using GA for the prediction of the direction of the Japanese stock index price. A similar approach (GA + ANN) was also proposed by Ebadati and Mortazavi (2018) in the context of stock market forecasting. In the study by Lam (2004), the rule extraction technique GLARE was used to compensate for a parameter misspecification and noise in the data during the training of the neural network. Furthermore, the study by Wang et al. (2011) proposed a novel prediction model called the wavelet de-noising-based backpropagation (WDBP) neural network. In this model, a wavelet transform was applied to decompose the data into multi-layer signals. Subsequently, this backpropagation neural network was applied to newly generated low-frequency signals in each layer to forecast future prices. Chiang et al. (2016) proposed an adaptive intelligent stock-trading decision support system that utilized a first wavelet transformation and then a particle swarm optimization (PSO)-based neural network to predict the stock price direction. Here, PSO was integrated into the neural network training algorithm to address the shortcomings of the backpropagation algorithm.

Armano et al. (2005) presented a hybrid model that integrates an extended classifier system (XCS), a GA, and a neural network to predict stock index prices. In this NXCS model, the XCS used a GA to optimize a rule-based system to overcome the reinforcement learning problems in a feed-forward neural network. Two of the challenges faced in stock price prediction is the volatility and the inherent noise in the data. A common consequence for neural networks using such data is that they easily over-fit, which reduces the prediction capability of these models. Ticknor (2013) addressed these problems by employing Bayesian regularization in the training of the neural network. In this approach, Bayesian regularization was used to assign the weights of the network, which then allowed the network to optimize the model. Laboissiere et al. (2015) analyzed the capabilities of the ANN as well as the selection of the most influential variables for the prediction of the daily minimum and maximum stock prices. In this study, a correlation-based

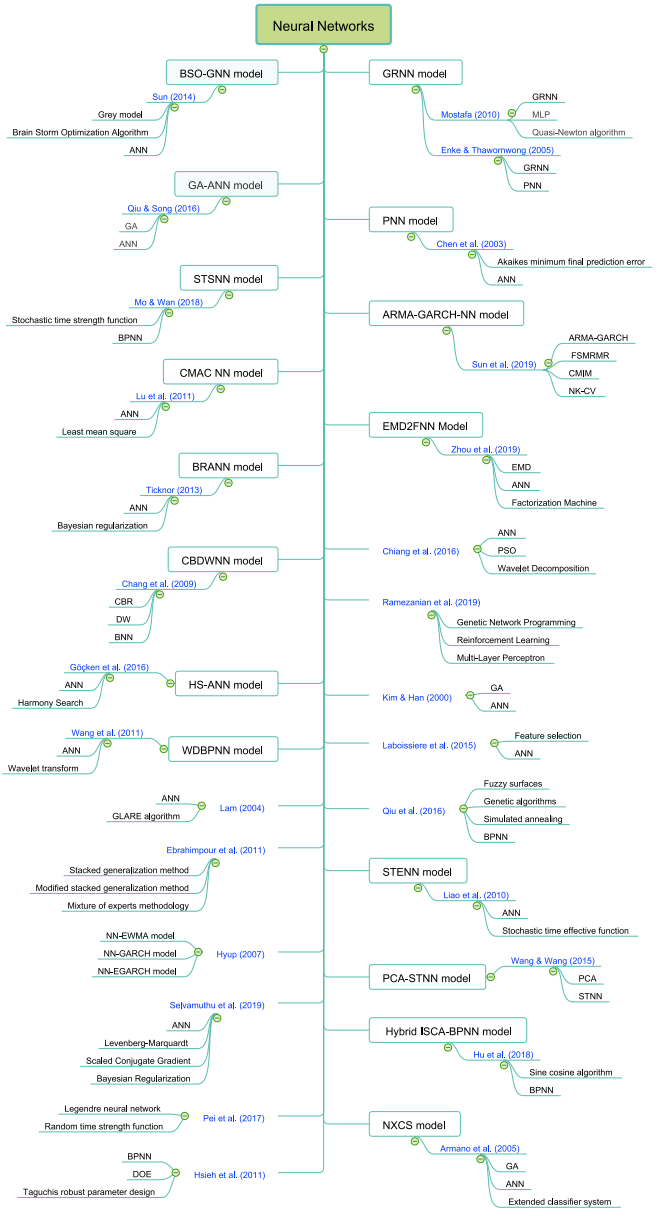

Fig. 30. Studies based on neural networks.

feature selection method was applied to select the most critical variables for three Brazilian companies. Gocken et al. (2016) used harmony search (HS) for the first time with an ANN to achieve better results in stock market forecasting compared with using neural networks alone.

Liao and Wang (2010) attempted to improve the forecasting ability of neural networks using a stochastic time-effective function and introduced a novel prediction method called the STENN model. The purpose of the stochastic time effective function was to avoid the loss of useful information in the selection of time-variant data during the training. STENN approach was developed further using principal component analysis (PCA) by Wang and Wang (2015) to forecast the SSE, HS300, S&P500, and DJIA indices. The empirical results obtained with the proposed PCA-STNN model indicated that his model outperformed the traditional BPNN, PCA-BPNN, and STENN models. A study by Mo and Wang (2018) also considered improving the predictive performance of the STENN model by introducing a return scaling cross-correlation function of an exponential parameter to forecast the return scaling cross-correlations between the Shenzhen Stock Exchange (SZSE) Component Index and the Shanghai Stock Exchange (SSE) Composite Index.

Chen et al. (2003) attempted to predict the direction of stock price return using a probabilistic neural network (PNN) model. In the proposed approach, they deployed a statistical procedure, Akaike's minimum final prediction error (FPE), to select potential input features from the data and calibrate the specifications of the model. Enke and Thawornwong (2005) examined the effectiveness of the generalized regression neural network (GRNN) and PNN models for level estimation and classification with respect to their capability to predict the monthly direction of the S&P 500 stock index return. They used the PNN and the GRNN for classification and level estimation, respectively. The results in this study indicate that the trading strategies using neural network in the context of classification achieve higher profits in the same market conditions compare with level-estimation-based neural network models as well as linear regression models and a buy-and-hold strategy. The GRNN model has also been used in Mostafa (2010) to predict the movement of the Kuwait Stock Exchange closing price. The author found the forecasting performance of the GRNN method to be superior to that of the multi-layer perceptron (MLP) neural network model. It was also shown that a quasi-Newton training algorithm achieves higher accuracy in prediction compared using quick propagation and conjugate gradient descent training algorithms.

Hsieh et al. (2011) integrated the Taguchi method (frequently used in design of experiment (DOE), based models) with a BPNN to improve the quality of the optimization process in the network and achieved more robust predictions. Sun (2014) developed a novel hybrid forecasting model called BSO-GNN based on the brain storm optimization (BSO) algorithm, the Grey model, and an ANN. Taking advantage of the ability of the Grey model to handle data with small samples, this study first introduced a Grey neural network (GNN) model to forecast daily stock index prices. The BSO was utilized to overcome some drawbacks in the parameter optimization of the GNN approach. Furthermore, Hu et al. (2018) presented an improved sine cosine algorithm (ISCA) for weight optimization in the BPNN model, introducing the novel ISCA-BPNN approach to predict the movements of the opening price of the S&P 500 and the Dow Jones Industrial Average Index (DJIA). Following a more econometric approach, Sun et al. (2019) developed a machine learning algorithm named ARMA-GARCH-NN to detect intra-day patterns to predict stock market shocks. The proposed method was formed using an ARMA-GARCH model (that estimated the market shocks), the forward selection minimal-redundancy-maximal-relevance criterion (FSMRMR), conditional mutual information maximization (CMIM), an ANN, and a nearest-k cross-validation (NK-CV). Both FSMRMR and CMIM were utilized to select the most relevant features from the data, and NK-CV was used as a novel cross-validation technique.

Hyup Roh (2007) studied the predictive power of the ANN, combining it with time-series models such as exponentially weighted moving average (EWMA), GARCH, and EGARCH. The author concluded that integrating time-series models with ANN can help in effectively forecasting the volatility of stock index prices. A study by Chang et al. (2009) combined dynamic time window (DTW), case-based reasoning (CBR), and an ANN to create an integrated approach termed CBDWNN for forecasting stock trading signals. In this approach, a case base dynamic window was operated to reduce false alarms with respect to buying and selling signals in the backpropagation neural network. Lu and Wu (2011) proposed an efficient cerebellar model articulation controller neural network (CAMC NN) for forecasting stock index prices. On account of its noise resistance, fast learning, and generalization capability, this study used the CAMC NN scheme to predict stock index prices. Also, the least mean squares (LMS) algorithm was applied as the learning rule in this scheme to update the weights. An effort by Qiu et al. (2016) addressed the problem of the non-linearity of historical time-series data by applying the ANN with a GA or simulated annealing (SA) and fuzzy surfaces. GA and SA were introduced here to achieve optimal weights and biases to enhance the predictive performance of

the ANN. In addition, the fuzzy surface was utilized to select relevant input variables.

Ebrahimpour et al. (2011) experimented with three variants of ANNs: stack generalization, modified stack generalization, and a mixture of multi-layer perception (MLP) Experts[10] for forecasting the Tehran stock market. The mixture of MLP experts was effective compared with the other methods in the experiment. Pei et al. (2017) investigated the prediction of stock prices using an improved Legendre neural network with a random time strength function (LeNNRT). As they reported, timely effectiveness and randomness in the learning process for updating weights of the network were the key advantages of using the LeNNRT method.

Lastly, we focus on three studies that proposed modern ANN structures for stock market predictions. Emphasizing the challenges posed by non-stationary and high noise in stock data, Zhou et al. (2019) introduced a new hybrid model named the empirical mode decomposition based neural network (EMD2NN) model for stock market trend prediction. In this model, both empirical model decomposition (EMD) and the factorization machine (FM) were operated for analyzing non-stationary data. Moreover, Ramezanian et al. (2019) applied a model including genetic network programming (GNP), reinforcement learning, and a multi-layer perceptron (MLP) neural network for the classification and prediction of stock returns. GNP was applied here to optimize the initial rules for the MLP classification process. The aim of reinforcement learning was to strengthen the optimization process and achieve better rules. Finally, Selvamuthu et al. (2019) used an ANN with three algorithms, Levenberg–Marquardt (LM), scaled conjugate gradient (SGC), and Bayesian regularization, using high-frequency data as well as tick data for the prediction of the Indian stock market.

5.4.2. Support vector machines

Since several authors have highlighted many difficulties with ANNs, such as locally optimal solutions, over-fitting (Kim, 2003; Yeh et al., 2011), and time complexity (Das & Padhy, 2018) during the active period, it is unsurprising that many studies in the field of stock market forecasting also focused on other prediction models. Among these, one of the most widely used methods are support vector machines (SVMs) for classification and support vector regression (SVR) for regression. Since the SVMs are based on the structural risk minimization principle, it can reduce the generalization error (Chai et al., 2015; Pai & Lin, 2005). Accordingly, there have been a lot of papers with SVM- and SVR-based models in financial market forecasting, which are discussed in more detail in the following paragraphs.

Lu (2013) proposed a novel approach by combining nonlinear independent component analysis (NLICA), SVR, and PSO for stock index forecasting. Here, NLICA was used to extract the relevant features dealing with non-linearity properties, and PSO was applied to optimize the parameters in the SVR. Rustam and Kintandani (2019) also examined an SVR with PSO for the prediction of Indonesian stock prices. However, the hybrid method of Das and Padhy (2018) that combines SVM and teaching-learning-based optimization (TLBO) outperformed the PSO+SVR model in the prediction of the closing price of the COMDEX commodity futures index.

In a different approach, Yeh et al. (2011) suggested a new method based on SVR, termed as MKSVR, co-operating with a multiple-kernel (MK) learning algorithm that tackled problems concerning the manual adjustment of the hyper-parameters of the kernel functions in the SVR. Similarly, a hybrid multi-kernel support vector machine (MKSVM) approach was proposed by Gowthul Alam and Baulkani (2019). In this model, factor analysis was used to filter out irrelevant features from the stock index data, and fruit fly optimization (FFO) was implemented to tune the parameters. This proposed method showed promising results in comparison to the MKSVM, PSO+MKSVM, and GA+MKSVM models in the prediction of daily price directions of stocks from the NYSE, DJIA, and S&P 500. Yu et al. (2009) presented another forecasting model, the evolving least squares support vector machine (LSSVM) learning paradigm with a mixed kernel, to predict stock market trends. In this method, GA was used twice, once to select essential features from the data, and once to optimize the parameters of LSSVM. The evolving LSSVM model is efficient and robust compared to other LSSVM models, especially with the use of different kernels individually, such as polynomial, radial basis kernel (RBF), sigmoid kernel, and mixed kernel. Chai et al. (2015) also proposed the LSSVM method in combination with an empirical mode decomposition (EMD) to forecast the CSI 300 index. In the EMD-LSSVM process, four-parameter optimization techniques (simplex, grid search (GS), PSO, and GA) were tested, and the results showed that the GS-based EMD-LSSVM method outperformed other methods in the prediction of stock price movements.

The study by Kao et al. (2013) aimed at the selection of a suitable wavelet sub-series for a forecasting model with improved forecasting accuracy, introducing a novel prediction model called Wavelet-MARS-SVR. This model used a wavelet transform (that decomposed the financial time-series data), multivariate adaptive regression splines (MARS) (that selected significant variables coming from the wavelet transformation), and SVR (that performed the prediction). Kazem et al. (2013) developed a forecasting model (called SVR-CFA) based on chaotic mapping, a firefly algorithm (FA), and SVR. To overcome shortcomings such as trapping in local optima and the slow convergence in the FA, chaotic mapping was applied instead of a random approach. The evidence in this study shows that the chaotic firefly algorithm is a better option than a GA to optimize the SVR hyper-parameters. Xiong et al. (2014) also used FA in the proposed model in that study, the multi-output support vector regression (MSVR), to predict interval-valued stock index prices over short and long time horizons. Later, in the study of Zhang, Teng, and Chen (2019), the FA was modified by introducing a dynamic adjustment strategy and an opposition-based chaotic strategy. In that context, the modified FA (MFA) was combined with an SVR to propose a novel hybrid forecasting model (SVR-MFA) for stock prices. The forecasting results with the SVR-MFA were more robust and also outperform the results of the SVR-CFA (Kazem et al., 2013) as well as the other optimization techniques (PSO-GA) applied with SVR models for the prediction of six variables in the Shanghai stock market.

The problem of the intuitively high level of noise in the analysis of financial data using SVM was addressed by Lu et al. (2009) from a different perspective by applying independent component analysis (ICA) together with SVM. The purpose of including the ICA was to detect and filter out noise from the financial time-series data and to then apply SVM subsequently. Moreover, one interesting attempt to forecast stock index prices using the exchange rate can be found in Zhang, Shao et al. (2019). They proposed a data-driven method based on SVM to forecast the movement of the Chinese stock index (SZ50) price using four data sources: technical indicators, exchange rates, a US market index, and financial news data. They demonstrated that using exchange rates for the prediction of this stock index price was reasonable compared with the other features used as input in that study. To select the relevant features, they applied a term frequency–inverse document frequency (TF–IDF) strategy from text mining for financial news data and a recursive strategy for other data sources.

Chen et al. (2017) developed an approach containing an SVM, the Adaboost.M2 algorithm, a naive Bayesian classifier, and the queen genetic algorithm (a type of GA) (QGA) to forecast stock market trading signals. Using the data of financial indicators, they tested two approaches: the Adaboost QGA-SVM, with and without financial news data. Across the boosting, the naive Bayesian classifier was applied to classify financial news by clarifying its effects on the stock price. This study achieved the highest accuracy with the Adaboost QGA-SVM

[10] Ebrahimpour et al. (2007) introduced the model, Mixture of MLP Experts (MMLPE), as a face detection method. This MMLPE method employs multi-layer perceptrons (MLPs) as expert and gating networks and uses a novel learning algorithm to adapt to the MLPs.

method without using financial news in terms of predictive performance. A hybrid model containing an SVM in combination with the ARIMA model was presented by Pai and Lin (2005) to examine its forecasting ability for stock prices. As mentioned earlier, this is one of the most cited articles in the applications of SVM models. Chen and Hao (2018) proposed another SVM-based prediction model (called PCA-WSVM), integrating PCA into weighted SVM to predict stock trading signals. A robust tool for mixed frequency issues and multi-output for stock price prediction was proposed by Pan et al. (2017). They called it the multiple output support vector machine unrestricted mixed data sampling (MSVM-UMIDAS) model. They revealed that the utilization of mixed frequency-independent variables presents various types of information significant to the stock price, and the use of multiple outputs takes into account many practical issues concerning stock price predictions.

Sedighi et al. (2019) proposed a novel upgraded prediction model deploying an artificial bee colony (ABC) algorithm, an adaptive neuro-fuzzy inference system (ANFIS), and an SVM. In the proposed ABC-ANFIS-SVM approach, the ABC algorithm was used to optimize the technical indicators, the ANFIS to forecast long-run stock price fluctuations, and the SVM to establish a link between the technical indicators and the stock price, as well as to further improve the forecasting accuracy of the model. The results obtained with the proposed method were more precise in the prediction of the US stock market than those from the 20 existing forecasting models.

The SVM-based studies discussed so far are summarized with their key topics in Fig. 31. It is noteworthy that, in this figure, we have not included the studies Cao and Tay (2001), Huang et al. (2005), Kim (2003), Rosillo et al. (2014), Tay and Cao (2001) that only utilized the SVM method without merging it with at least one other soft computing technique for making stock market predictions.

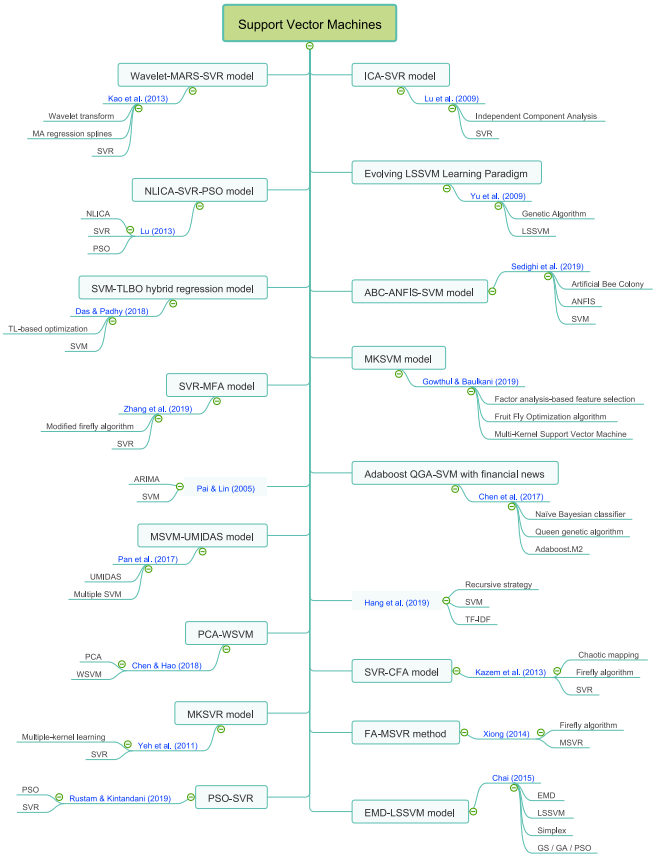

Fig. 31. Studies based on support vector machines.

5.4.3. Fuzzy Theory

In this section, we concentrate on fuzzy-theory-based techniques for modeling and predicting the stock market in the selected literature. Since Prof. Zadeh (Zadeh, 1965) introduced fuzzy sets, they have been applied in many real-world applications successfully. Fuzzy theory provides advances in real-time problems dealing with uncertainty. Over the last decade, techniques from fuzzy set theory have gained more attention in stock market research.

An early work by Wang (2002) addressed the difficulty of using a large volume of stock data as well as the uncertainty of the differences between two continuous time series for stock market predictions. To overcome these issues, the author constructed a system that minimized the size of the stock price data in terms of the storage requirements (e.g., MB) and worked through fuzzification methods in combination with Grey theory. The method they created was called the fuzzy Grey system, which can predict the stock price promptly at a specific time. Lai et al. (2009) used a novel framework combining k-means clustering, a GA, and a fuzzy decision tree (FDT) to forecast the stock price of the Taiwan Stock Exchange. The proposed GAFDT model first used k-mean clustering to obtain sub-clusters of stocks in the Taiwan Stock Exchange Corporation (TSEC) and then the GA to determine the optimal fuzzy term for each input index in the FDT. Generating decision rules via the FDT, the prediction was performed as the last step. Anbalagan and Maheswari (2015) investigated the ability of a fuzzy meta graph (FM) to forecast the stock market and to deal with issues such as non-stationarity and non-linearity of the time-series data. Efendi et al. (2018) argued that existing stock market prediction models are still poor at overcoming randomness and issues such as volatility and uncertainty of the stock prices in data preparation. Taking this as a motivation, they developed a novel forecasting method called fuzzy random auto-regression (FR-AR) model. In the FR-AR, the fuzzy random variable handles the low-high and stock price data, while the auto-regressive component deals with the stationarity of the data.

Fuzzy time series (FTS) has rapidly gained traction in modeling and predicting stock markets. The idea of FTS is to split the universe of discourse for fuzzy sets (partitions/intervals) from time-series data and then learn how each division acts by observing rules across time-series patterns (Song & Chissom, 1993). Accordingly, Chu et al. (2009) proposed a new FTS model (called the fuzzy dual-time series model) to predict a stock index using the stock index itself and its volume as dual factors based on the assumption that both affect the future price of a stock index. As discussed by Javedani Sadaei and Lee (2014), finding the effective length of each interval to determine linguistic variables is often challenging when using FTS in stock market forecasting. Addressing this issue, they proposed a model-based systematic, descriptive, and well-structured approach using FTS for forecasting stock market data. A similar study was established in Chen et al. (2014) for predicting the stock market using an FTS-based approach. Essentially, they deployed Pearson correlation coefficients to select the essential technical indicators and granular spread partition (GSP) to identify the partitions of the universe of discourse and the interval lengths in the FTS training process. Chen and Chen (2015) also applied an FTS-based granular computing approach with entropy-based discretization and binning-based partition methods to predict stock market prices. Another noteworthy attempt at the prediction of the stock market can be found in Ye et al. (2016), who used a multi-order FTS, technical analysis, and a GA. They started by constructing a multi-variable time series across the technical analysis and then employed the GA to determine the precise domain partition and improved the forecasting performance using a multi-order (first-order, second-order, and third-order) FTS.

Rubio et al. (2017) further developed the classical FTS for stock market forecasting by introducing fuzzy logical relationships (FLRs) and fuzzy logical groups (FLGs) for assigning weights and providing

trapezoidal fuzzy numbers as the predictions for future values of a stock index. The study by Cagcag Yolcu and Alpaslan (2018) evaluated the steps formed by FTS in a single simultaneous process by introducing a novel hybrid approach (called H-FTSM) to predict the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX). The H-FTSM model employed the FTS, a PSO (that identified the optimal fuzzy relations), a single multiplicative neuron model (that determined the fuzzy rules), and fuzzy C-means clustering. The proposed H-FTSM achieved promising results in the prediction of the TAIEX compared with the 12 benchmarks, including (Chen & Chen, 2015; Chen et al., 2014), used in their study. Similarly, Zhang, Zhang et al. (2019) proposed a forecasting model based on FTS, fuzzy C-means clustering, and a multi-factor BPNN for stock market prediction. This approach differs from the H-FSTM in that it uses a GA to optimize fuzzy relations and a BPNN to establish the fuzzy relations between two continuous time series of fuzzy sets.

In fuzzy modeling tasks, Takagi–Sugeno–Kang (TSK) fuzzy systems are well-known techniques that have accomplished promising performances in a variety of applications. Chang and Liu (2008) developed a TSK type fuzzy rule-based system to predict the stock price using technical indicators as inputs. The framework of this TSK fuzzy system was formed using step-wise regression (to select the essential factors), k-means clustering (to partition data into clusters), and a fuzzy inference system (to generate rules and set the parameters). A comprehensive study was conducted by Chang and Fan (2008) in which a novel prediction method that integrates a Haar wavelet transform and a TSK fuzzy rule-based system for forecasting the stock market was proposed. In this study, they also employed k-means clustering to avoid rule explosion and then generated fuzzy rules for each cluster. The K-nearest neighbor (KNN) method with simulated annealing (SA) was also utilized as a sliding window to further tune the prediction results of the TSK model. Moreover, Chang et al. (2016) modeled a TS fuzzy rule-based model in combination with an SVR to forecast stock trading. The SVR was used to support the learning of trading signals in TSK model training. The piece-wise linear representation (PLR) method was also used to select relevant features to improve the predictive power in the proposed approach.

We found three studies that applied adaptive network-based fuzzy inference system ANFIS to predict the stock market. Atsalakis and Valavanis (2009a) presented an ANFIS model premised on two phases, the CON-ANFIS (to control the stock market process model) and the PR-ANFIS (to predict the trend), and achieved promising results challenging the weak form of the EMH. Yang et al. (2011) proposed a fuzzy inter-transaction class association rule (interCAR) mining method based on GNP to forecast the Tokyo Stock Exchange. The idea of this model was to prevent the loss of information during data transformation.

A different path was taken by Wei et al. (2011), who used a correlation matrix to select essential features, a subtractive clustering method to determine the linguistic values for the technical indicator partitions using a data discretization approach, and ANFIS to obtain the linguistic term rules from the technical indicators as well as to optimize the parameters of the FIS using an adaptive network to generate the predictions. A similar type of model based on ANFIS was presented by Kaur et al. (2016). They first used an ordered weighted average (OWA) operator to aggregate high-dimensional data points into a single feature. Then, ANFIS with fuzzy c-mean clustering was applied to produce accessible rules to predict stock index prices.

A distinct fuzzy model can be found in Hadavandi et al. (2010) where the authors attempted to achieve the best stock market predictions by applying an integrated method using genetic fuzzy systems (GFS) and an ANN. In addition, they used step-wise regression analysis (SRA) to find the most influential factors impacting the stock price and a self-organizing map (SOM) neural network to cluster the row data as inputs for the GFS model. Lastly, two other recent studies based on fuzzy theory are discussed. Pal and Kar (2019) proposed a hybrid model combining data discretization, the cumulative probability

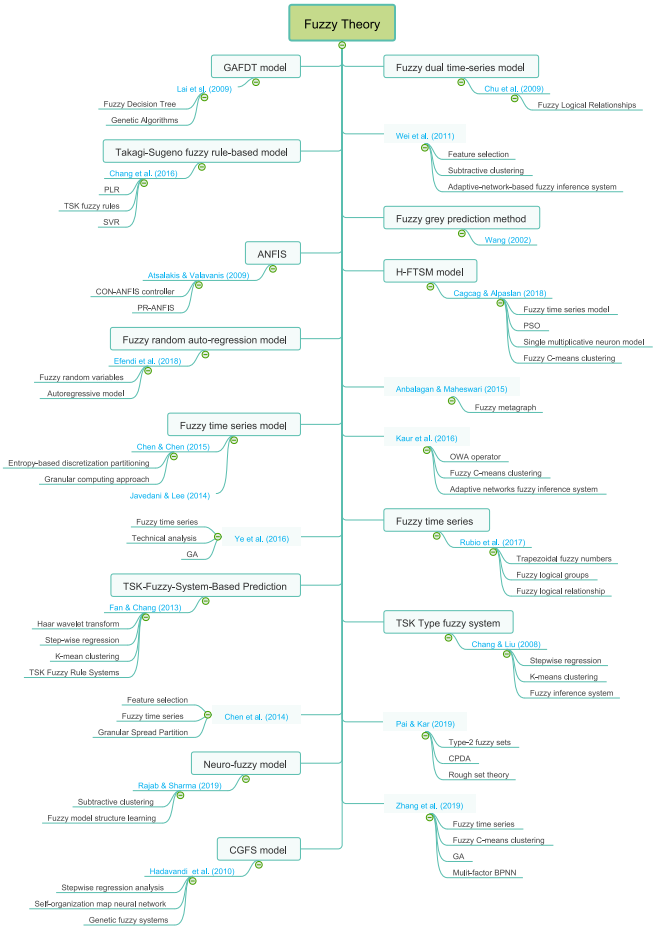

Fig. 32. Studies based on fuzzy theory.

distribution approach (CPDA), and rough set theory. They first applied a modified type-2 fuzzy-set-based methodology to discretize time-series data points to linguistic values, and then CPDA to find the intervals for the linguistic values. Lastly, rule reducts were generated using rough set theory (RGH) to be applied in the prediction step. Another study by Rajab and Sharma (2019) proposed a neuro-fuzzy approach based on a compact rule base and constrained learning for forecasting the stock market. This model was implemented in three essential steps: (1) relevant technical indicators were chosen by using Pearson's correlation coefficient, (2) a fuzzy rule base generation was performed by using subtractive clustering, and (3) the rule base reduction and optimization of the rule base were driven by using fuzzy model structure learning. To present a brief view of fuzzy approaches, a mapping image of all these models is presented in Fig. 32.

5.4.4. Deep learning

It is apparent that, recently, there has been an elevated interest in deep learning-based approaches in the field of stock market prediction (see Fig. A.40). Deep neural networks (DNNs) offer flexibility in modeling using network structures, model parameters, activation functions, as well as various deep learning algorithms. For instance, DNNs have been used for stock market prediction with high-frequency data due to their capability of extracting features from large data sets. Studies using deep learning are outlined in Fig. 33. Next, we shortly go through each of these models.

In our review, Chong et al. (2017) were the first to use a DNN-based systematic approach for stock market analysis and prediction. They examined the influence of three unsupervised feature extraction

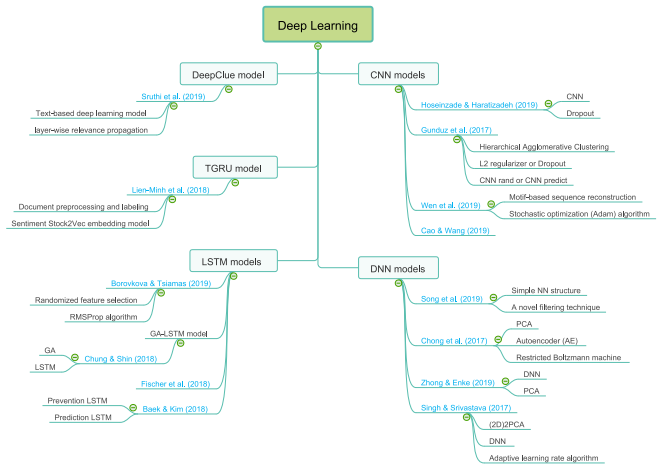

Fig. 33. Studies based on deep learning.

methods, PCA, an auto-encoder, and a restricted Boltzmann machine, on the DNN in the prediction of stock price movements. Singh and Srivastava (2017) demonstrated that the DNN model could improve the forecasting accuracy in stock market predictions. To achieve an effective performance with a DNN, they used 2-directional 2-dimensional PCA ($(2D)^2$PCA) to extract relevant features, and an adaptive learning rate algorithm (ADADELTA) as the optimization technique for learning rate annealing and momentum training. Unlike the radial basis function neural network (RBFNN) and recurrent neural network (RNN) methods, this proposed method achieved an improved accuracy in the prediction of the stock of Google (now: Alphabet Inc.) traded on the NASDAQ stock exchange.

Furthermore, a comprehensive study by Zhong and Enke (2019) used a big data analytic process with a DNN to forecast the daily direction of the SPDR S&P 500 ETF. They detected a pattern with the DNN model using data transformed via PCA, while the number of hidden layers was increased from 12 to 1000. Another distinct approach comes from Song et al. (2019), who studied stock price fluctuations by applying an improved deep learning technique to forecast future stock prices. Instead of using price-based-features, they constructed 715 features from technical indicators to design a DNN model. Then the DNN-based prediction model was implemented using a simple neural network structure and a novel filtering technique (to filter stock prices and identify patterns of fluctuations).

Another research focus in deep learning is the convolutional neural network (CNN), which has been applied extensively in the field of image processing, speech recognition, natural language processing, and other areas. Gunduz et al. (2017) used a CNN approach with specifically ordered input features to forecast intra-day movements of the Borsa Istanbul 100 (BIST 100) index. In the prediction process, the correlation between features and hierarchical agglomerative clustering was used to order the features. At the same time, three different techniques (L2 regularizer, drop-out, and early stopping) were tested to prevent overfitting. They also studied the effectiveness of feature correlations by investigating a correlation-based approach (CNN-corr) and compared it to the randomly ordered features approach (CNN-rand) in the prediction. The results demonstrated that the CNN-corr model was superior in comparison. However, this model was outperformed by the proposed CNN approach presented by Hoseinzade and Haratizadeh (2019) in the prediction of six stock index price movements. The interesting fact with this study was that they extracted relevant features through the developed methods: 2D-CNNpred (two-dimensional representation of features) or 3D-CNNpred (three-dimensional representation of features). They also used the drop-out method for the training of the model.

Cao and Wang (2019) demonstrated that a CNN could deal with categorical and continuous variables in financial forecasting and can achieve effective prediction results. They constructed two models, a CNN and a CNN-SVM, to predict stock index prices and showed that both models were effective according to the empirical results. Moreover, Wen et al. (2019) introduced a novel approach to reconstruct time-series data through frequent patterns (i.e., leveraging motifs) and then applied a CNN model to learn the underlying patterns and reconstructed time-series sequences, supporting the prediction of stock price directions. In this proposed method, the stochastic optimization algorithm (Adam) was applied to train the network.

The long short-term memory (LSTM) network is one of the more advanced deep learning algorithms. According to the literature, Fischer and Krauss (2018) performed the first study on the utility of the LSTM model for stock market prediction. Subsequently, Chung and Shin (2018) integrated it with a GA to optimize the LSTM topology and to determine the time window size to improve the capability of the LSTM network for financial market predictions. Another work by Baek and Kim (2018) concentrated on the problem of over-fitting by proposing a novel forecasting approach relying on an LSTM network. The proposed ModAugNet includes two modules: (1) an over-fitting prevention LSTM module and (2) a prediction LSTM module.

Furthermore, Borovkova and Tsiamas (2019) developed an ensemble framework of LSTM networks to predict intra-day directional movements of the stock price. In addition, a randomized procedure for feature selection and RMSProp as the training algorithm were used in the proposed framework. With a combination of CNN and LSTM, Chen et al. (2019) proposed a stock price trend prediction model (TPM) based on the encoder–decoder mechanism. This proposed method consists of two phases. First, it applied a piece-wise linear regression method (PLR) (that extracted long-term temporal features) and a CNN (that extracted short-term spatial market features) as a dual feature extraction method. Second, a encoder–decoder framework formed by an LSTM was applied to select and merge relevant features and then perform trend prediction.

A novel deep learning approach to predict the stock market using both historical stock prices and financial news data can be found in Lien Minh et al. (2018). In this study, mainly two novel approaches were used. First, a two-stream gated recurrent unit (TGRU) model for stock price trend forecasting; second, a sentiment Stock2Vec embedding model associated with financial news data as well as a sentiment dictionary. Moreover, they initially pre-processed the news article data (from Reuters and Bloomberg) and stock prices through labeling and word embedding. This approach was found to be effective for the prediction of the S&P 500 index price directions and VN-index price trends. Furthermore, Shi et al. (2019) applied a DeepClue model in combination with a pixel-based layer-wise relevance propagation. They used stock-related tweets together with historical price data and financial news articles from Reuters and Bloomberg regarding the US stock market.

5.4.5. Feature selection

Stock market researches are often associated with numerous variables in the form of historical data on stocks, markets, macro-economic variables, and other forms of potentially relevant features. Some stock market studies performed an in-depth analysis of feature relevance/predictor importance to obtain a better understanding those variables that are relevant and, hence crucial for a particular prediction task. Feature selection reduces the set of available variables to only those variables considered relevant for a particular task, hence, often resulting in faster algorithm training, a reduction in over-fitting on account of the absence of irrelevant and/or redundant variables, and a potential improvement of the model accuracy (Guyon & Elisseeff, 2003). Several studies attempted to utilize feature selection methods to achieve a high generalization performance of stock prediction models.

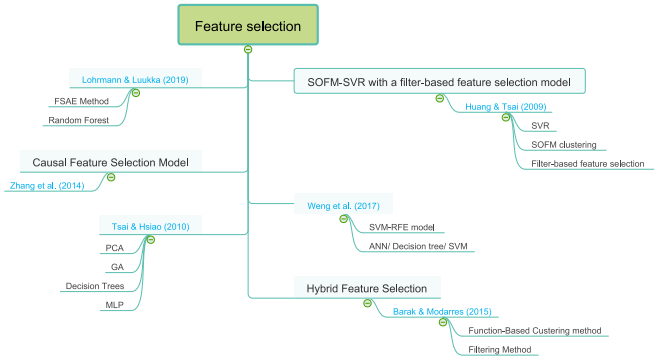

Fig. 34. Studies based on feature selection.

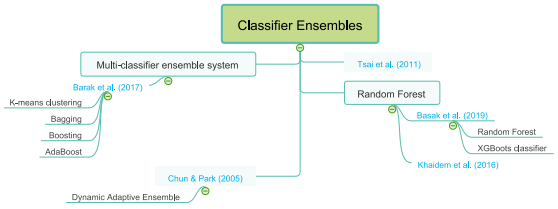

Fig. 35. Studies based on classifier ensembles.

With this review, we found six studies that are mapped with the corresponding techniques in Fig. 34.

Zhang et al. (2014) proposed a novel causal feature selection method to determine the relevant features for effective stock predictions. Huang and Tsai (2009) and Barak and Modarres (2015) use filtering methods for feature selection in their applications. In Huang and Tsai (2009), filter-based feature selection was used to reduce the training time and to improve the accuracy of the self-organizing feature map (SOFM)-SVR model proposed for the prediction of Taiwan index futures. Similarly, Barak and Modarres (2015) selected the most relevant features to forecast risk and return using a filter and function-based clustering methods. Another interesting attempt at stock forecasting using a combination of multiple feature selection methods can be found in Tsai and Hsiao (2010). They first selected PCA,[11] GA and decision trees (CART) as feature selection methods. Then, prediction was performed with ANN to find which of these feature selection methods allows them to achieve better performance in the forecast. They also tested the ability of combinations (by the union, intersection, and multi-intersection methods) of these feature selection methods in the prediction of stock prices. The combination of PCA and GA resulted in the best prediction performance in this study. Lohrmann and Luukka (2019) applied the fuzzy similarity and entropy (FSAE) (Lohrmann et al., 2018) based feature selection method to predict intra-day S&P 500 returns with a random forest. In this study, 136 features related to currencies, commodities, and technical indicators were used as input data, and irrelevant features were filtered out for effective forecasting of the return. Weng et al. (2017) investigated how different online data sources and technical indicators can be used to build an efficient forecasting system applying feature selection. They used the Recursive Feature Elimination (RFE) method with an SVM and an ANN and applied it in forecasting the next day movement of Apple Inc's stock.

5.4.6. Classifier ensembles

In general, an ensemble model is a learning algorithm that combines and trains a set of classifiers and then aggregates the separate votes of their predictions into a single prediction/decision. The stock market studies in our review that used classifier ensembles for their predictions are summarized in Fig. 35. The focus of the ensemble models in this section is on classifier ensembles (supervised learning).

The first investigation of classifier ensembles into the stock market prediction was published by Chun and Park (2005). In their analysis, dynamic adaptive ensemble case-based reasoning (DAE CBR) was proposed as a new learning technique. The key idea of this method was to determine the parameter combinations, optimizing them, and then applying CBR to perform the prediction. Tsai et al. (2011) also examined the predictive performance of classifier ensembles for stock return predictions. They considered two types of ensembles: (1) homogeneous classifier ensembles (using an ensemble of neural networks) and (2) heterogeneous classifier ensembles (deploying an ensemble of decision trees, logistic regression, and neural networks). They also considered hybrid methods of majority voting and bagging.

A comprehensive study by Barak et al. (2017) proposed a fusion model (called multi-classifier ensemble system) that used multiple diverse base classifiers for stock market predictions. In the fusion framework, various classifiers were operated together with the Bagging, Boosting, and AdaBoost models to get a common output. Before performing the classification, the data was clustered using k-mean clustering to select the best combination of classifiers from KNN, MLP, random forest, decision tree, SVM, decision tree naive Bayes (DTNB), LAD Tree, BF tree, CART tree, Rep tree, the Bayes, to mention a few.

According to the literature, the random forest model that belongs to the class of ensemble learning methods was also used to create financial forecasting models. Examples include (Khaidem et al., 2016), who introduced an improved methodology based on a random forest to predict the direction of the stock price, and the latest study by Basak et al. (2019), which utilized a random forest to forecast the direction of stock prices (Facebook and Apple Inc.) and compared its predictive performance with the XGBoost classifier. In addition, Lohrmann and Luukka (2019) used a random forest for intra-day return predictions of the S&P 500 index.

5.4.7. Other models

In this section, we discuss other forecasting models that were left outside the previous categorization since these methods were not commonly used in the reviewed literature. Such methods include, but are not limited to, KNN, Bayesian networks, and clustering. All of the techniques discussed in this section are mapped in Fig. 36.

As the oldest study in our literature review, Leung et al. (2000) examined the predictability of the movements of a stock index price using multivariate classification techniques and level estimation approaches. They tested probability-based classifiers, including linear discriminant analysis and logit, probit, and probabilistic neural networks in comparison to level estimation methods such as exponential smoothing, a multivariate transfer function, vector auto-regression with a Kalman filter, and a multilayered feed-forward neural network. Moreover, Enke et al. (2011) proposed a three-stage stock market forecasting model in which a multiple regression analysis (that defined the relevant financial and economic variables), differential evolution-based type-2 fuzzy clustering (that created the prediction model), and a fuzzy type-2 neural network (that performed the prediction) were used. Son et al. (2012) conducted a study on the prediction of high-frequency (five minute periodicity) KOSPI 200 index data using linear regression, logistic regression, an ANN, and an SVM, all with and without dimensionality reduction. The corresponding results indicated the superiority of the SVM model in the prediction compared with the other methods. Similarly, Patel et al. (2015) compared the predictive power of four classification methods, an ANN, an SVM, a random forest, and a Naive-Bayes classifier, on the Indian stock market.

In the context of this literature review, clustering was rarely used in financial market predictions. In particular, there were only two relevant

[11] Usually PCA is considered as a feature extraction method in general.

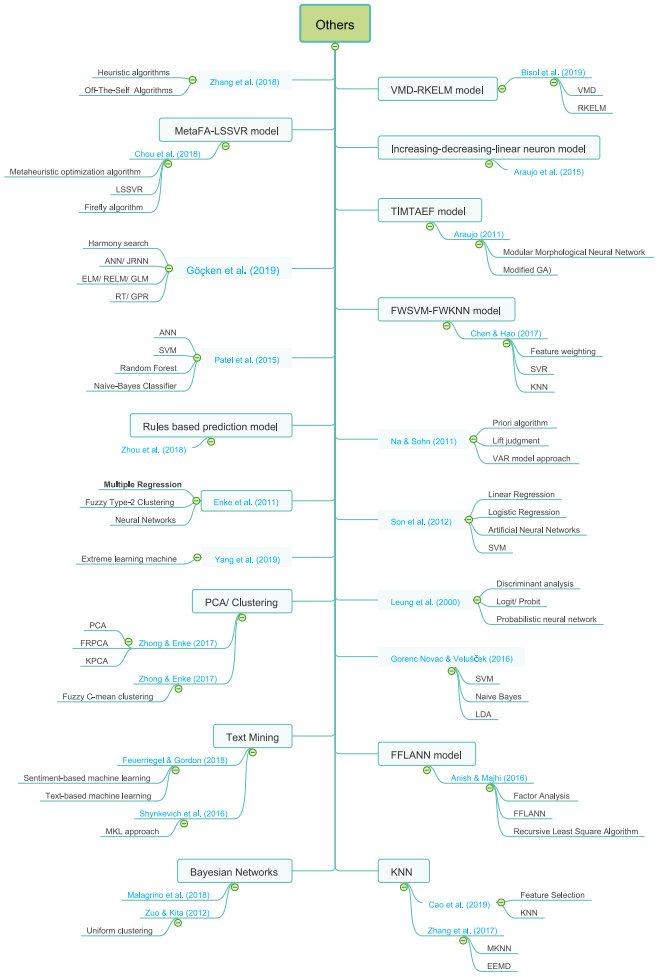

Fig. 36. Studies that used other machine learning techniques.

studies in the set of articles for this study. The first study by Vilela et al. (2019) proposed a two-stage forecasting model using K-means or fuzzy C-means clustering to segment the data in the first stage and SVR in the second stage to forecast the future stock price. The second study by Zhong and Enke (2017a) also used a clustering-based prediction model to forecast the daily direction of the stock return. They applied fuzzy C-means clustering to prepare the data. Subsequently, they applied PCA to each of the clusters and the entire data set and then combined these data. With this aggregated data set, the prediction was performed using ANNs and logistic regression. It was pointed out that PCA+ANNs was able to accomplish the high accuracy in the forecast. Using the same data, they also presented another study in Zhong and Enke (2017b) that compared the effectiveness of three-dimensionality reduction techniques (PCA, fuzzy robust PCA, and kernel-based PCA) in the prediction of the daily direction of the stock return. As in their first paper, they also found in the second publication that ANNs with a PCA approach outperformed the benchmark algorithms in the forecast.

Another comparably rare approach in the stock market prediction literature is the Hidden Markov Model (HMM), which is a broadly applied tool to forecast time-series data. Hassan et al. (2007) proposed a fusion model that integrates a Hidden Markov Model (HMM), a GA, and an ANN for stock market forecasting. In the proposed model, the GA is used to optimize the initial parameters in the HMM, while the ANN is deployed to transform the daily stock prices to independent sets of values as the inputs to the HMM. Then, the optimized HMM is used to predict a one-day ahead stock price. Gorenc Novak and Velušček (2016) presented a stock price prediction strategy based on daily high prices using an SVM, an LDA, and an NB classifier. They specifically developed simple trading strategies (called Guided D-threshold trading strategies), which guided the classifiers to increase the performance.

The K-nearest neighbor (KNN) classifier is among the simplest and most widely applied machine learning algorithms in classification. Zhang et al. (2017) proposed a new multi-dimensional k-nearest neighbor (MKNN) model-based ensemble empirical mode decomposition (EEMD) approach to predict the high and closing prices of stocks simultaneously. Within this study, the combination of the MKNN and EEMD (EEMD-MKNN model) was efficient and performed well in comparison to an EMD-KNN, a KNN, and an ARIMA model in the prediction of the prices of four well-known stocks. In addition, the study by Cao et al. (2019) applied a KNN model for forecasting next day stock price patterns. They used feature selection based on the Kruskal–Wallis test and a novel pattern network construction model to extract patterns from the data. Along with the selected features, the KNN and an SVM were used for the prediction.

Malagrino et al. (2018) introduced a Bayesian network model for forecasting the next day direction of the Brazilian IBOVESPA stock index price. In addition, Zuo and Kita (2012) investigated the use of a Bayesian network in the prediction of stock prices. They first discretized the continuous P/E ratio data through clustering, and the prediction was then performed using a Bayesian network.

Recently, there has been a considerable interest in financial market prediction systems using information extracted from financial news articles. For instance, Shynkevich et al. (2016) developed a multiple kernel approach (MKL) to predict stock price movements using information from various financial news categories (different kernels for each news category). They tested the proposed model for large cap stocks from the “Health Care” sector (in the S&P 500). They illustrated that growing the number of relevant news categories as the input data for the financial prediction significantly improves the performance of the forecasting model. Another attempt at forecasting a stock index based on text mining can be found in Feuerriegel and Gordon (2018). Their investigation was based on 75,927 ad hoc announcements from companies as data sources, and it highlighted that text-based machine learning methods succeeded in improving forecasting accuracy compared with the benchmarks used in this study.

The study by Araújo et al. (2015) presented a hybrid forecasting model named increasing decreasing linear neuron (IDLN) to overcome the random walk dilemma in high-frequency financial data forecasting. This IDLN model consists of a series of linear operators (e.g., finite impulse response), nonlinear operators (dilation and erosion morphological), and nonlinear decreasing operators (anti-dilation and anti-erosion morphological). Another work links email communication within a company to the company’s stock price (Zhou et al., 2018). These authors developed a novel data mining model based on internal communication patterns to predict the stock price of the company. Their study revealed that there was a significant impact of corporate communication on stock price. A recent study by Yang et al. (2019) suggested a novel hybrid stock selection model that deploys an extreme learning machine (ELM). In essence, this model consisted of two steps: stock prediction (that predicted future stock returns) and stock scoring (that selected highly valued stocks to generate an equally weighted portfolio).

Based on the time-series data of numerous stock market indices in the world, Na and Sohn (2011) developed a forecasting approach using an associated rules analysis to predict the change in KOSPI prices. They used the Apriori algorithm to generate the associated rules and lift judgment (Wang et al., 2004) to reduce possible biases in the generated rules. A vector auto-regression (VAR) model was also used to identify the relationship between the Korean and the US stock markets. With the proposed method, their experiment achieved accurate predictions of the changes in the KOSPI. To eliminate the random walk (RW) dilemma for financial market prediction, Araújo

(2011) presented a novel method called translation-invariant morphological time-lag added evolutionary forecasting (TIMTAEF). This hybrid approach was a combination of a modular morphological neural networks (MMNN) and a modified GA (MGA). Furthermore, Anish and Majhi (2016) proposed a feedback type of the functional link artificial neural network (FLANN) with recursive least squares (RLS) training as a novel forecasting approach. This model first used factor analysis to identify relevant features, and then RLS was performed during training with the FFLANN to reduce the training time. Chen and Hao (2017) introduced a novel hybrid FWSVM-FWKNN (feature weighted SVM and feature weighted KNN) model to forecast Chinese stock indices. In this framework, each feature's relative importance was estimated by first computing information gain, and then the weight was set for each feature. Subsequently, these weights were used to calculate the inner products of kernel functions in SVM to forecast the direction of the stock price movement. Lastly, the weights were recalculated and used to measure the Euclidean distance in KNN to predict the stock index price.

A different path was taken by Zhang et al. (2018a), who proposed a novel stock trend prediction system based on an unsupervised heuristic algorithm to forecast the price movement and its growth rate intervals within the predefined prediction time horizon across a predefined prediction duration. This heuristic algorithm was utilized in this context to classify row transaction data into four classes, Up, Down, Flat, and Unknown, within the predefined fixed time interval. Then, prediction was performed using a random forest and decision trees on the Weka platform.

To forecast stock prices in the Taiwan stock market, Chou and Nguyen (2018) developed a prediction system based on a sliding-window meta-heuristic optimization that included a meta-heuristic firefly algorithm (FA) and a least-squares SVR (LSSVR). From the empirical analysis, they recommended to use their model for highly non-linear-time-series data whose patterns are hard to detect using traditional methods. Recent research by Bisoi et al. (2019) proposed a predictive model that can predict both stock prices and stock price movements. This method first uses variational mode decomposition (VMD) to decompose the data and, afterwards, a robust kernel-based extreme learning machine (RKELM) to perform the day ahead price prediction. Lastly, we discussed the study by Göçken et al. (2019), where seven hybrid soft computing models were tested to forecast the stock market. Each of these models was used in conjunction with HS to optimize the parameters. The sef of models included an HS-NN, an HS-JRNN (Jordan recurrent NN), an HS-ELM, an HS-RELM (recurrent extreme learning machine), an HS-GLM (generalized linear model), an HS-RT (regression tree), and an HS-GPR (Gaussian process regression).

5.4.8. Distribution of applied machine learning techniques by year

Fig. 37 illustrates the distribution of the application of various machine learning techniques in stock market prediction studies by year. The earliest work in our review was conducted in 2000 using an ANN. In 2001, empirical works deploying SVMs appeared. It is noteworthy that these two methods were the most popular approaches for stock market predictions in our review throughout the rest of the years. In addition, a sharp increase in the application of deep learning for stock market prediction can be observed since 2017, and the highest number of articles (eight to be exact) that applied deep learning was published in 2019. In summary, the approaches referred to the most in recent attempts on financial market forecasting were ANNs, SVMs, fuzzy theory, deep learning, and feature selection.

5.4.9. Optimization and feature analysis techniques

In Table 8, the main optimization and feature operation techniques, together with the corresponding references in which these methods were applied, can be found. In this context, the term "optimization" essentially refers to optimizing the parameters and the structure of a model, whereas "feature operation" refers to the selection, generation, extraction, denoising, and discretization of features. It is apparent from the table that the GA was the most widely used technique for optimization as well as for feature selection. To extract relevant features, PCA was extensively applied as a dimensionality reduction method, while wavelet transform was deployed to denoise features.

5.4.10. Review of validation methods

Proper validation is a crucial aspect of machine learning, which is also applicable in the context of stock market prediction studies. Surprisingly, only 49 articles in the selected literature mentioned (explicitly) that they validated their forecasting approach using a specific validation technique. Fig. 38 displays the frequencies of each validation method implemented in the reviewed articles. In the figure, "CV" refers to cross-validation.

The results in the figure indicate that the "Holdout method" and "5-fold CV" are the commonly used validation methods in stock market prediction studies. The "Holdout method" and "5-fold CV' were applied in 20% (10) and 18% (9) studies, respectively for validation of the used predictive models. The "Walk-forward method". which is often used in time-series analyses, was ranked third among the preferred validation methods in 10% (5) studies. Notice that some studies (see, e.g., Chang & Fan, 2008; Mo & Wang, 2018) mentioned that they applied "Cross-validation", but did not name the cross-validation technique they used explicitly.

5.5. Review of performance metrics

Evaluating the predictive performance of machine learning methods has been a crucial part of all studies in our review collection, since these are the measures based on which the performance of each algorithm is evaluated and benchmarked with. Fig. 39[12] shows the frequency of different types of performance metrics from each study. It is noteworthy that some of these metrics did not evaluate prediction performance. All metrics were classified into one of the four categories: "Accuracy-base", "Error-based", "Return-based", and "Statistical tests".

Fig. 39 highlights that the root mean square error (RMSE) as an error-based metric was the most frequently used evaluation metric for measuring the forecasting performance of a model. Considering error-based measures, the mean absolute percentage error (MAPE) was the second most used measure for model evaluation in this category. Other widely applied metrics in this category are the mean absolute error (MAE) and the mean squared error (MSE), appearing 29 and 16 times in the literature, respectively. Accuracy as an evaluation metric was deployed in 43 studies, which ranks it second to the RMSE among all evaluation metrics. Most of the studies in our review were based on classification-based prediction tasks. To show the prediction performance concerning the number correct true positive classifications, *hit ratio* was used as an evaluation metric in 19 studies. In addition, there were 13 studies that used the *sharp ratio* to present the performance of the return-based predictions in a risk-adjusted way.

Our review also identified several efforts to deploy statistical tests to determine whether the proposed method yielded statistically significant results compared to the benchmarks in the prediction. In such studies, a paired t-test was the most widely used test found in 17 studies, while Kruskal Wallis, Wilcoxon, and other tests were only implemented in a few cases. Moreover, for return-based predictions, relevant measures such as the average return, standard deviation of the daily return, rate of return (ROR), and return on investment (ROI) were reported as the results.

[12] The labels for some evaluation metrics are displayed in abbreviated form but the complete names of these metrics can be found in Table B.1 in Appendix.

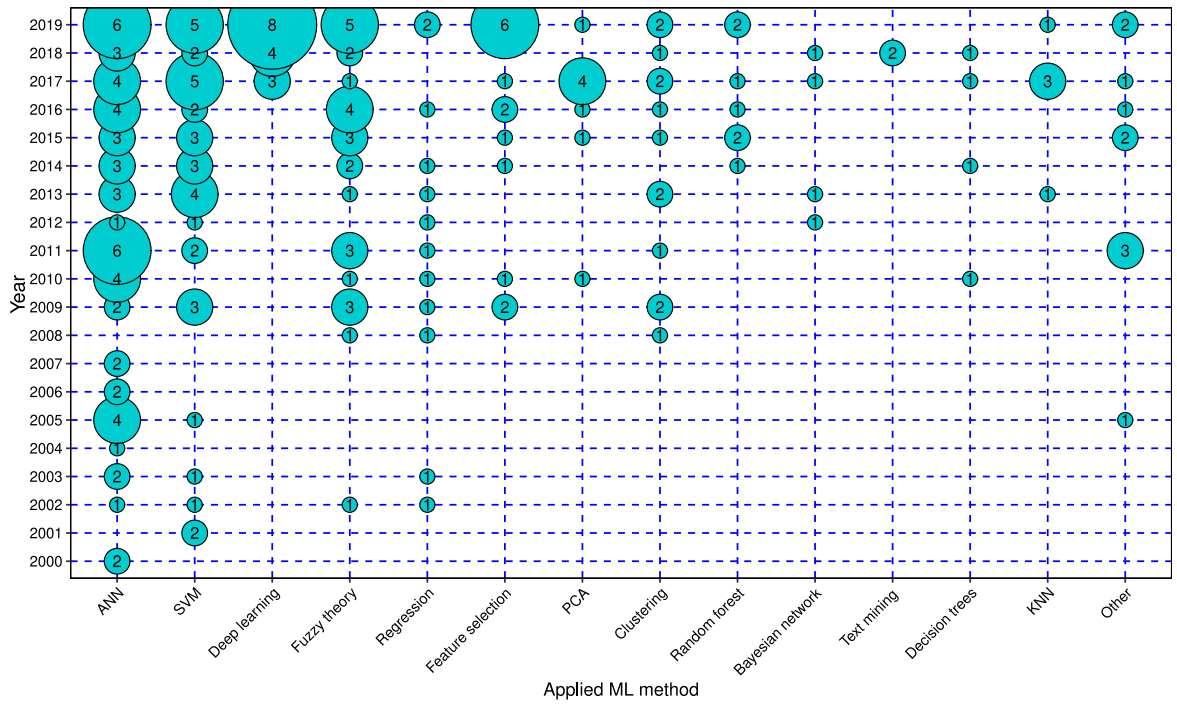

Fig. 37. Yearly distribution of the applied machine learning approaches.

Table 8
Main optimization and feature operation methods.

Method (counts)	Model optimization	Feature operations
GA (18)	Araújo (2011), Armano et al. (2005), Chai et al. (2015), Chen et al. (2017), Chung and Shin (2018), Ebadati and Mortazavi (2018), Gocken et al. (2016), Hassan et al. (2007), Lai et al. (2009), Pan et al. (2017), Qiu and Song (2016), Qiu et al. (2016), Ye et al. (2016), Yu et al. (2009), Zhang, Zhang et al. (2019)	Kim and Han (2000), Tsai and Hsiao (2010), Yu et al. (2009)
PCA (8)		Anish and Majhi (2016), Chong et al. (2017), Singh and Srivastava (2017), Tsai and Hsiao (2010), Wang and Wang (2015), Zhong and Enke (2017a, 2017b, 2019)
Wavelet transform (6)		Anish and Majhi (2016), Bisoi et al. (2019), Chang and Fan (2008), Chiang et al. (2016), Kao et al. (2013), Wang et al. (2011)
PSO (5)	Cagcag Yolcu and Alpaslan (2018), Chai et al. (2015), Chiang et al. (2016), Lu (2013), Rustam and Kintandani (2019)	
Grid search (4)	Chai et al. (2015), Chen and Hao (2017, 2018), Lu et al. (2009)	
Simulated annealing (3)	Chang and Fan (2008), Chang and Liu (2008), Qiu et al. (2016)	
Firefly algorithm (3)	Kazem et al. (2013), Xiong et al. (2014), Zhang, Teng, and Chen (2019)	
HS (2)	Göçken et al. (2019), Gocken et al. (2016)	
GNP (2)	Ramezanian et al. (2019), Yang et al. (2011)	
Bayesian regularization (2)	Selvamuthu et al. (2019), Ticknor (2013)	
Self-organizing map (2)		Hadavandi et al. (2010), Huang and Tsai (2009)
EMD (2)		Chai et al. (2015), Zhou et al. (2019)

5.6. Machine learning approaches by performances

There were some challenges related to the results of different studies were presented, which make them challenging to compare. For instance, some studies offered prediction results with the proposed/used machine learning model for several different time horizons (see, e.g., Chen & Hao, 2017; Das & Padhy, 2018; Pan et al., 2017), for various stock indices and stocks (see, e.g., Borovkova & Tsiamas, 2019; Chiang et al., 2016; Hoseinzade & Haratizadeh, 2019), for different stages of the applied method (see e.g., Nermend & Alsakaa, 2017; Tsai & Hsiao, 2010; Yu et al., 2009), for each class separately (see, e.g., Chang et al., 2009; Lohrmann & Luukka, 2019; Zhang et al., 2018a), and with other different settings (see, e.g., Hsieh et al., 2011; Rustam & Kintandani, 2019; Yang et al., 2019; Zhang, Shao et al., 2019). Moreover, some studies (see, e.g. Chang & Wu, 2015; Ebadati & Mortazavi, 2018; Liao & Wang, 2010; Na & Sohn, 2011; Wang, 2002) implemented their performance evaluation using very different approaches that are not common in machine learning applications. Finally, most of the selected studies have used more than three performance metrics from numerous types, as demonstrated in Fig. 39.

Even though such issues made the presentation of the results difficult in entirely, we picked the articles with the best prediction performance (highest accuracy or lowest error) for each machine learning approach. Table 9 summarizes those studies by the machine learning approaches ANN, SVM/ SVR, fuzzy theory, deep learning, feature selection, and other models.

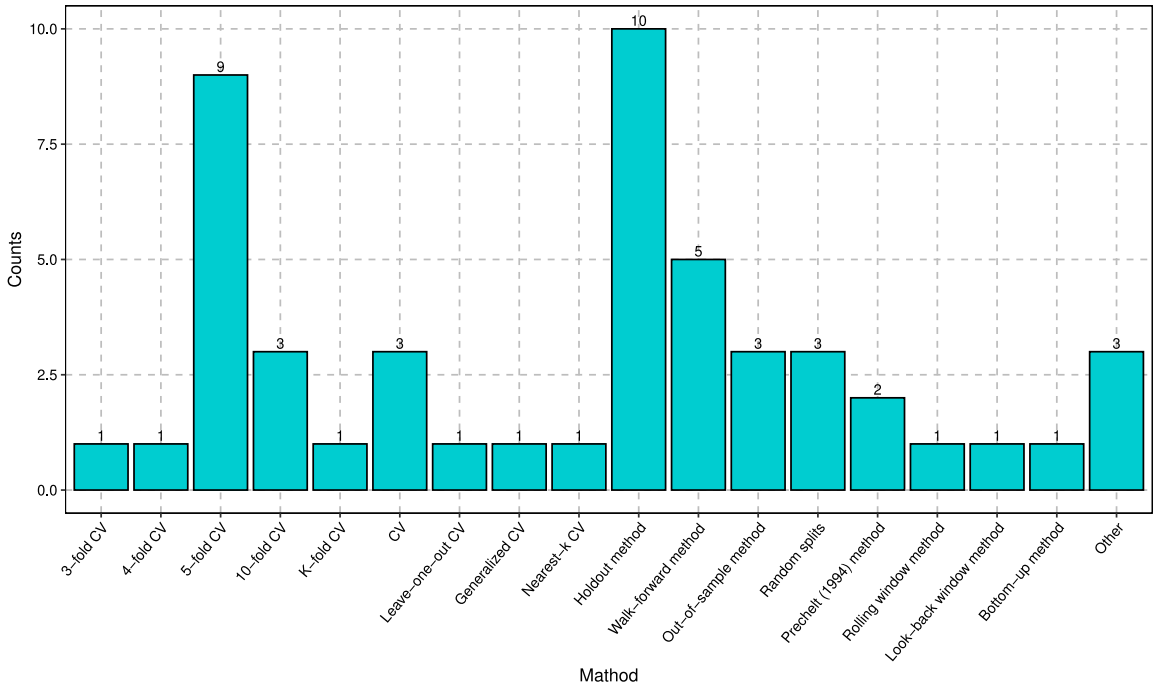

Fig. 38. Validation techniques used in the reviewed articles.

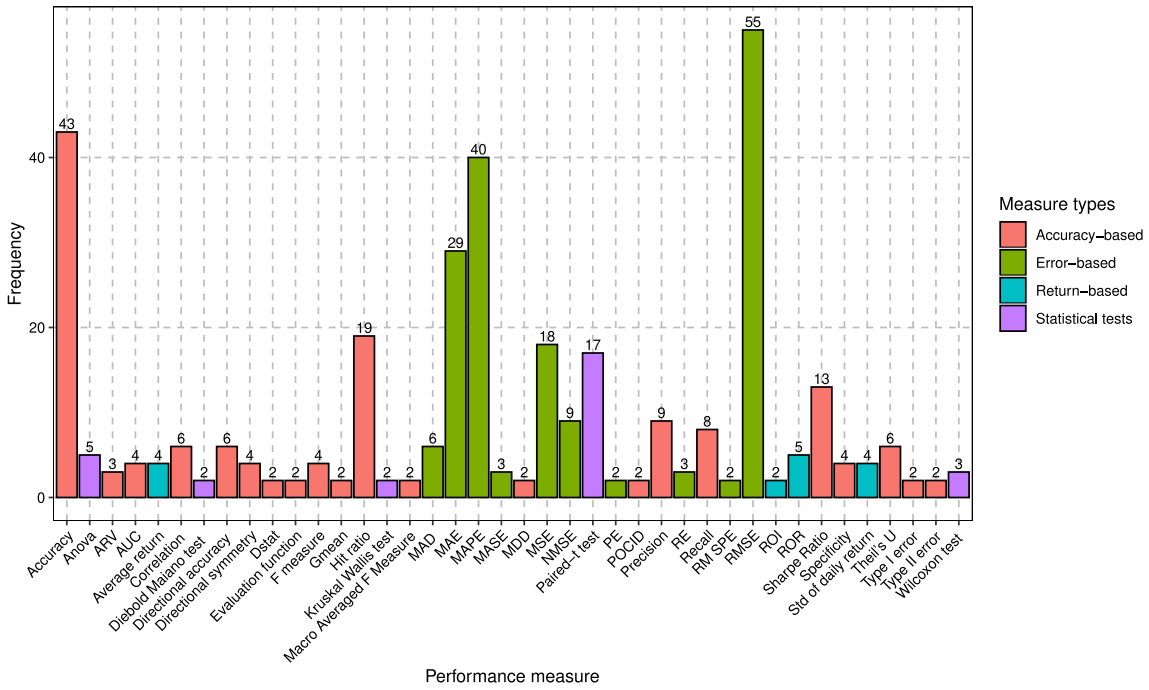

Fig. 39. Performance metrics used in the evaluation of the forecasting performance of the proposed models.

5.7. A review of most recent related studies

The focus of our study was on the related articles published in the last two decades. However, the most recent studies were not included since we concentrated our research on the period between 2000 and 2019. Hence, this sub-section attempts to briefly review the most recent and relevant articles found from the databases we utilized in this research. Table 10 summarizes recent prominent works related to our research that employed machine learning techniques for stock price or return predictions. Among the various machine learning methods utilized for stock market forecasting, there has been a rapid rise in the use of deep learning techniques in the last two years. This significant growth is visible in Figs. A.40 and 37. Another interesting fact is that LSTM networks have witnessed a growing interest in stock market predictions among the variations of deep learning models.

According to Liu and Long (2020), a combination of empirical wavelet transforms (EWT), dpLSTM (dropout LSTM), PSO, and outlier robust extreme learning machine (ORELM) is significantly effective than BPNN (Sun & Gao, 2015), LSTM (Horata et al., 2013), and RF (Basak et al., 2019) models in forecasting indices in the USA and Chinese markets. Jiang et al. (2020) also proposed a stacking framework by integrating tree-based methods (RF, extremely randomized

Table 9
Selected studies with the highest performance for each machine learning approach.

Main approach	Study	Target	Input data	Benchmarks	Measure	Result
ANN	Hu et al. (2018)	S&P 500, DJIA	Indices data, Google trends	BPNN, GWO-BPNN, PSO-BPNN, WOA-BPNN, SCA-BPNN	Hit ratio	86.81% (S&P 500), 88.98% (DJIA)
	Qiu et al. (2016)	Nikkei 225 index	Technical indicators	GA feature discretization, SVM, BPNN, classification model	Hit ratio	81.27%
SVM/SVR	Sedighi et al. (2019)	DOW 30, NASDAQ 100, S&P 500	Technical indicators	19 benchmarks	RMSE	0.0092 (average over all indices)
	Zhang, Teng, and Chen (2019)	6 securities from SSE	Closing prices	LSTM model and 11 more	RMSE	1.62e−06 (1), 4.33e−06 (2), 0.000420(3), 1.07e−05(4), 0.005916(5), 0.003501(6)
Fuzzy theory	Zhang, Zhang et al. (2019)	SSECI, TAIEX	Closing prices	FCM, PSO, and 9 more	RMSE	1.663 (SSECI), 1.2170 (TAIEX)
	Chang and Liu (2008)	TSE index, MediaTek Inc	Technical indicators	BPN, MRA	Accuracy	97.6% (TSE index) and 98.08% (MediaTek)
Deep Learning	Lien Minh et al. (2018)	S&P 500	Financial news, index prices	GRU, LSTM with and TGRU Glove and Word2Vec with Stock2Vec	Accuracy	66.32%
	Singh and Srivastava (2017)	NASDAQ	Technical indicators	RBFNN+(2D)2PCA, RNN+(2D)2PCA	RMSE	1.01%
Feature selection	Weng et al. (2017)	AAPL stock	Market data, technical indicators, Wikipedia traffic, Google news counts	7 cases with different data sources	Accuracy	85.8%
	Barak and Modarres (2015)	TSE return	Financial ratios and fundamental index	BF tree, LAD tree, and 14 more	Accuracy	80.24%
Other	Zhou et al. (2018)	Enron stock	Stock price & emails data	Decision tree	Accuracy	86.67%

Table 10
A review of most recent related articles published during 2020 and 2021.

Study	Target	Prediction type	Input data	Approach	Model specifications	Benchmarks	Performance metrics
Liu and Long (2020)	S&P 500, CMSB, DJI	Price	Closing prices	Deep learning	EWT + dpLSTM (dropout LSTM) +PSO + ORELM	BPNN, LSTM, RF	RMSE, MAPE, MAE, SDE
Shen and Shafiq (2020)	Chinese stock market	Trend (up/down)	Closing prices, fundamental data	Deep learning	LSTM + feature expansion + REF + PCA	SVM, MLP,NB,RF,LR, ARIMA	Accuracy, F1 score, TPR, TNR
Jiang et al. (2020)	S&P 500, DOW 30, Nasdaq	Movement (1/0)	Technical and macro-economic variables	Other (Stacking method)	(RF, ERT, XGBoost, LightGBM) + (RNN, bidirectional RNN, RNN-LSTM and GRU) + LR- or Lasso-based meta classifier	RF, ERT, XGBoost, LightGBM, RNN, BRNN, LSTM, GRU	Accuracy, precision, recall, F score, and AUC
Yuan, Yuan et al. (2020)	Chinese A-share market	Trend (+1/−1)	Technical, economic, and fundamental variables	Other	(SVM, RF, ANN) + feature selection	Without feature selection	Accuracy, sharp ratio, annualized return
Li et al. (2020)	CSI 300 stock index	Price	Internet messages, open & close prices	Deep learning	LSTM + corpus-based approach + Näive Bayes	SVM, LR, Näive Bayes	Accuracy, precision, recall, F1 score
Lu et al. (2020)	Shanghai Composite Index	Price	Basic technical variables	Deep learning	CNN+LSTM	MLP, CNN, RNN, LSTM, CNN-RNN	MAE, RMSE, R^2
Lee and Kim (2020)	S&P 500 index, KOSPI200, FTSE100	Price	Prices of indices	Deep learning	ConvLSTM + CNN + LSTM	CNN, LSTM, SingleNet, and SMA	MSE, MAPE, MAE
Kim et al. (2020)	S&P 500 index	Direction	Prices of 55 stocks	Other	(LR, MLP, RF, XGBoost, LSTM) + ETE	Without ETE	Accuracy and adjusted accuracy
Carta et al. (2021b)	S&P 500 index	Movement (high/low)	Daily news and time series	Classifier ensembles	Decision tree + lexicon generation + feature extraction	MLP, Gradient Boosting, RF, and other two	Accuracy, precision, recall, F1 score
Yun et al. (2021)	KOSPI 200	Direction	Technical indicators	Classifier ensembles	GA + XGBoost	RF, Decision Trees, SVM-RBF, KNN, and LSTM	Accuracy, precision, recall, F1 score
Jing et al. (2021)	Five stocks in SSE	Price	Stock prices and textual data	Deep learning	CNN + LSTM	SVR, GA-SVR, CNN, GA-SVR, LSTM	MAPE
Wang et al. (2021)	S&P 500, DJI, SSE, NYSE, N225, FTSE, NASDAQ	Price	Prices of indices	Deep learning	Reservoir computing (RC) model	LSTM, RNN, and EMD2FNN	RMSE, MAE, MAPE, R^2

trees [ERT], XGBoost, and light gradient boosting machine [LightGBM]), deep learning methods (RNN, bidirectional RNN, RNN-LSTM and gated recurrent unit [GRU]), and logistic regression (LR)- or Lasso-based meta classifiers for stock index prediction. Moreover, two special deep learning techniques, LSTM and CNN, have been frequently applied in recent studies (Jing et al., 2021; Lee & Kim, 2020; Lu et al., 2020), demonstrating that their mixture works more effective than each of them alone.

Some other studies employed different data processing and classification algorithms to enhance the prediction performance of LSTM networks for the selected market. For instance, Shen and Shafiq (2020) proposed a deep learning approach based on LSTM, feature expansion, recursive feature elimination (RFE), and PCA to predict the Chinese stock market. Their results indicate that the proposed method performs better than other forecasting models, such as SVM, MLP, RF, LR, and ARIMA, achieving the highest accuracy of 93%. Liu and

Long (2020) examined the performance of LSTM networks in predicting the CSI 300 stock index price through messages data from internet stock message boards. For sentiment extraction, they employed a corpus-based approach and the Naive-Bayes classifier as text mining techniques. This proposed model achieved an accuracy of 80.2% in the prediction, outperforming all baseline models used. From a different perspective, a recent study by Wang et al. (2021) developed a novel deep learning model of reservoir computing through random, small-world, and scale-free networks to examine predictive performance in the following day predictions of the seven major stocks selected. The result of the study shows that the developed approach outperformed the most frequently used deep learning techniques, including LSTM, RNN, and EMD2FNN (Zhou et al., 2019), achieving the lowest RMSE in the prediction.

Apart from that, Kim et al. (2020) investigated the effectiveness of the use of effective transfer entropy (ETE) with well-known machine learning techniques, including LR, MLP, RF, XGBoost, and LSTM, to forecast the direction of the S&P 500 index. Similarly, Yuan, Yuan et al. (2020) compared the predictive capability of SVM, RF, and ANN models with and without feature selection in predicting the Chinese A-share market. The best performance (52.75%) accuracy was reported with the RF–RF (random forest for feature selection + prediction) model in the experiment. As for classifier ensemble approaches, Carta et al. (2021b) proposed a decision tree-based machine learning approach (with news aggregation, lexicon generation, and feature extraction) for stock market forecasting. Yun et al. (2021) also presented a hybrid approach based on GA and XGBoost techniques to predict the direction of the KOSPI 200 index. During the experiment, they specifically introduced a new feature engineering process comprising feature expansion, data preparation, and feature selection. The proposed method achieved the best accuracy of 93.8% compared with benchmark approaches.

5.8. Recent progress

The findings from our review confirm an increasing interest in the stock market prediction research area over the last five years. As evident from Figs. 37 and A.40, the main trend in recent studies is the growing use of deep learning methods, for instance LSTM, CNN and RNN. Moreover, as shown in 10, the use of classifier ensembles instead of single classifiers also seems to be common in recent research works. However, ANN- and SVM- based approaches can still be found in the literature, but they appear to lose an increasing share to deep learning and classifier ensembles.

Moreover, over the last five years, the application of feature selection and feature extraction methods, which are used in the data pre-processing step and, thus, are methods that support setting up simpler, more interpretable, and potentially more accurate models, has increased. Due to their benefits such as a decrease in computational expense and model complexity, removal of variables that act as noise, and potential improvements in model performance, It can be expected that these methods will solidify their popularity in future studies. From the data perspective, technical, fundamental, and historical price data-based studies are still common, but rising interest in the use of the textual data extracted from different sources, for example, financial news and social media data, is apparent. In particular, a special focus appears to be placed on sentiment analysis (text mining) for internet messages/message boards, news, and tweets data for model building.

6. Study limitations

The focus of our review was on supervised and unsupervised machine learning methods. A minority of articles may have included statistical forecasting methods such as ARIMA or GARCH (e.g., as benchmarks for the proposed method), but these did not represent the emphasis in our review. In addiction, multi-criteria decision making (MCDM) methods, which are also often encountered in the financial forecasting literature, were excluded from our study. Finally, the focus on machine learning also meant that many other soft-computing methods, e.g., for regression analysis, were also excluded from our study. Another potential limitation was that we focused exclusively on journal papers in the search result, and we excluded other sources of studies such as conference papers, peer-reviewed workshop publications, and technical reports. However, we acknowledge that it was possible to miss relevant articles because of the selection of the search terms used in our study.

7. Discussion and conclusions about the review

In this paper, a systematic review and analysis of the machine learning literature was conducted for stock market prediction. We analyzed 138 journal articles published between 2000 and 2019. From the empirical evidence obtained, we identified highly cited articles, the yearly distribution of keywords, and other useful information. Next, we explored the dependent and independent variables of the financial data sets in the set of journal articles in detail.

The investigation of the markets, indices, and stocks in the data revealed that indices and stocks in the USA were the most investigated ones, first and foremost, the S&P 500 index. However, on the regional level, Asia was investigated the most. In particular, Taiwan, China and South Korea together were more often considered than the USA. Hence, it is also unsurprising that the TAIEX (Taiwan) was the second most investigated index in our review. In addition, we found a moderate correlation between the frequency with which a stock index was considered in the literature and its market capitalization, indicating that larger indices tend to be covered more often in the scientific literature.

From an individual stock point of view, stocks linked to health care, information technology, and consumer discretionary were most frequently found in the literature. Moreover, we reviewed 2173 unique variables which were used in the selected literature. The largest type by the number of variables was “Technical Indicator” with 1348 variables. We found that, especially for lower dimensional data with less than 20 variables, the average share of technical indicators among all variable types was, with approximately 80% clearly the largest. However, there was a tendency that with a larger variable set, the share of “Macro-Economy” and “Fundamental Indicator” specific variables was increasing. Studies exceeding 500 variables were, in our review, focused on the bag of words approach from text mining, where relevant keywords are extracted from, e.g., financial news or tweet data.

Our analysis showed that the most frequently used machine learning-based prediction models for stock market forecasting were based on the ANN, the SVM, and fuzzy theory. This finding is somewhat different from the previous evidence presented by Henrique et al. (2019), as these authors had not reported on fuzzy theory-based methods. As another finding, we identified that deep learning techniques have received much attention in the last three years, appearing in 15 related articles. In addition, fuzzy-set-theory-based approaches have clearly more attention in the last decade compared with the previous one. In addition, GAs were mostly used when efforts were made to improve the parameter optimization for machine learning methods. For feature extraction methods, PCA and wavelet transform were the most popular methods.

From the review of the most recent studies, we found that the rapid rise in the use of deep learning techniques has continued even in 2020 and 2021 years. All deep learning-based papers found in this new analysis have applied improved LSTM models to predict stock market variables. Based on this analysis, we can conclude that LSTM networks have been much more effective than the other forms of deep learning (for example, CNN and RNN), showing more robust predictions.

In summary, we provide the following contributions for future research:

(1) A systematic literature review and bibliometric analysis of the recent literature on machine learning in stock market forecasting, (2)

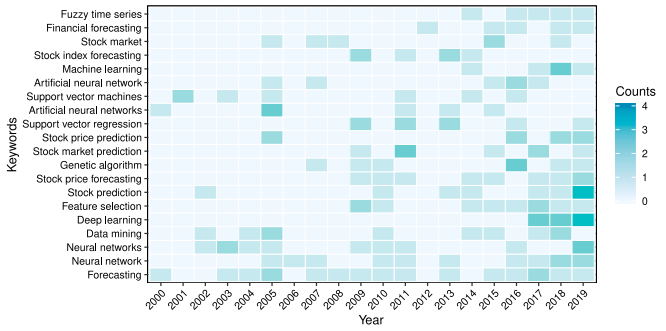

Fig. A.40. Author's keywords dynamic by year.

an in-depth analysis of the stocks and financial markets covered in the literature as well as the types and specific variables used for predictions, (3) a detailed presentation of existing forecasting approaches based on machine learning and where they have been used in the literature, and (4) information on recent trends concerning the use of machine learning methods (such as deep learning) in stock market forecasting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the Finnish Foundation for Share Promotion (Pörssisäätiö), Finland.

Appendix A

A.1. Author's keywords dynamic by year

Fig. A.40 shows the use of these keywords from year to year. This illustrates that researches in the stock market analysis increasingly concentrated on new machine learning approaches such as deep learning. According to Fig. A.40, *neural networks* and *support vector machines* based models have been more popular in earlier studies. Over time researchers continued to used these techniques but increasingly combined several machine learning techniques (e.g., genetic algorithm, feature selection) with the parent model rather than using it alone. Especially the rise of deep learning-based studies has been apparent in the last three years. This indicates that deep learning is receiving increased interest in stock market forecasting.

A.2. Citation performance

With the aim to understand the impact of the articles in this study, we also collected the number of citations for each article. Fig. A.41 presents the 15 most cited articles in each of the two 10 year periods during the last 20 years. The number of citations presented in the figure was collected via Google Scholar, which tracks the citation information of publications.

According to Fig. A.41 the research effort of Kim (2003) who applied an SVM to predict the stock index price and achieved promising results in comparison to a BPNN model and case-based reasoning, has been the paper with the largest number of citations among the most cited articles on this subject for last 20 years. It had 1504 citations at the time we conducted the search. The study by Tay and Cao (2001) which also utilized an SVM for financial market forecasting had 1242 citations, ranking it second among the most cited papers. Next to that, the paper by Huang et al. (2005) is placed third with 905 citations. This paper has investigated the applicability of SVMs in the prediction of the weekly movement of the stock price. The results were compared to linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and elman backpropagation neural networks (EBNN). In the paper, it was concluded that the SVM outperforms the benchmark methods deployed. In summary, the top 3 most cited articles for the last 20 years were based on SVM approaches for financial market forecasting.

Next, turning our focus towards the last decade, the research of Kara et al. (2011) has been the most cited one with 518 citations. In their paper, these authors are applying both neural networks and SVM methods to forecast the Istanbul stock market indices. Addressing the problems in the prediction of stock price movements, the study in Patel et al. (2015) compares the performance of neural networks, SVMs, random forests, and Naive Bayes. This study has gained 386 citations in the last 5 years. We should point out that the study of Fischer and Krauss (2018) has been becoming more and more popular among the latest research studies since it gained 268 citations within a short time period of about 2 years. This study deployed long short-term memory (LSTM) neural networks as one of the most advanced deep learning techniques to forecast stock price movements and achieved competitive results compared with a random forest and logistic regression. This also indicates what we mentioned previously in the analysis of the keyword, particularly, that deep learning is gaining more attention in recent researches in financial market forecasting.

Table B.1
List of performance metrics in full names.

Abbreviation	Definition
AUC	Area under curve
Dstat	Directional statistic
MDD	Maximum drawdown
POCID	Prediction of change in direction
ARV	Average relative variance
Gmean	Geometric mean
RMSE	Root mean square error
RE	Relative error
MAPE	Mean absolute percentage error
MASE	Mean absolute scaled error
ROI	Return of the investment
DA	Directional accuracy
RE	Relative error
ROR	Rate of return
PE	Prediction error
MAE	Mean absolute error
NMSE	Normalized mean square error
RelMAE	Relative mean absolute error
RM SPE	Root mean square percentage error

Appendix B

See Table B.1.

Appendix C

See Table C.1.

Appendix D

See Table D.1.

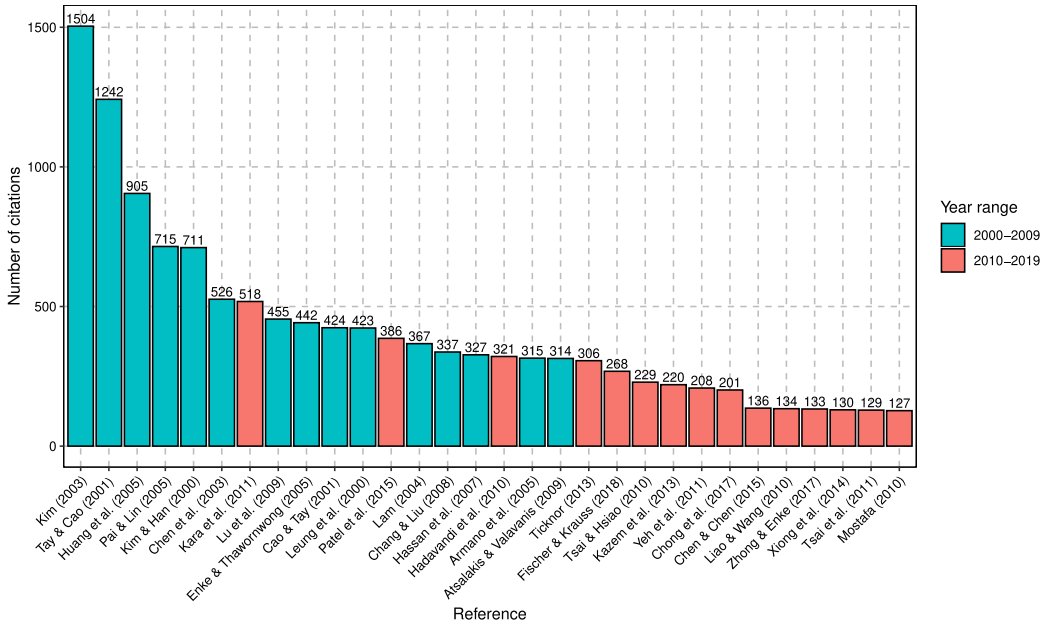

Fig. A.41. Top 30 articles among the mostly cited.

Table C.1
Indices by full name, exchange and market.

Index Abbreviation	Index Name	Exchange	Market
ASE	Athens Stock Exchange Composite Index	Athens Stock Exchange	Greece
ATX	Austrian Traded Index	Vienna Stock Exchange	Austria
BIST 100	Borsa Istanbul 100 Index	Istanbul Stock Exchange	Turkey
BSE SENSEX	Bombay Stock Exchange Sensitive Index	Bombay Stock Exchange	India
Bursa Malaysia	Bursa Malaysia (Kuala Lumpur Stock Exchange)	Bursa Malaysia	Malaysia
CDAX	Composite Deutscher Aktienindex	Deutsche Börse	Germany
CME	Chicago Mercantile Exchange Futures	Chicago Mercantile Exchange	USA
CNX NIFTY 50	CNX National Fifty 50	National Stock Exchange of India	India
CNX NIFTY 500	CNX National Fifty 500	National Stock Exchange of India	India
COMIT Index	Banca Commerciale Italiana Index	Borsa Italiana	Italy
CSI 300	China Securities Index 300	Shanghai & Shenzhen Stock Exchange	China
CSI 500	China Securities Index 500	Shanghai & Shenzhen Stock Exchange	China
DAX	Deutscher Aktienindex	Deutsche Börse	Germany
DJIA	Dow Jones Industrial Average Index	NYSE & NASDAQ	USA
Euro STOXX 600	European Stocks 600 by STOXX Ltd.	European Stock Market	European Union
FITX	Taiwan Stock Exchange Index Futures	Taiwan Stock Exchange	Taiwan
FOREX	Foreign Exchange Rates Market	–	USA
FTSE 100	FTSE 100 Index	London Stock Exchange	United Kingdom
GEM	Growth Enterprise Market	Hong Kong Stock Exchange	China
HSI	Hang Seng Index	Hong Kong Stock Exchange	China
IBEX 35	Indice Bursatil Espanol 35	Bolsa de Madrid	Spain
IBOVESPA	Indice Bovespa	Brasil Bolsa Balcão	Brazil
JKSE	Jakarta Stock Exchange Composite Index	Jakarta Stock Exchange	Indonesia
KOSPI	Korea Composite Stock Price Index	Korea Exchange	South Korea
KOSPI 200	Korea Composite Stock Price Index	Korea Exchange	South Korea
KSE	Kuwait Stock Exchange Index	Kuwait Stock Exchange	Kuwait
LSE	London Stock Exchange	London Stock Exchange	United Kingdom
MCX COMDEX	MCX Commodity Index Futures	Multi Commodity Exchange of India	India
NASDAQ	NASDAQ Composite Index	Nasdaq Stock Exchange	USA
Nikkei 225	Nihon Keizai Shinbun 225 Index	Tokyo Stock Exchange	Japan
NYSE	New York Stock Exchange Composite Index	New York Stock Exchange	USA
QE	Qatar Stock Exchange Index	Qatar Stock Exchange	Qatar
Russell 2000	Russell 2000 Index	NYSE & NASDAQ, OTC Markets	USA
S&P 400	Standard & Poor's 400 Index	NYSE & NASDAQ, Investors Exchange	USA
S&P 500	Standard & Poor's 500 Index	NYSE & NASDAQ, CBOE Exchange	USA
S&P 600	Standard & Poor's 600 Index	NYSE & NASDAQ, OTC Markets	USA
SSE	Shanghai Stock Exchange Composite Index	Shanghai Stock Exchange	China
SSE 50	Shanghai Stock Exchange 50 Index	Shanghai Stock Exchange	China

(*continued on next page*)

Table C.1 (*continued*).

Index Abbreviation	Index Name	Exchange	Market
STI	FTSE Straits Times Index	Singapore Exchange	Singapore
SZSE	Shenzhen Stock Exchange Component Index	Shenzhen Stock Exchange	China
TAIEX	Taiwan Capitalization Weighted Stock Index	Taiwan Stock Exchange	Taiwan
TEJ Data	Taiwan Economic Journal Data Set	–	Asia Multiple
TSE	Tehran Stock Exchange Index (TEDPIX)	Tehran Stock Exchange	Iran
TSX	S&P/TSX Composite Index	Toronto Stock Exchange	Canada
WIG 20	Warszawski Indeks Giełdowy 20	Warsaw Stock Exchange	Poland
VN index	Vietnam Ho Chi Minh Stock Index	Ho Chi Minh City Stock Exchange	Vietnam

Table D.1
Selected indices by market capitalization.

Index	Market Cap. in B. USD	Value Date	Source
ASE	46.8	10.4.2020	https://www.athexgroup.gr/en/web/guest/stocks/-/map/m/-1/6/
ATX	85.2	31.3.2020	https://www.wienerborse.at/uploads/u/cms/files/press/media-fact-sheet-the-vienna-stock-exchange.pdf
BIST 100	46.2	10.4.2020	https://www.borsaistanbul.com/en/data/data/consolidated-data
BSE SENSEX	1638.0	12.3.2020	https://www.business-standard.com/article/markets/bse-market-cap-at-lowest_-level-since-june-2017-lost-rs-33-trn-in-one-month-120031200393_1.html
Bursa Malaysia	317.2	31.3.2020	https://www.ceicdata.com/en/malaysia/bursa-malaysia-market-capitalization/bursa-malaysia-market-capitalization
CNX NIFTY 50	906.0	9.4.2020	Sum of Market Capitalization of the 50 Constituent (Google financial information from 09.04.2020)
CSI 300	4165.7	31.3.2020	http://www.csindex.com.cn/uploads/indices/detail/files/en/000300factsheeten.pdf?t=1586637379
DAX	1138.1	10.4.2020	https://www.finanzen.net/index/dax/marktkapitalisierung
DJIA	7252.3	9.4.2020	https://markets.businessinsider.com/index/dow_jones
FTSE 100	1425.9	31.3.2020	https://www.ftserussell.com/analytics/factsheets/home/search
Hang Seng	2235.2	28.2.2020	https://www.hsi.com.hk/static/uploads/contents/en/dl_centre/factsheets/hsie.pdf
IBOVESPA	1030.0	10.4.2020	https://www.tradinghours.com/exchanges/bovespa
KOSPI	979.2	31.3.2020	https://www.ceicdata.com/en/korea/korea-exchange-kospi-market-market-_capitalization/market-cap-kospi-total
LSE	3625.4	31.3.2020	https://www.londonstockexchange.com/statistics/markets/main-market/main-market.htm
NASDAQ	12644.5	31.3.2020	https://www.ceicdata.com/en/united-states/nasdaq-turnover-and-market-_capitalization/market-capitalization-nasdaq
Nikkei 225	2939.7	10.4.2020	https://indexes.nikkei.co.jp/en/nkave/archives/summary
NYSE	30213.5	9.4.2020	Extrapolated using the NYSE Composite Index value & old Market Cap from: https://www.nyse.com/market-cap
S&P 500	24560.4	9.4.2020	https://markets.businessinsider.com/index/s&p_500
BFBFBFSSE	4724.6	10.4.2020	http://english.sse.com.cn/
SZSE	3341.4	3.4.2020	http://www.szse.cn/English/about/news/szse/P020200407522583193167.pdf
TAIEX	1018.9	10.4.2020	https://www.twse.com.tw/en/
TSE	519.0	8.4.2020	https://tse.ir/pages/Files/Peyvast/pf_72821.pdf

References

Ahmad, M. O., Dennehy, D., Conboy, K., & Oivo, M. (2018). Kanban in software engineering: A systematic mapping study. *Journal Of Systems And Software*, *137*, 96–113. http://dx.doi.org/10.1016/j.jss.2017.11.045.

Ahmadi, E., Jasemi, M., Monplaisir, L., Nabavi, M. A., Mahmoodi, A., & Amini Jam, P. (2018). New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the support vector machine and heuristic algorithms of imperialist competition and genetic. *Expert Systems With Applications*, *94*, 21–31.

Ahmed, Y. A., Ahmad, M. N., Ahmad, N., & Zakaria, N. H. (2019). Social media for knowledge-sharing: A systematic literature review. *Telematics And Informatics*, *37*, 72–112. http://dx.doi.org/10.1016/j.tele.2018.01.015.

Altay, E., & Satman, M. H. (2005). Stock market forecasting: Artificial neural network and linear regression comparison in an emerging market. *Journal Of Financial Management And Analysis*, *18*(2), 8–33.

Ambreen, T., Ikram, M., & Niazi, M. (2018). Empirical research in requirements engineering: trends and opportunities. *Requirements Engineering*, *23*, 63–95.

Anbalagan, T., & Maheswari, S. U. (2015). Classification and prediction of stock market index based on fuzzy metagraph. *Procedia Computer Science*, *47*(C), 214–221.

Anish, C. M., & Majhi, B. (2016). Hybrid nonlinear adaptive scheme for stock market prediction using feedback FLANN and factor analysis. *Journal Of The Korean Statistical Society*, *45*(1), 64–76. http://dx.doi.org/10.1016/j.jkss.2015.07.002.

Araújo, R. D. A. (2011). Translation invariant morphological time-lag added evolutionary forecasting method for stock market prediction. *Expert Systems With Applications*, *38*(3), 2835–2848.

Araújo, R. D. A., Oliveira, A. L., & Meira, S. (2015). A hybrid model for high-frequency stock market forecasting. *Expert Systems With Applications*, *42*(8), 4081–4096.

Armano, G., Marchesi, M., & Murru, A. (2005). A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences*, *170*(1), 3–33.

Atsalakis, G. S., & Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems With Applications*, *36*(7), 10696–10707. http://dx.doi.org/10.1016/j.eswa.2009.02.043.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques - part II: Soft computing methods. *Expert Systems With Applications*, *36*, 5932–5941. http://dx.doi.org/10.1016/j.eswa.2008.07.006.

Badolia, L. (2016). *How can i get started investing in the stock market*. Educreation Publishing, 2016.

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. http://dx.doi.org/10.1016/j.eswa.2018.07.019.

Barak, S., Arjmand, A., & Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market. *Information Fusion*, *36*, 90–102. http://dx.doi.org/10.1016/j.inffus.2016.11.006.

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems With Applications*, *42*(3), 1325–1339. http://dx.doi.org/10.1016/j.eswa.2014.09.026.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *North American Journal Of Economics And Finance*, *47*, 552–567. http://dx.doi.org/10.1016/j.najef.2018.06.013.

Bataa, E., Vivian, A., & Wohar, M. (2019). Changes in the relationship between short-term interest rate, inflation and growth: evidence from the UK, 1820-2014. *Bulletin Of Economic Research*, *71*, 616–640.

Bisoi, R., Dash, P. K., & Parida, A. K. (2019). Hybrid variational mode decomposition and evolutionary robust kernel extreme learning machine for stock price and movement prediction on daily basis. *Applied Soft Computing*, *74*, 652–678. http://dx.doi.org/10.1016/j.asoc.2018.11.008.

Bodie, Z., Kane, A., & Marcus, A. J. (2009). *Investments* (8th Ed.). Irwin: McGraw-Hill.

Bondt, W. F. M. D., & Thaler, R. (1985). Does the stock market overreact? *The Journal Of Finance*, *40*, 793–805.

Bondt, W. F. M. D., & Thaler, R. H. (1990). Do security analysts overreact? *The American Economic Review*, *80*, 52–57.

Borovkova, S., & Tsiamas, I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal Of Forecasting*, *38*(6), 600–619. http://dx.doi.org/10.1002/for.2585.

Braun, V. (2006). Using thematic analysis in psychology. *Qualitative Research In Psychology*, *3*, 77–101.

Bustos, S. M., Anderson, J. V., Miniconi, M., Nowak, M., Roszczynska-Kurasinska, M., & Brée, D. (2011). Pricing stocks with yardsticks and sentiments. *Algorithmic Finance*, *1*, 183–190.

Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems With Applications*, *156*, Article 113464.

Cagcag Yolcu, O., & Alpaslan, F. (2018). Prediction of TAIEX based on hybrid fuzzy time series model with single optimization process. *Applied Soft Computing*, *66*, 18–33. http://dx.doi.org/10.1016/j.asoc.2018.02.007.

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers And Operations Research*, *32*(10).

Cao, H., Lin, T., Li, Y., & Zhang, H. (2019). Stock price pattern prediction based on complex network and machine learning. *Complexity*, *2019*.

Cao, L., & Tay, F. E. (2001). Financial forecasting using support vector machines. *Neural Computing And Applications*, *10*, 184–192.

Cao, J., & Wang, J. (2019). Stock price forecasting model based on modified convolution neural network and financial time series analysis. *International Journal Of Communication Systems*, *32*, 1–13.

Carosia, A. E. O., Coelho, G. P., & Silva, A. E. A. (2019). Analyzing the Brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, *34*, 1–19.

Carta, S., Consoli, S., Piras, L., Podda, A. S., & Recupero, D. R. (2021a). Event detection in finance using hierarchical clustering algorithms on news and tweets. *PeerJ Computer Science*, *7*.

Carta, S. M., Consoli, S., Piras, L., Podda, A. S., & Recupero, D. R. (2021b). Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. *IEEE Access*, *9*, 30193–30205. http://dx.doi.org/10.1109/ACCESS.2021.3059960.

Chai, J., Du, J., Lai, K. K., & Lee, Y. P. (2015). A hybrid least square support vector machine model with parameters optimization for stock forecasting. *Mathematical Problems In Engineering*, *2015*.

Chang, P. C., & Fan, C. Y. (2008). A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting. *IEEE Transactions On Systems, Man And Cybernetics Part C: Applications And Reviews*, *38*, 802–815. http://dx.doi.org/10.1109/TSMCC.2008.2001694.

Chang, P. C., & Liu, C. H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems With Applications*, *34*(1), 135–144. http://dx.doi.org/10.1016/j.eswa.2006.08.020.

Chang, P. C., Liu, C. H., Lin, J. L., Fan, C. Y., & Ng, C. S. (2009). A neural network with a case based dynamic window for stock trading prediction. *Expert Systems With Applications*, *36*(3 PART 2), 6889–6898. http://dx.doi.org/10.1016/j.eswa.2008.08.077.

Chang, P. C., & Wu, J. L. (2015). A critical feature extraction by kernel PCA in stock trading model. *Soft Computing*, *19*(5), 1393–1408.

Chang, P. C., Wu, J. L., & Lin, J. J. (2016). A Takagi-Sugeno fuzzy model combined with a support vector regression for stock trading forecasting. *Applied Soft Computing*, *38*, 831–842. http://dx.doi.org/10.1016/j.asoc.2015.10.030.

Chavarnakul, T., & Enke, D. (2018). Aintelligent technical analysis based equivolume charting for stock trading using neural networks. *Expert Systems With Applications*, *34*, 1004–1017.

Chen, M. Y., & Chen, B. T. (2015). A hybrid fuzzy time series model based on granular computing for stock price forecasting. *Information Sciences*, *294*, 227–241. http://dx.doi.org/10.1016/j.ins.2014.09.038.

Chen, Y. J., Chen, Y. M., & Lu, C. L. (2017). Enhancement of stock market forecasting using an improved fundamental analysis-based approach. *Soft Computing*, *21*(13), 3735–3757. http://dx.doi.org/10.1007/s00500-016-2028-y.

Chen, Y. S., Cheng, C. H., & Tsai, W. L. (2014). Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *41*(2), 327–347. http://dx.doi.org/10.1007/s10489-014-0520-6.

Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems With Applications*, *80*, 340–355.

Chen, Y., & Hao, Y. (2018). Integrating principle component analysis and weighted support vector machine for stock trading signals prediction. *Neurocomputing*, *321*, 381–402. http://dx.doi.org/10.1016/j.neucom.2018.08.077.

Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan stock index. *Computers And Operations Research*, *30*(6), 901–923. http://dx.doi.org/10.1016/S0305-0548(02)00037-0.

Chen, Y., Lin, W., & Wang, J. Z. (2019). A dual-attention-based stock price trend prediction model with dual features. *IEEE Access*, *7*, 148047–148058.

Chiang, W. C., Enke, D., Wu, T., & Wang, R. (2016). An adaptive stock index trading decision support system. *Expert Systems With Applications*, *59*, 195–207. http://dx.doi.org/10.1016/j.eswa.2016.04.025.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems With Applications*, *83*, 187–205. http://dx.doi.org/10.1016/j.eswa.2017.04.030.

Chou, J., & Nguyen, T. (2018). Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression. *IEEE Transactions On Industrial Informatics*, *14*(7), 3132–3142.

Chu, H. H., Chen, T. L., Cheng, C. H., & Huang, C. C. (2009). Fuzzy dual-factor time-series for stock index forecasting. *Expert Systems With Applications*, *36*(1), 165–171. http://dx.doi.org/10.1016/j.eswa.2007.09.037.

Chun, S. H., & Park, Y. J. (2005). Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction. *Expert Systems With Applications*, *28*(3), 435–443. http://dx.doi.org/10.1016/j.eswa.2004.12.004.

Chung, H., & Shin, K. S. (2018). Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability*, *10*(10), http://dx.doi.org/10.3390/su10103765.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *The Journal Of Finance*, *53*, 1839–1885.

Das, S. P., & Padhy, S. (2018). A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. *International Journal Of Machine Learning And Cybernetics*, *9*(1), 97–111. http://dx.doi.org/10.1007/s13042-015-0359-0.

de Faria, E. L., Albuquerque, M. P., Gonzalez, J. L., Cavalcante, J. T., & Albuquerque, M. P. (2009). Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods. *Expert Systems With Applications*, *36*(10), 12506–12509. http://dx.doi.org/10.1016/j.eswa.2009.04.032.

Ebadati, O. M., & Mortazavi, M. T. (2018). An efficient hybrid machine learning method for time series stock market forecasting. *Neural Network World*, *28*(1), 41–55. http://dx.doi.org/10.14311/NNW.2018.28.003.

Ebrahimpour, R., Kabir, E., & Yousef, M. R. (2007). Face detection using mixture of MLP experts. *Neural Processing Letters*, *26*, 69–82.

Ebrahimpour, R., Nikoo, H., Masoudnia, S., Yousefi, M. R., & Ghaemi, M. S. (2011). Mixture of mlp-experts for trend forecasting of time series: A case study of the tehran stock exchange. *International Journal Of Forecasting*, *27*(3), 804–816. http://dx.doi.org/10.1016/j.ijforecast.2010.02.015.

Efendi, R., Arbaiy, N., & Deris, M. M. (2018). A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. *Information Sciences*, *441*, 113–132.

Enke, D., Grauer, M., & Mehdiyev, N. (2011). Stock market prediction with multiple regression, fuzzy type-2 clustering and neural networks. *Procedia Computer Science*, *6*, 201–206. http://dx.doi.org/10.1016/j.procs.2011.08.038.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems With Applications*, *29*(4), 927–940.

Fadlalla, A., & Amani, F. (2014). Predicting next trading day closing price of Qatar exchange index using technical indicators and artificial neural networks. *Intelligent Systems In Accounting, Finance And Management*, *21*, 209–223.

Fama, E. F. (1965). The behavior of stock market prices. *The Journal Of Business*, *38*, 34–105.

Fama, E. (1970). Efficient capital markets: A review of the theory. *The Journal Of Finance*, *25*, 383–417.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal Of Financial Economics*, *49*, 283–306.

Farias Nazário, R. T., e Silva, J. L., Sobreiro, V. A., & Kimura, H. (2017). A literature review of technical analysis on stock markets. *Quarterly Review Of Economics And Finance*, *66*, 115–126. http://dx.doi.org/10.1016/j.qref.2017.01.014.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*(June), 88–97. http://dx.doi.org/10.1016/j.dss.2018.06.008.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal Of Operational Research*, *270*, 654–669.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, *34*, Article 100190. http://dx.doi.org/10.1016/j.cosrev.2019.08.001.

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2019). Stock price prediction using hybrid soft computing models incorporating parameter tuning and input variable selection. *Neural Computing And Applications*, *31*(2), 577–592.

Gocken, M., ÖZÇALICI, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating meta-heuristics and artificial neural networks for improved stock price prediction. *Expert Systems With Applications*, *44*, 320–331.

Gorenc Novak, M., & Velušček, D. (2016). Prediction of stock price movement based on daily high prices. *Quantitative Finance*, *16*(5), 793–826. http://dx.doi.org/10.1080/14697688.2015.1070960.

Gowthul Alam, M. M., & Baulkani, S. (2019). Local and global characteristics-based kernel hybridization to increase optimal support vector machine performance for stock market prediction. *Knowledge And Information Systems*, *60*(2), 971–1000. http://dx.doi.org/10.1007/s10115-018-1263-1.

Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowledge-Based Systems*, *137*, 138–148. http://dx.doi.org/10.1016/j.knosys.2017.09.023.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal Of Machine Learning Research*, *3*, 1157–1182.

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, *23*(8), 800–808. http://dx.doi.org/10.1016/j.knosys.2010.05.004.

Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Systems With Applications*, *33*(1), 171–180.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems With Applications*, *124*, 226–251.

Horata, P., Chiewchanwattana, S., & Sunat, K. J. N. (2013). Robust extreme learning machine. *Neurocomputing*, *102*, 31–44.

Hoseinzade, E., & Haratizadeh, S. (2019). CNNPred: CNN-based stock market prediction using a diverse set of variables. *Expert Systems With Applications*, *129*, 273–285. http://dx.doi.org/10.1016/j.eswa.2019.03.029.

Hsieh, L. F., Hsieh, S. C., & Tai, P. H. (2011). Enhanced stock price variation prediction via DOE and BPNN-based optimization. *Expert Systems With Applications*, *38*(11), 14178–14184. http://dx.doi.org/10.1016/j.eswa.2011.04.229.

Hu, H., Tang, L., Zhang, S., & Wang, H. (2018). Predicting the direction of stock markets using optimized neural networks with Google Trends. *Neurocomputing*, *285*, 188–195. http://dx.doi.org/10.1016/j.neucom.2018.01.038.

Huang, J. Y., & Liu, J. H. (2020). Using social media data mining technology to improve stock price forecast accuracy. *Journal Of Forecasting*, *39*, 104–116.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers And Operations Research*, *32*, 2513–2522.

Huang, C. L., & Tsai, C. Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems With Applications*, *36*, 1529–1539.

Hyup Roh, T. (2007). Forecasting the volatility of stock price index. *Expert Systems With Applications*, *33*(4), 916–922. http://dx.doi.org/10.1016/j.eswa.2006.08.001.

Javedani Sadaei, H., & Lee, M. H. (2014). Multilayer stock forecasting model using fuzzy time series. *The Scientific World Journal*, *2014*, http://dx.doi.org/10.1155/2014/610594.

Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems With Applications*, *184*, Article 115537. http://dx.doi.org/10.1016/j.eswa.2021.115537.

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics And Its Applications*, *541*(258), Article 122272. http://dx.doi.org/10.1016/j.physa.2019.122272.

Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems With Applications*, *178*, Article 115019.

Kao, L. J., Chiu, C. C., Lu, C. J., & Chang, C. H. (2013). A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems*, *54*(3), 1228–1244.

Kara, Y., Acar Boyacioglu, M., & Baykan, O. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange. *Expert Systems With Applications*, *38*, 5311–5319.

Kaur, G., Dhar, J., & Guha, R. K. (2016). Minimal variability OWA operator combining ANFIS and fuzzy c-means for forecasting BSE index. *Mathematics And Computers In Simulation*, *122*, 69–80. http://dx.doi.org/10.1016/j.matcom.2015.12.001.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, *13*(2), 947–958.

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *Applied Mathematical Finance*, 1–20.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*, 307–319.

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems With Applications*, *19*, 125–132.

Kim, S., Ku, S., Chang, W., Chang, W., Chang, W., & Song, J. W. (2020). Predicting the direction of US stock prices using effective transfer entropy and machine learning techniques. *IEEE Access*, *8*, 111660–111682. http://dx.doi.org/10.1109/ACCESS.2020.3002174.

Kitchenham, B. (2004). *Procedures for performing systematic reviews*. Keele University, UK and National ICT Australia.

Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information And Software Technology*, *55*, 2049–2075.

Klemm, P., Bunnell, D., Cullen, M., Soneji, R., Gibbons, P., & Holecek, A. (2003). Online cancer support groups: a review of the research literature.. *Computers, Informatics, Nursing : CIN*, *21*, 136–142. http://dx.doi.org/10.1097/00024665-200305000-00010.

Kubat, M. (2017). *An introduction to machine learning* (2nd Ed.). Springer Publishing Company, Incorporated.

Kumar, D., Sarangi, P. K., & Verma, R. (2021). A systematic review of stock market prediction using machine learning and statistical techniques. *Materials Today: Proceedings*, http://dx.doi.org/10.1016/j.matpr.2020.11.399.

Laboissiere, L. A., Fernandes, R. A. S., & Lage, G. G. (2015). Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. *Applied Soft Computing*, *35*, 66–74. http://dx.doi.org/10.1016/j.asoc.2015.06.005.

Lai, R. K., Fan, C. Y., Huang, W. H., & Chang, P. C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems With Applications*, *36*(2 PART 2), 3761–3773. http://dx.doi.org/10.1016/j.eswa.2008.02.025.

Lam, M. (2004). Neural network techniques for financial performance prediction: Integrating fundamental and technical analysis. *Decision Support Systems*, *37*(4), 567–581. http://dx.doi.org/10.1016/S0167-9236(03)00088-5.

Lee, S. W., & Kim, H. Y. (2020). Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation. *Expert Systems With Applications*, *161*, Article 113704.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems*, *32*, 361–377.

Leung, M. T., Daouk, H., & Chen, A. S. (2000). Forecasting stock indices: A comparison of classification and level estimation models. *International Journal Of Forecasting*, *16*(2), 173–190. http://dx.doi.org/10.1016/S0169-2070(99)00048-5.

Li, Y., Bu, H., Li, J., & Wu, J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal Of Forecasting*, *36*(4), 1541–1562. http://dx.doi.org/10.1016/j.ijforecast.2020.05.001.

Li, B., Chan, K., Ou, C. X., & Sun, R. (2017). Discovering public sentiment in social media for predicting stock movement of publicly listed companies. *Information Systems*, *69*, 81–92.

Liao, Z., & Wang, J. (2010). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems With Applications*, *37*(1), 834–841. http://dx.doi.org/10.1016/j.eswa.2009.05.086.

Lien Minh, D., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404. http://dx.doi.org/10.1109/ACCESS.2018.2868970.

Lim, K. P., & Brooks, R. (2011). The evolution of stock market efficiency over time: a survey of the empirical literature. *Journal Of Economic Surveys*, *25*, 69–108.

Liu, H., & Long, Z. (2020). An improved deep learning model for predicting stock market price time series. *Digital Signal Processing: A Review Journal*, *102*, Article 102741. http://dx.doi.org/10.1016/j.dsp.2020.102741.

Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal Of Portfolio Management*, *30*, 15–129.

Lohrmann, C., & Luukka, P. (2019). Classification of intraday S&P500 returns with a random forest. *International Journal Of Forecasting*, *35*(1), 390–407. http://dx.doi.org/10.1016/j.ijforecast.2018.08.004.

Lohrmann, C., Luukka, P., Jablonska-Sabuka, M., & Kauranne, T. (2018). A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection. *Expert Systems With Applications*, *110*, 216–236.

Lu, C. J. (2013). Hybridizing nonlinear independent component analysis and support vector regression with particle swarm optimization for stock index forecasting. *Neural Computing And Applications*, *23*(7–8), 2417–2427. http://dx.doi.org/10.1007/s00521-012-1198-5.

Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, *47*(2), 115–125.

Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-based model to forecast stock prices. *Complexity*, *2020*, http://dx.doi.org/10.1155/2020/6622927.

Lu, C. J., & Wu, J. Y. (2011). An efficient CMAC neural network for stock index forecasting. *Expert Systems With Applications*, *38*(12), 15194–15201. http://dx.doi.org/10.1016/j.eswa.2011.05.082.

Maknickiene, N., & Maknickas, A. (2013). Financial market prediction system with Evolino neural network and delphi method. *Journal Of Business Economics And Management*, *14*(2), 403–413.

Malagrino, L. S., Roman, N. T., & Monteiro, A. M. (2018). Forecasting stock market index daily direction: A Bayesian network approach. *Expert Systems With Applications*, *105*, 11–22. http://dx.doi.org/10.1016/j.eswa.2018.03.039.

Malkiel, B., & Mullainathan, S. (2005). Market efficiency versus behavioral finance. *Journal Of Applied Corporate Finance*, *17*, 124–136.

Mo, H., & Wang, J. (2018). Return scaling cross-correlation forecasting by stochastic time strength neural network in financial market dynamics. *Soft Computing*, *22*(9), 3097–3109. http://dx.doi.org/10.1007/s00500-017-2564-0.

Mostafa, M. M. (2010). Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait. *Expert Systems With Applications*, *37*(9), 6302–6309. http://dx.doi.org/10.1016/j.eswa.2010.02.091.

Murphy, J. J. (1999). *Technical analysis of the financial markets: a comprehensive guide to trading methods and applications, Vol. 2*. Prentice Hall Press.

Na, S. H., & Sohn, S. Y. (2011). Forecasting changes in Korea composite stock price index (KOSPI) using association rules. *Expert Systems With Applications*, *38*(7), 9046–9049. http://dx.doi.org/10.1016/j.eswa.2011.01.025.

Nermend, Y., & Alsakaa, K. (2017). Back-propagation artificial neural networks in stock market forecasting . An application to the warsaw stock exchange WIG20. *The IEB International Journal Of Finance*, *15*, 88–99.

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal Of Industrial Engineering International*, *9*(1), 1–9. http://dx.doi.org/10.1186/2251-712X-9-1.

O'Connor, N., & Madden, M. G. (2006). A neural network approach to predicting stock exchange movements using external factors. *Knowledge-Based Systems*, *19*, 371–378.

Olson, D., & Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal Of Forecasting*, *19*(3), 453–465. http://dx.doi.org/10.1016/S0169-2070(02)00058-4.

Pai, P. F., & Lin, C. S. (2005). ARIMA,artificial neural networks,stock prices,support vector machines,time series forecasting. *Omega*, *33*, 497–505.

Pal, S. S., & Kar, S. (2019). Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory. *Mathematics And Computers In Simulation*, *162*, 18–30.

Pan, Y., Xiao, Z., Wang, X., & Yang, D. (2017). A multiple support vector machine approach to stock index forecasting with mixed frequency sampling. *Knowledge-Based Systems*, *122*, 90–102. http://dx.doi.org/10.1016/j.knosys.2017.01.033.

Patel, P. B., & Marwala, T. (2006). Forecasting closing price indices using neural networks. In *IEEE International Conference On Systems, Man, And Cybernetics* (pp. 2351–2356).

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems With Applications*, *42*, 259–268.

Pei, A., Wang, J., & Fang, W. (2017). Predicting agent-based financial time series model on lattice fractal with random Legendre neural network. *Soft Computing*, *21*(7), 1693–1708. http://dx.doi.org/10.1007/s00500-015-1874-3.

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems With Applications*, *135*, 60–70. http://dx.doi.org/10.1016/j.eswa.2019.06.014.

Qiu, M., & Song, Y. (2016). Predicting the direction of stock market index movement using an optimized artificial neural network model. *PLoS ONE*, *11*(5), 1–11. http://dx.doi.org/10.1371/journal.pone.0155133.

Qiu, M., Song, Y., & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. *Chaos, Solitons And Fractals*, *85*, 1–7. http://dx.doi.org/10.1016/j.chaos.2016.01.004.

Rajab, S., & Sharma, V. (2019). An interpretable neuro-fuzzy approach to stock price forecasting. *Soft Computing*, *23*(3), 921–936. http://dx.doi.org/10.1007/s00500-017-2800-7.

Ramezanian, R., Peymanfar, A., & Ebrahimi, S. B. (2019). An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in tehran stock exchange market. *Applied Soft Computing*, *82*, Article 105551. http://dx.doi.org/10.1016/j.asoc.2019.105551.

Rather, A. M., Sastry, V. N., & Agarwal, A. (2017). Stock market prediction and Portfolio selection models: a survey. *Opsearch*, *54*(3), 558–579. http://dx.doi.org/10.1007/s12597-016-0289-y.

Rosillo, R., Giner, J., & De Fuente, D. L. (2014). The effectiveness of the combined use of VIX and support vector machines on the prediction of sandp 500. *Neural Computing And Applications*, *25*(2), 321–332.

Rubio, A., Bermúdez, J. D., & Vercher, E. (2017). Improving stock index forecasts by using a new weighted fuzzy-trend time series method. *Expert Systems With Applications*, *76*, 12–20. http://dx.doi.org/10.1016/j.eswa.2017.01.049.

Rustam, Z., & Kintandani, P. (2019). Application of support vector regression in Indonesian stock price prediction with feature selection using particle swarm optimisation. *Modelling And Simulation In Engineering*, *2019*, http://dx.doi.org/10.1155/2019/8962717.

Safer, A. M. (2002). The application of neural networks to predict abnormal stock returns using insider trading data. *Applied Stochastic Models In Business And Industry*, *18*(4), 381–389.

Sedighi, M., Jahangirnia, H., Gharakhani, M., & Fard, S. F. (2019). A novel hybrid model for stock price forecasting based on metaheuristics and support vector machine. *Data*, *4*(2), 1–28. http://dx.doi.org/10.3390/data4020075.

Selvamuthu, D., Kumar, V., & Mishra, A. (2019). IndiaN stock market prediction using artificial neural networks on tick data. *Financial Innovation*, *5*(1), http://dx.doi.org/10.1186/s40854-019-0131-7.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal Of Financial Studies*, *7*(2), http://dx.doi.org/10.3390/ijfs7020026.

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), http://dx.doi.org/10.1186/s40537-020-00333-6.

Shi, L., Teng, Z., Wang, L., Zhang, Y., & Binder, A. (2019). DeepClue: VIsual interpretation of text-based deep stock prediction. *IEEE Transactions On Knowledge And Data Engineering*, *31*, 1094–1108. http://dx.doi.org/10.1109/TKDE.2018.2854193.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *The Journal Of Economic Perspectives*, *17*, 83–104.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems*, *85*, 74–83. http://dx.doi.org/10.1016/j.dss.2016.03.001.

Singh, R., & Srivastava, S. (2017). Stock prediction using deep learning. *Multimedia Tools And Applications*, *76*(18), 18569–18584.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal Of Business Research*, *104*, 333–339.

Son, Y., Noh, D. J., & Lee, J. (2012). Forecasting trends of high-frequency KOSPI200 index data using learning classifiers. *Expert Systems With Applications*, *39*(14), 11607–11615. http://dx.doi.org/10.1016/j.eswa.2012.04.015.

Song, Q., & Chissom, B. S. (1993). Fuzzy time series and its models. *Fuzzy Sets And Systems*, *54*, 269–277.

Song, Y., Lee, J. W., & Lee, J. (2019). A study on novel filtering and relationship between input-features and target-vectors in a deep learning model for stock price prediction. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *49*(3), 897–911. http://dx.doi.org/10.1007/s10489-018-1308-x.

Sun, Y. (2014). A hybrid approach by integrating brain storm optimization algorithm with grey neural network for stock index forecasting. *Abstract And Applied Analysis*, *2014*.

Sun, Y., & Gao, Y. (2015). An improved hybrid algorithm based on PSO and BP for stock price forecasting. *Open Cybernetics And Systemics Journal*, *9*, 2565–2568.

Sun, J., Xiao, K., Liu, C., Zhou, W., & Xiong, H. (2019). Exploiting intra-day patterns for market shock prediction: A machine learning approach. *Expert Systems With Applications*, *127*, 272–281.

Tay, F. E., & Cao, L. J. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*, 309–317.

Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems With Applications*, *40*(14), 5501–5506.

Timmermann, A., & Granger, C. W. (2004). Efficient market hypothesis and forecasting. *International Journal Of Forecasting*, *20*, 15–27.

Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269. http://dx.doi.org/10.1016/j.dss.2010.08.028.

Tsai, C. F., Lin, Y. C., Yen, D. C., & Chen, Y. M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, *11*(2), 2452–2459. http://dx.doi.org/10.1016/j.asoc.2010.10.001.

Turner, T. (2007). *A beginner's guide to day trading online* (2nd Ed.). Adams Media.

Vilela, L. F., Leme, R. C., Pinheiro, C. A., & Carpinteiro, O. A. (2019). Forecasting financial series using clustering methods and support vector regression. *Artificial Intelligence Review*, *52*(2), 743–773. http://dx.doi.org/10.1007/s10462-018-9663-x.

Wang, Y. F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems With Applications*, *22*(1), 33–38. http://dx.doi.org/10.1016/S0957-4174(01)00047-1.

Wang, Y. F., Chuang, Y. L., Hsu, M. H., & Keh, H. C. (2004). A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection. *Expert Systems With Applications*, *26*, 427–434.

Wang, W. J., Tang, Y., Xiong, J., & Zhang, Y. C. (2021). Stock market index prediction based on reservoir computing models. *Expert Systems With Applications*, *178*, Article 115022. http://dx.doi.org/10.1016/j.eswa.2021.115022.

Wang, J., & Wang, J. (2015). Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks. *Neurocomputing*, *156*, 68–78.

Wang, J.-Z., Wang, J.-J., Zhang, Z.-G., & Guo, S.-P. (2011). Forecasting stock indices with back propagation neural network. *Expert Systems With Applications*, *38*, 14346–14355.

Wei, L. Y., Chen, T. L., & Ho, T. H. (2011). A hybrid model based on adaptive-network-based fuzzy inference system to forecast Taiwan stock market. *Expert Systems With Applications*, *38*(11), 13625–13631. http://dx.doi.org/10.1016/j.eswa.2011.04.127.

Wen, M., Li, P., Zhang, L., & Chen, Y. (2019). Stock market trend prediction using high-order information of time series. *IEEE Access*, *7*, 28299–28308.

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems With Applications*, *79*, 153–163. http://dx.doi.org/10.1016/j.eswa.2017.02.041.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *EASE '14*, *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. (38), New York, NY, USA.

Xiong, T., Bao, Y., & Hu, Z. (2014). Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting. *Knowledge-Based Systems*, *55*, 87–100. http://dx.doi.org/10.1016/j.knosys.2013.10.012.

Yang, F., Chen, Z., Li, J., & Tang, L. (2019). A novel hybrid stock selection method with stock prediction. *Applied Soft Computing*, *80*, 820–831. http://dx.doi.org/10.1016/j.asoc.2019.03.028.

Yang, Y., Mabu, S., Shimada, K., & Hirasawa, K. (2011). Fuzzy intertransaction class association rule mining using genetic network programming for stock market prediction. *IEEJ Transactions On Electrical And Electronic Engineering*, *6*(4), 353–360. http://dx.doi.org/10.1002/tee.20668.

Ye, F., Zhang, L., Zhang, D., Fujita, H., & Gong, Z. (2016). A novel forecasting method based on multi-order fuzzy time series and technical analysis. *Information Sciences*, *367–368*, 41–57. http://dx.doi.org/10.1016/j.ins.2016.05.038.

Yeh, C. Y., Huang, C. W., & Lee, S. J. (2011). A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Systems With Applications*, *38*(3), 2177–2186. http://dx.doi.org/10.1016/j.eswa.2010.08.004.

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions On Evolutionary Computation*, *13*(1), 87–102.

Yuan, K., Liu, G., Wu, J., & Xiong, H. (2020). Dancing with trump in the stock market. *ACM Transactions On Intelligent Systems And Technology*, *11*, 1–22.

Yuan, X., Yuan, J., Jiang, T., & Ain, Q. U. (2020). Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. *IEEE Access*, *8*, 22672–22685. http://dx.doi.org/10.1109/ACCESS.2020.2969293.

Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems With Applications*, *186*, Article 115716. http://dx.doi.org/10.1016/j.eswa.2021.115716.

Zadeh, L. A. (1965). Fuzzy sets. *Information And Control*, *8*, 338–353.

Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. *Expert Systems With Applications*, *97*, 60–69. http://dx.doi.org/10.1016/j.eswa.2017.12.026.

Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E. W., & Liu, M. (2014). A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, *142*, 48–59.

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics And Its Applications*, *477*, 161–173. http://dx.doi.org/10.1016/j.physa.2017.02.072.

Zhang, J., Shao, Y. H., Huang, L. W., Teng, J. Y., Zhao, Y. T., Yang, Z. K., & Li, X. Y. (2019). Can the exchange rate be used to predict the shanghai composite index? *IEEE Access*, *8*, 2188–2199.

Zhang, J., Teng, Y. F., & Chen, W. (2019). Support vector regression with modified firefly algorithm for stock price forecasting. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *49*(5), 1658–1674.

Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., & Yu, P. S. (2018). Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, *143*, 236–247.

Zhang, W., Zhang, S., Zhang, S., Yu, D., & Huang, N. N. (2019). A novel method based on FTS with both GA-FCM and multifactor BPNN for stock forecasting. *Soft Computing*, *23*(16), 6979–6994.

Zhong, X., & Enke, D. (2017). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, *267*, 152–168.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems With Applications*, *67*, 126–139. http://dx.doi.org/10.1016/j.eswa.2016.09.027.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, *5*(1), http://dx.doi.org/10.1186/s40854-019-0138-0.

Zhou, P. Y., Chan, K. C., & Ou, C. X. (2018). Corporate communication network and stock price movements: Insights from data mining. *IEEE Transactions On Computational Social Systems*, *5*(2), 391–402. http://dx.doi.org/10.1109/TCSS.2018.2812703.

Zhou, F., min Zhou, H., Yang, Z., & Yang, L. (2019). EMD2FNN: A Strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. *Expert Systems With Applications*, *115*, 136–151. http://dx.doi.org/10.1016/j.eswa.2018.07.065.

Zuo, Y., & Kita, E. (2012). Stock price forecast using Bayesian network. *Expert Systems With Applications*, *39*(8), 6729–6737.