International Journal of
Financial Studies

MDPI

Review

Stock Market Analysis: A Review and Taxonomy of Prediction Techniques

Dev Shah, Haruna Isah * and Farhana Zulkernine

School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada; dshah@cs.queensu.ca (D.S.); farhana@cs.queensu.ca (F.Z.)
* Correspondence: isah@cs.queensu.ca

Received: 4 March 2019; Accepted: 15 May 2019; Published: 27 May 2019

Abstract: Stock market prediction has always caught the attention of many analysts and researchers. Popular theories suggest that stock markets are essentially a random walk and it is a fool's game to try and predict them. Predicting stock prices is a challenging problem in itself because of the number of variables which are involved. In the short term, the market behaves like a voting machine but in the longer term, it acts like a weighing machine and hence there is scope for predicting the market movements for a longer timeframe. Application of machine learning techniques and other algorithms for stock price analysis and forecasting is an area that shows great promise. In this paper, we first provide a concise review of stock markets and taxonomy of stock market prediction methods. We then focus on some of the research achievements in stock analysis and prediction. We discuss technical, fundamental, short- and long-term approaches used for stock analysis. Finally, we present some challenges and research opportunities in this field.

Keywords: stock exchanges; stock markets; analysis; prediction; statistics; machine learning; pattern recognition; sentiment analysis

JEL Classification: G10; G15; G17; D53

1. Introduction

Financial markets are one of the most fascinating inventions of our time. They have had a significant impact on many areas like business, education, jobs, technology and thus on the economy (Hiransha et al. 2018). Over the years, investors and researchers have been interested in developing and testing models of stock price behaviour (Fama 1995). However, analysing stock market movements and price behaviours is extremely challenging because of the markets dynamic, nonlinear, nonstationary, nonparametric, noisy, and chaotic nature (Abu-Mostafa and Atiya 1996). According to Zhong and Enke (2017), stock markets are affected by many highly interrelated factors that include economic, political, psychological, and company-specific variables. Technical and fundamental analysis are the two main approaches to analyse the financial markets (Park and Irwin 2007; Nguyen et al. 2015). To invest in stocks and achieve high profits with low risks, investors have used these two major approaches to make decisions in financial markets (Arévalo et al. 2017).

According to Hu et al. (2015), fundamental analysis is mainly based on three essential aspects (i) macroeconomic analysis such as Gross Domestic Product (GDP) and Consumer Price Index (CPI) which analyses the effect of the macroeconomic environment on the future profit of a company, (ii) industry analysis which estimates the value of the company based on industry status and prospect, and (iii) company analysis which analyses the current operation and financial status of a company to evaluate its internal value. Different valuation approaches exist for fundamental analysis. The average growth approximation technique compares Stock-A with other stocks in the same category to better

understand valuations, i.e., assuming two companies have the same growth rate, the one with the lower Price-to-Earnings (P/E) ratio is considered to be better. Hence the fair price is the earnings times target P/E. The P/E method is the most commonly used valuation method in the stock brokerage industry (Imam et al. 2008). The constant growth approximation technique such as Gordon's growth model (Gordon and Shapiro 1956; Gordon 1959) is one of the best-known classes of dividend discount models. It assumes that dividends of a company will increase at a constant growth rate forever but less than the discount rate. Dutta et al. (2012) demonstrated the utility of fundamental analysis through the use of financial ratios to separate good stocks from poor stocks. The authors compared their one-year return against the benchmark—i.e., Nifty—which gives an accuracy of 74.6%. This is one of the few papers which focus on using fundamental features (i.e., company-specific ratios) to identify stocks for investments.

Furthermore, Hu et al. (2015) grouped the domains of technical analysis into sentiment, flow-of-funds, raw data, trend, momentum, volume, cycle, and volatility. Sentiment represents the behaviours of various market participants. Flow-of-funds is a type of indicator used to investigate the financial status of various investors to pre-evaluate their strength in terms of buying and selling stocks, then, corresponding strategies, such as short squeeze, can be adopted. Raw data include stock price series and price patterns such as K-line diagrams and bar charts. Trend and momentum are examples of price-based indicators, trend is used for tracing the stock price trends while momentum is used to evaluate the velocity of the price change and judge whether a trend reversal in stock price is about to occur. Volume is an indicator that reflects the enthusiasm of both buyers and sellers for investing, it is also a basis for predicting stock price movements. The cycle is based on the theory that stock prices vary periodically in the form of a long cycle of more than 10 years containing short cycles of a few days or weeks. Finally, volatility is often used to investigate the fluctuation range of stock prices and to evaluate risk and identify the level of support and resistance.

Sentiments can drive short-term market fluctuations which in turn cause disconnects between the price and true value of a company's shares but over long periods of time, however, the weighing machine kicks in as a company's fundamentals ultimately cause the value and market price of its shares to converge. A prominent example comes from the Nobel Laureate Robert Shiller, who showed that stock prices are extremely volatile over the short term but somewhat predictable by their price-to-earnings over long periods (Shiller 1980). Diamond (2000) explained what returns to expect from the stock markets considering the economic scenario and suggested that in the future, returns could be substantially lower. Shiller (2000) also suggested that stocks are overvalued, and the bubble will burst anytime. In the year 2000, rightly so, we witnessed the dotcom bubble burst.

Stock market price prediction is a tricky thing. Several theories regarding stock markets have been conceptualized over the years. They either try to explain the nature of stock markets or try to explain whether the markets can be beaten. One such popular and most debated theory given by Fama (1970) is the Efficient Market Hypothesis (EMH) which states that at any point in time, the market price of a stock incorporates all information about that stock. In other words, the stock is accurately valued until something changes. There are three variants of EMH (i) the weak form which is consistent with the random walk hypothesis (Fama 1995), and that stock prices move randomly while price changes are independent of each other hence, it is not possible to beat the market by earning abnormal returns on the basis of technical analysis; (ii) the semi-strong form which states that prices adjusted rapidly according to market and public information such as dividend, earnings announcements, and political or economic events, hence it is not possible to earn abnormal returns on the basis of fundamental analysis; and, finally, (iii) the strong form which states that prices reflect market, public, and private information as such no investor has monopolistic access to information (Naseer and Tariq 2015).

According to EMH, price changes are unpredictable and forecasting a financial market is a hopeless effort. However, (Abu-Mostafa and Atiya 1996) argued that the existence of so many price trends in financial markets and the undiscounted serial correlations among fundamental events and economic figures affecting the markets are two of many pieces of evidence against the EMH. Researchers and

investors disagree with EMH both empirically and theoretically, thereby shifting the focus of discussion from EMH to the behavioural and psychological aspects of market players (Naseer and Tariq 2015). According to Zhong and Enke (2017), financial variables, such as stock prices, stock market index values, and the prices of financial derivatives are therefore thought to be predictable. Many widely accepted empirical studies show that financial markets are to some extent predictable (Chong et al. 2017). Criticism of EMH has given rise to an increasing number of studies that question the validity of EMH and introduce new and successful approaches that combine technical analysis indicators and chart patterns with methodologies from econometrics, statistics, data mining, and artificial intelligence (Arévalo et al. 2017).

Many new technologies and methods have been proposed over the years to try and predict stock prices via many avenues, thanks to the challenging and ever-changing landscape of stock markets (Chen and Chen 2016). In this paper, we focus on two topics, namely, stock analysis and stock prediction. We look at the research in the past, but mainly focus on modern techniques, highlighting some of the main challenges they pose and recent achievements for stock analysis and prediction. Finally, we discuss potential challenges and possible future research directions. We organize the rest of this paper as follows. Section 2 provides a background review and taxonomy of the various approaches to stock market analysis. Section 3 describes a literature study on stock markets analysis and prediction. Section 4 discusses and compares the approaches mentioned in Section 3. Section 5 provides an overview of challenges and additional areas for future research. Finally, Section 6 concludes the paper.

2. Taxonomy of Stock Market Analysis Approaches

Recent advancements in stock analysis and prediction fall under four categories—statistical, pattern recognition, machine learning (ML), and sentiment analysis. These categories mostly fall under the broader category of technical analysis, however, there are some machine learning techniques which also combine the broader categories of technical analysis with fundamental analysis approaches to predict the stock markets. Figure 1 shows a taxonomy of popular stock prediction techniques. These techniques have gained popularity and have shown promising results in the field of stock analysis in the recent past.

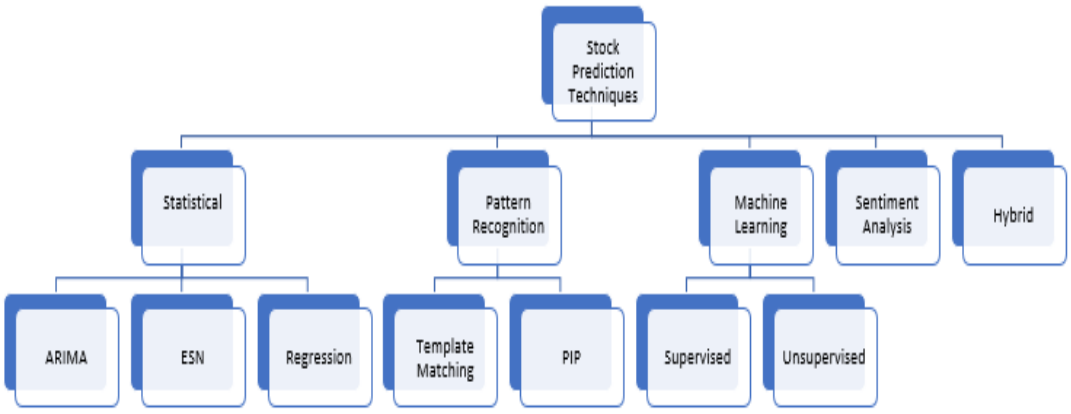

Figure 1. Taxonomy of stock prediction techniques.

Before the advent of machine learning techniques, statistical techniques which often assumes linearity, stationarity, and normality provided a way to analyse and predict stocks. Time series in stock market analysis is a chronological collection of observations such as daily sales totals and prices of stocks (Fu et al. 2005). According to Zhong and Enke (2017), one group of statistical approaches which fall into the category of univariate analysis, due to their use of time series as input variables, are the Auto-Regressive Moving Average (ARMA), the Auto-Regressive Integrated Moving Average (ARIMA), the Generalized Autoregressive Conditional Heteroskedastic (GARCH) volatility, and the

Smooth Transition Autoregressive (STAR) model. The ARIMA model is a widely used technique for stock market analysis (Hiransha et al. 2018). ARMA combines Auto-Regressive (AR) models which try to explain the momentum and mean reversion effects often observed in trading markets and Moving Average (MA) models which try to capture the shock effects observed in time series. A key limitation of the ARMA model is that it does not consider volatility clustering, a key empirical phenomenon in many financial time series. ARIMA is a natural extension to the class of ARMA models and can reduce a non-stationary series to a stationary series. The ARIMA (Box et al. 2015) is fitted to time series data to forecast future points. Zhong and Enke (2017) further describe another group of statistical approaches which usually utilize multiple input variables, these include Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and regression algorithms.

Pattern recognition is synonymous to machine learning but with respect to stock analysis, these two techniques are, however, applied in very different ways. Pattern recognition focuses on the detection of patterns and trends in data (Fu and Young 1986; Wang and Chan 2007; Parracho et al. 2010). Patterns in stock markets are recurring sequences found in Open-High-Low-Close (OHLC) candlestick charts which traders have historically used as buy and sell signals (Velay and Daniel 2018). Technical analysis relies on patterns found directly in stock data; it involves the visual analysis of charts constructed over time to show variations in price, volume, or other derived indicators such as price momentum (Nesbitt and Barrass 2004). Charting is a technique of technical analysis for comparing market price and volume history to chart patterns for predicting future price behaviour based on the degree of match (Leigh et al. 2002). Familiar chart patterns typically derived from their shapes are gaps, spikes, flags, pennants, wedges, saucers, triangles, head-and-shoulders, and various tops and bottoms (Park and Irwin 2007). Patterns of stock prices have the capacity to inform an investor of the future evolution of that stock (Parracho et al. 2010). Two widely used pattern recognition methods are Perceptually Important Points (PIP), which involve reducing time-series dimensions (i.e., the number of data point) by preserving the salient points and template matching, a technique used to match a given stock pattern with a pictographic image for object identification (Chen and Chen 2016). According to (Velay and Daniel 2018), many studies have found some correlation between patterns and future trends.

Machine learning has been extensively studied for its potentials in the prediction of financial markets (Shen et al. 2012). Machine learning tasks are broadly classified into supervised and unsupervised learning. In supervised learning, a set of labelled input data for training the algorithm and observed output data are available. However, in unsupervised learning, only the unlabelled or observed output data is available. The goal of supervised learning is to train an algorithm to automatically map the input data to the given output data. When trained, the machine would have learned to see an input data point and predict the expected output. The goal of unsupervised learning is to train an algorithm to find a pattern, correlation, or cluster in the given dataset. It can also act as a precursor for supervised learning tasks (Bhardwaj et al. 2015). Several algorithms have been used in stock price direction prediction. Simpler techniques such as the single decision tree, discriminant analysis, and naïve Bayes have been replaced by better-performing algorithms such as Random Forest, logistic regression, and neural networks (Ballings et al. 2015). With nonlinear, data-driven, and easy-to-generalize characteristics, multivariate analysis through the use of deep Artificial Neural Networks (ANNs) has become a dominant and popular analysis tool in the financial market analysis (Zhong and Enke 2017). Recently, deep nonlinear neural network topologies are beginning to attract attention in time series prediction (Bao et al. 2017).

Sentiment analysis is another approach which has lately been used for stock market analysis (Bollen et al. 2011). It is the process of predicting stock trends via automatic analysis of text corpuses such as news feeds or tweets specific to stock markets and public companies. The sentiment classification techniques are mainly divided into machine learning approach and lexicon-based approach which is further divided into dictionary-based or corpus-based approaches (Bhardwaj et al. 2015). Seng and Yang (2017) demonstrated the potential of using sentiment signals from an unstructured text for improving the efficiency of models for predicting volatility trends in the stock market.

The hybrid approach applies a combination of multiple different approaches for improved performance, for example, a hybrid of statistical and pattern recognition approaches, or a hybrid of statistical and machine learning approaches. The next section describes some of the recent work on different algorithms and methods under each class in Figure 1, and additionally discusses several hybrid approaches which have been used in stock market prediction.

3. Literature Survey

Based on the taxonomy shown in Figure 1, this paper presents a literature study on some of the most popular techniques that have been applied for stock prediction.

3.1. Statistical Approach

Numerous statistical techniques have been tried and tested for stock market analysis and prediction. The Exponential Smoothing Model (ESM) is a popular smoothing technique which is applied on time series data, it essentially uses the exponential window function for smoothing time series data and analyse the same (Billah et al. 2006). De Faria et al. (2009) compared the ANN and adaptive ESM model for predicting the Brazilian stock indices. Their experiment revealed the predictive power of ESM and the results for both methods show similar performances although the neural network model i.e., the multilayer feedforward network slightly outperformed the adaptive ESM in terms of the Root Means Square Error (RMSE). Dutta et al. (2012) took an interesting path by selecting the financial ratios as independent variables for a logistic regression model and then analysed the relationship between these ratios and the stock performances. The paper focused on a classification task for predicting companies which are good or poor based on their one-year performance. The results show that the financial ratios—like net sales, book value PE, price-to-book (P/B), EBITDA, etc.—classify the companies into good and poor classes with an accuracy of 74.6%, which is a good indication of why company health matters in stock analysis and prediction.

Devi et al. (2013) tried to address some issues not currently addressed in most of the stock analysis literature, such as the dimensionality and expectancy of a naïve investor. The authors essentially utilize the historical data of four Indian midcap companies for training the ARIMA model. The Akaike Information Criterion Bayesian Information Criterion (AICBIC) test was applied to predict the accuracy of the model. Testing the model on individual stocks and the Nifty 50 Index showed that the Nifty Index is the way to go for naïve investors because of low error and volatility.

Ariyo et al. (2014) explore the extensive process of building ARIMA models. To identify the optimal model out of all the ARIMA models generated, the authors chose criteria like the standard error of regression, adjusted R-square, and Bayesian information criteria. The best ARIMA model, based on the above criteria, did a satisfactory job in predicting the stock prices of Nokia and Zenith Bank. Furthermore, Ariyo et al. (2014) made a solid case not to undermine the powers of ARIMA models in terms of stock analysis because it can compete reasonably well against the emerging forecasting techniques available today for short term prediction.

Bhuriya et al. (2017) implemented variants of regression models to predict the stock price of Tata Consultancy Services stock based on five features i.e., open, high, low, close price, and volume. The paper compares the performances of the linear, polynomial, and Radial Basis Function (RBF) regression models based on the confidence values of the predicted results. In addition, Bhuriya et al. (2017) reported that the linear regression model outperformed the other techniques and achieved a confidence value of 0.97.

3.2. Pattern Recognition

Pattern recognition techniques do pattern matching to identify future trends based on historical templates. Fu et al. (2005) suggested an approach to identifying patterns in time series data more efficiently using human visualization concept of PIP. The results from their experiments suggest that the PIP approach not only reduces dimensionality but also allows for early detection of patterns when

compared to template matching, because it uses a subsequence pattern matching approach by slicing time series data using the sliding window approach.

Leigh et al. (2008) challenged the EMH (Fama 1970) theory by showing that profits obtained using the heuristic method would be better than trading randomly. They utilized a bull flag pattern, which indicates a rise in prices in the near future and built a recognizer for identifying this pattern using template matching. The technique was applied on 9000 trading days of NYSE closing prices and the results show that the trading approach beats the average market profit most of the times, hence reinforcing the credibility of the technical analysis.

Parracho et al. (2010) proposed an approach to combine template matching with Genetic Algorithms (GA) for creating an algorithmic trading system. Template matching is utilized to identify upward trends and the GA helps in identifying the optimal values for the parameters used in template matching, i.e., fit buy, fit sell, noise removal, and window size. The trading strategy is trained on the S&P 500 stock data from 1998–2004 and tested on the 2005–2010 data. The results show that it outperforms the buy and hold strategy on an index and gets decent results for the individual stocks as well when compared to the buy and hold strategy.

Phetchanchai et al. (2010) proposed an innovative approach to analyse financial time series data by considering the zigzag movement in the data. In order to identify the Zigzag movements, the PIP technique was selected and the Zigzag based Mary tree (ZM-tree) was used for organizing these important points. The proposed technique illustrates a better performance in dimensionality reduction than existing techniques like Specialized Binary Trees (SB-Tree).

Cervelló-Royo et al. (2015) proposed a chart pattern based trading rule using the flag pattern. The study extends previous work by introducing two new parameters, stop loss and take profit, which allows the dynamic modelling of the closing of operations. It also employs intraday data to allow considerable width in the number of observations in the sample. Furthermore, Cervelló-Royo et al. (2015) considered both the opening and the closing prices to widen the information scope when deciding whether or not to start an operation. According to the authors, the results confirmed the positive performance of the flag pattern over the intraday data of the Dow Jones Industrial Average (DJIA) for a time horizon of more than 13 years. The results were also validated using two leading European indexes: the German stock index or Deutscher Aktienindex (DAX) and the Financial Times Stock Exchange (FTSE). It also provides empirical evidence which confronts the EMH (Fama 1970) indicating how it is possible to develop an investment strategy capable of beating the market in the mean-variance sense.

Chen and Chen (2016) proposed a hybrid approach to identify bull flag patterns on the Taiwan Capitalization Weighted Stock Index (TAIEX) and National Association of Securities Dealers Automated Quotations (NASDAQ) indices. The authors developed a methodology that combines the advantages of two traditional pattern recognition methods (PIP and template matching). Their proposed hybrid approach outperformed the other models like Rough Set Theory (RST), GA, and a hybrid model of GA and RST (Cheng et al. 2010) by a good margin in terms of total index returns.

Arévalo et al. (2017) offer a robust mechanism to dynamically trade DJIA based on filtered flag pattern recognition using template matching, based on the initial work of Cervelló-Royo et al. (2015). The authors impose multiple filters before considering the flag patterns as actionable for making trades, based on Exponential Moving Averages (EMA) and price ranges of the detected patterns. Their approach performs much better than the base approach of Cervelló-Royo et al. (2015) and the buy and hold strategy, resulting in higher profit and lower risk.

Kim et al. (2018) build a Pattern Matching Trading System (PMTS) based on Dynamic Time Warping (DTW) algorithm in order to trade index futures on the Korea Composite Stock Price Index (KOSPI 200). Taking the morning 9:00–12:00 p.m. time series data as input for the sliding windows, the authors then use DTW in order to match with known patterns. This forms the basis of the trading strategy to be carried in the afternoon's session on the same trading day. Their approach generates good annualized returns and shows that most patterns are more profitable near the clearing time.

3.3. Machine Learning

Many machine learning techniques have been explored for stock price direction prediction (Ballings et al. 2015). ANN and Support Vector Regression (SVR) are two widely used machine learning algorithms for predicting stock price and stock market index values (Patel et al. 2015). A literature survey of supervised and unsupervised machine learning methods applied in stock market analysis will be presented next.

3.3.1. Supervised Learning

Supervised learning techniques like Support Vector Machine (SVM) and Decision Trees can learn to predict stock market prices and trends based on historical data and provide meaningful analysis of historical price. Bernal et al. (2012) implemented a subclass of Recurrent Neural Networks (RNN) known as Echo State Networks (ESN) to predict S&P 500 stock prices using price, moving averages, and volume as features. The technique outperforms the Kalman Filter technique with a meagre test error of 0.0027. In order to generalize and validate their result, Bernal et al. (2012) examined the algorithm on 50 other stocks and reported that their results performed well against state of the art techniques.

Ballings et al. (2015) benchmark ensemble methods consisting of Random Forest, AdaBoost, and Kernel Factory against single classifier models such as Neural Networks, Logistic Regression, Support Vector Machines, and K-Nearest Neighbor using data from 5767 publicly listed European companies. The authors used five times two-fold cross-validation and Area Under the Curve (AUC) as a performance measure for predict long term stock price direction and reported Random Forest as the top algorithm.

Milosevic (2016) proposed an approach for long term prediction of stock market prices through a classification task where a stock is 'good' if the stock price increases by 10% in a year otherwise it is a 'bad'. Furthermore, Milosevic (2016) performed a manual feature selection, selected 11 relevant fundamental ratios, and applied several machine learning algorithms to stock prediction. It follows that Random Forest achieved the best F-Score of 0.751 against techniques such as SVM and Naïve Bayes.

Another technique that has grappled the attention of data scientists is the eXtreme Gradient Boosting (XGBoost). Dey et al. (2016) predicted the direction of stocks based on XGBoost algorithm using the technical indicators as the features. The results show that XGBoost beats the other techniques in performance achieving an accuracy of 87–99% for long term prediction of Apple and Yahoo stocks.

Long Short-Term Memory (LSTM) network have shown a lot of promises for time series prediction; Di Persio and Honchar (2017) applied three different Recurrent Neural Network models namely a basic RNN, the LSTM, and the Gated Recurrent Unit (GRU) on Google stock price to evaluate which variant of RNN performs better. It was evident from the results that the LSTM outperformed other variants with a 72% accuracy on a five-day horizon and the authors also explained and displayed the hidden dynamics of RNN. Roondiwala et al. (2017) implemented an LSTM network to predict Nifty prices with features like OHLC. Their results show that the LSTM achieves an RMSE of 0.00859 for the test data in terms of daily percentage changes.

Yang et al. (2017) proposed an ensemble of multi-layer feedforward networks for Chinese stock prediction. Three component networks were trained using training algorithms like backpropagation and Adam. The ensemble was formed using the bagging approach (Efron and Tibshirani 1994). The results obtained demonstrate that the Chinese markets are partially predictable and achieve a satisfactory accuracy, precision, and recall.

Zhang et al. (2018) propose a stock price trend prediction system that can predict both stock price movement and its interval of growth (or decline) rate within predefined prediction durations. They trained a random forest model from historical data from the Shenzhen Growth Enterprise Market (China) to classify multiple clips of stocks into four main classes (up, down, flat, and unknown) according to the shapes of their close prices. Their evaluation shows that the proposed system is robust to the market volatility and outperforms some existing predictions methods in terms of accuracy and return per trade.

Hossain et al. (2018) propose a deep learning-based hybrid model that consists of two well-known DNN architectures: LSTM and GRU. The authors trained a prediction model using S&P 500 time series dataset spanning about 66 years (1950 to 2016). The approach involves passing the input data to the LSTM network to generate a first level prediction and then passing the output of LSTM layer to the GRU layer to get the final prediction. The proposed network achieved a Mean Squared Error (MSE) of 0.00098 in prediction with outperforming previous neural network approaches.

Recently, Lv et al. (2019) synthetically evaluated various ML algorithms and observed the daily trading performance of stocks under transaction cost and no transaction cost. They utilized 424 S&P 500 index component stocks (SPICS) and 185 CSI 300 Index Component Stocks (CSICS) between 2010 and 2017 and compared traditional machine learning algorithms with advanced deep neural network (DNN) models. The traditional machine learning algorithms are SVM, Random Forest, Logistic Regression, naïve Bayes, Classification and Regression Tree (CART), and eXtreme Gradient Boosting while the DNN architectures include Multilayer Perceptron (MLP), Deep Belief Network (DBN), Stacked Autoencoders (SAE), RNN, LSTM, and GRU. Their results show that traditional machine learning algorithms have a better performance in most of the directional evaluation indicators without considering the transaction cost, however, DNN models show better performance considering transaction cost.

3.3.2. Unsupervised Learning

Unsupervised learning helps to identify correlations in an uncorrelated dataset like stock markets. Powell et al. (2008) drew a comparison between the supervised technique SVM and unsupervised technique K-means. They perform Principal Component Analysis (PCA) to reduce the dimensions or features. Both models are tested on S&P 500 data and the results show that both techniques have similar performance, SVM achieves 89.1% and K-means achieves 85.6% respectively. They also state how different distance measures for clustering affect the prediction accuracy and the best performance is shown by the Canberra distance metric.

Babu et al. (2012) proposed a clustering method called the HAK by combining the Hierarchical Agglomerative Clustering (HAC) and the reverse K-means clustering to predict the short-term impact on stocks after the release of financial reports. The study compared three different clustering techniques namely the HAC, K-means, and the reverse K-means. It also compared the proposed HRK against the three techniques and the SVM. Firstly, HAK takes financial reports and stock quotes as input and uses text analysis to convert each financial report into a feature vector. It then divides the feature vectors into clusters using HAC. Secondly, for each cluster, the K-means clustering method was applied to partition each cluster into sub-clusters and for each sub-cluster the centroids were computed. Finally, the centroids were used as the representative feature vectors to predict stock price movements. The experimental results show that the proposed technique outperforms SVM in terms of accuracy.

Wu et al. (2014) proposed a model based on the AprioriAll algorithm (association rule learning) and K-means. They converted stock data into charts using a sliding window and then the charts were clustered using K-means to extract chart patterns. Frequent patterns were extracted using AprioriAll to predict trends that are often associated (bought or sold together). The results show that their model outperforms other related work (Wang and Chan 2007; Chen 2011) in terms of average returns and also mutual funds.

Peachavanish (2016) proposes a clustering method to identify a group of stocks with the best trend and momentum characteristics at a given time, and therefore are most likely to outperform the market during a short time period. The author conducted an experiment on five-year historical price data of stocks listed on the Stock Exchange of Thailand (SET) and reported that the proposed method can outperform the market in the long run. Table 1 presents a summary of the existing literature on supervised ML approaches.

Table 1. Summary of literature on ML approaches applied to stock prediction.

Paper	Dataset	Features	Technique	Prediction Type	Metrics	Results
Leigh et al. (2008)	NYSE	Price	Template Matching	Daily	Average profits	3.1–4.59%
Bernal et al. (2012)	S&P 500	Price, MA, volume	ESN RNN	Daily	Test Error	0.0027
Milosevic (2016)	1700+ individual stocks	Price, 10 financial ratios	Random Forest vs. SVM vs. NB vs. Logistic Regression	Classification (good vs. bad)	Precision, Recall and F-score	0.751 (Random Forest)
Dey et al. (2016)	Apple, Yahoo	Technical indicators	XGBoost vs. SVM vs. ANN	Daily	Accuracy	85–99% (XGBoost)
Di Persio and Honchar (2017)	Google Stock	OHLCV	RNN vs. LSTM vs. GRU	Daily, Weekly	Log loss, accuracy	72%, 5 day (LSTM)
Yang et al. (2017)	Shanghai composite index	OHLCV	Ensemble of DNN's	Daily	Accuracy, relative error	71.34%
Zhang et al. (2018)	Shenzhen GE Market	Price trends	Random Forest	Classification (up, down, flat, and unknown)	Return per trade	75.1%

3.4. Sentiment Analysis

Sentiments can drive short-term market fluctuations which in turn causes a disconnect between the stock price and the true value of a company's share. Over long periods of time, however, the weighing machine kicks in as the fundamentals of a company ultimately cause the value and market price of its shares to converge. Sentiments are a big part of stock markets and the idea of analysing sentiments based on various data sources can give insights on how stock markets react to different kinds of news in the immediate and medium term. Hence a novel approach—sentimental analysis—has emerged which gauges the sentiment from data sources or sentiment behind the news to identify its impact on the markets.

Schumaker and Chen (2009) determined the effects of breaking news on stock prices within 20 min after the release. An SVM derivative model was proposed based on three different textual representations namely the Bag of Words (BoW) model, the Noun Phrases model, and the Named Entities model. First, news data was gathered and stored in a database using the three different textual representations. Next, the authors fetched the closing prices of the respective stocks for the last 60 min and used Support Vector Regression to predict the price for the next 20 min based on the price and sentiment analysis. The experiments showed that the model significantly outperformed the simple linear regression model in terms of closeness, directional accuracy and simulated trading. The authors also stated that noun phrases method performed much better when compared to the bag of words and named entities models.

Bollen et al. (2011) were one of the first studies to have considered Twitter data to predict stock trends. They analysed the Twitter data using Google Profile of Mood States (GPOMS) and Opinion Finder to understand correlations and predict DJIA closing prices. In addition, Bollen et al. (2011) cross-validated the Opinion Finder and GPOMS time series against popular events like presidential election and thanksgiving to gauge the public mood. They applied a granger causative analysis to determine whether one time series is dependent on another. After applying a Self-Organizing Fuzzy Neural Network (SOFNN) on a dataset of approximately 10 million tweets and DJIA closing prices from February to December 2008, the technique achieved an accuracy of 87.6% in predicting daily values of DJIA.

Mittal and Goel (2012) based their work on the study by Bollen et al. (2011), however they tested the proposed approach on a larger dataset of over 400+ million tweets. They preprocessed missing days data consisting of weekends and public holidays by replacing the missing values with the average value. They then applied a portfolio management strategy based on the Bollen et al. (2011) technique. They adjusted sharp rise and fall in prices to get a steady daily directional trend; they also removed

data of volatile periods because of the inherent difficulty to predict such periods and made use of four classes of moods instead of two namely happy, alert, calm, and kind. Their technique achieved an accuracy of 75% and the portfolio management strategy achieved a decent profit over a 40-day testing period. Moreover, their work gave further insight that not only 'calmness' but also 'happiness' is granger causative over a range of three to four days.

Lee et al. (2014) proposed an approach to determine the importance of text analysis in stock market prediction. Form 8-K reports include important updates regarding the company. The authors created a system to predict whether a stock's price will go up, down, or stay the same by performing sentiment analysis on the 8-K reports of the respective stocks. Lee et al. (2014) ran their model first with only financial features and then again using financial and linguistic features (unigrams). It was observed that the most important feature was the 'earnings surprise' and that the text analysis helped to improve the model's accuracy by 10%. Their work also put forth an interesting finding that the effect of sentiment analysis on the 8-K reports diminishes quickly with time. Therefore, these would only be suitable for short term predictions.

Kalyanaraman et al. (2014) proposed a sentiment analysis model to gauge sentiments from news articles and feed the output from the model into two different machine learning algorithms. The authors made use of Bing API to get the news for a set of companies. They created their own dictionary to categorize positive and negative sentiments with respect to the stock market domain due to the lack of such open source dictionaries. The words in the article were compared against the dictionary and were tagged as positive and negative along with their frequencies. For example, a score of −3 would be assigned if a negative word appeared three times in the article. The output from the model was then fed into a linear regression model which used gradient descent for optimization. Results showed that the machine learning model using gradient descent was able to predict the sentiment of a news article with an accuracy of 60% when compared to manual analysis. Moreover, the ML model achieved an accuracy of 81.82% for predicting the actual stock prices.

Cakra and Trisedya (2015) tried to predict the price, price fluctuation, and margin percentage of Indonesian stocks using a simple sentiment analysis model coupled with classification techniques and a linear regression predictive model. The authors classified tweets into 'positive', 'negative', and 'neutral' classes. The work disregarded 'neutral' tweets as they were considered to be promotional and spam tweets. They retrieved lexicon sentiment semi-automatically from the data. First the corpus was tokenized into single words using the Indonesian dictionary. Next, formal words that were categorized as lexicons were chosen and informal words were manually checked by searching similar words. Then the positive and negative lexicons were separated. Their experiments suggested that the Random Forest algorithm produced the best result in classifying tweets amongst other algorithms with an accuracy of 60.39%. Cakra and Trisedya (2015) achieved an accuracy of 67.37% on price fluctuation prediction based on the classification of the tweets data using the Naive Bayes algorithm and 66.34% using the Random Forest algorithm.

Pagolu et al. (2016) implemented a sentiment analysis model based on Twitter data. The authors used N-gram and Word2vec (2-layer NN) to analyse the polarity of sentiments behind the tweets. They achieved an accuracy of around 70% and noted that the correlation between price and sentiments was 71.82%. The study concluded that with more data, the accuracy of the model would increase. Other approaches that utilized natural language and social media data include the study by Xu and Cohen (2018) which present StockNet, a neural network architecture for predicting stock price movement from tweets and historical stock prices. The model demonstrates a state-of-the-art performance and introduces recurrent, continuous latent variables for better treatment of stochasticity. Table 2 presents a summary of the literature study on sentiment analysis approaches.

Table 2. Summary of literature on sentiment analysis approaches applied to stock prediction.

Paper	Dataset	Technique	Prediction Type	Metrics	Results
Schumaker and Chen (2009)	News articles, S&P 500	Bag of words vs. noun phrases vs. noun entities → SVM	Daily	Returns, DA	2.57% (Noun phrases)
Bollen et al. (2011)	DJIA, Twitter data	Mood Indicators → SOFNN	Daily	Accuracy	87.14%
Lee et al. (2014)	8-K Reports, Stock prices, volatility	Ngram → Random Forest	Daily, long term	Accuracy	>10% (Increase in accuracy)
Kalyanaraman et al. (2014)	News articles (Bing API)	Dictionary approach → Linear Regression	Daily	Accuracy	81.82%
Pagolu et al. (2016)	MSFT price, Twitter data	Ngram + word vec → Random Forest	Daily	Accuracy	70.1%

3.5. Hybrid Approach

The hybrid approach applies a combination of multiple different approaches, for example, statistical and pattern recognition approaches, or statistical and machine learning approaches. Markowska-Kaczmar and Dziedzic (2008) implemented a supervised feedforward neural network technique to identify patterns in stock data and use the PIP technique to reduce the dimensionality and find only the important points of the patterns. The PIP technique was found to doing a fair job in discovering patterns on shortened time series datasets.

Tiwari et al. (2010) proposed an intriguing hybrid model by combining the statistical Hierarchical Hidden Markov model (HHMM) with supervised learning technique using decision trees to predict the Bombay Stock Exchange Sensitive Index (BSE SENSEX) trend based on its historical closing prices, dividends and earnings. After the extraction of features from the dataset, relevant features were selected by a decision tree. Then a set-based classifier was used for prediction while the HHMM was used to evaluate the predictions and for generating the final predictions which yield an accuracy of 92.1%.

Shen et al. (2012) proposed a prediction algorithm which combined statistical and SVM approaches. The technique exploits correlations amongst global markets and other products to predict the next day trend of stock prices. The authors chose varied datasets which might have an impact on each other such as the United States Dollar (USD), Japanese Yen (JPY), NASDAQ, gold, and oil prices, and drilled down important features via statistical methods like auto and cross correlation. Results of this study boasted a 77.6% prediction accuracy on the DJIA and up to 85% for longer-term predictions of more than 30 days.

Wang et al. (2012) present a Proposed Hybrid Model (PHM) which combines three models namely the ESM, ARIMA, and a Backpropagation Neural Network (BPNN) model, to leverage the strength of each model in predicting stock prices on a weekly basis. Results from this study show that the hybrid model outperforms the performances of the individual constituent sub-models and all traditional models when tested on the Shenzhen Integrated Index and DJIA with a directional accuracy of 70.16%.

Yoshihara et al. (2014) proposed an approach that considered the long-term effects of news events. The study employs a combination of the Recurrent Neural Networks Restricted Boltzmann Machine (RNN-RBM) and the Deep Belief Network (DBN) for predicting stock trends in a binary approach, i.e., up or down. Events represented as vectors using the bag of words model were given as inputs to the model. The results were compared against SVM and DBN and the proposed model achieved the lowest error rates amongst all.

Ding et al. (2015) proposed a hybrid approach which combined sentiment analysis and neural network models for the prediction of S&P 500 index. News events were represented as vectors and a deep Convolutional Neural Network (CNN) was trained to predict short and long-term influences of

those events on the index. The study uses more than 10 million events over seven years and achieves an accuracy of 64.21% on the S&P 500 index and 65.48% on the individual stock price prediction.

Rather et al. (2015) proposed a hybrid model consisting of both the linear and non-linear approaches for predicting stock prices. The work combines the results of the component models namely the ARIMA, ESN, and the RNN models. The weights of each of the constituent models, which represent their effects on the prediction result were determined using a genetic algorithm. The proposed hybrid model achieves the lowest Mean Absolute Error (MAE) and MSE compared to the constituent models and outperforms the RNN in terms of price prediction.

Creighton and Zulkernine (2017) extended the original work of Wang et al. (2012) by applying the hybrid approach to daily stock price prediction and on different indices such as the S&P 400 and the S&P 500. The study by Creighton and Zulkernine (2017) showed that for the daily predictions, the hybrid model did not outperform its constituent models and the BPNN model gave the most accurate predictions. The statistical ARIMA and ESM models including the combined hybrid PHM could predict better for longer time range but suffered from price fluctuations when applied to the daily predictions. Table 3 presents a summary of the literature study on hybrid approaches.

Table 3. Summary of literature on hybrid approaches applied to stock prediction.

Paper	Dataset	Features	Technique	Prediction Type	Metrics	Results
Wang et al. (2012)	DJIA and SJI Index	Price	ESM + BPNN + ARIMA	Weekly	Directional Accuracy	70.16%
Tiwari et al. (2010)	Sensex + 3 stocks	Price, EPS and DPS	HHMM + Decision Trees	Daily	Accuracy	92.1%
Shen et al. (2012)	Indices, commodities	Asset prices	Auto, cross correlation + SVM	Daily, monthly	Accuracy	77.6%
Rather et al. (2015)	NSE stocks	Price, mean, SD	ARIMA + ESM + RNN + GA	Daily	Avg MSE, MAE	0.0009, 0.0127
Yoshihara et al. (2014)	Nikkei stocks, news articles	Word vectors	Bag of Words → DBN + RNN-RBM vs. SVM vs. DBN	Long term	Test error rates	39% (Lowest)
Ding et al. (2015)	S&P 500	Historical events	NN (event embeddings) + CNN	Weekly, Monthly	Accuracy & MCC	64.21%

4. Discussion

4.1. Statistical

It is evident from the surveyed literature that, despite the emergence of many techniques for stock prediction, statistical methods like ARIMA, ESM, Regression and their variants continue to be of interest for stock market forecasting due to their performance. For example, De Faria et al. (2009) provided a nice comparison between Adaptive ESM and NN which show that both models perform equally well except for the hit rates for forecasting stock direction where the NN does better. However, the study by De Faria et al. (2009) failed to provide information regarding the dataset and features used for the models. Nevertheless, it shows the power of statistical models and how they are still competing with emerging techniques like deep learning and hybrid models. Statistical models, in general, assume that there is a linear correlation structure in time series data. This is a limitation that emerging techniques are overcoming through combining statistical and machine learning or other techniques (Wang et al. 2012; Shen et al. 2012). One class of statistical model that is useful for understanding the risk or volatility in stock trading is ARIMA, this was demonstrated by Devi et al. (2013), that the Nifty index is a much better indicator of the volatility of stocks. Ariyo et al. (2014) predicted stock prices based on ARIMA and their results were convincing for stocks like Nokia and Zenith bank. However, the metrics that

lead to the results could have been explained better and testing on more stocks would have provided a better picture.

4.2. Pattern Recognition

In general, pattern recognition techniques show promises but on their own do not give convincing results on stock prediction Velay and Daniel (2018). These techniques can be powerful for analysing and mining patterns rather than predicting the actual values. Therefore, instead of using pattern recognition techniques in isolation for stock prediction, it would be better if they were just used for identifying trends or in combination with prediction techniques. A recent pattern recognition model developed by Chen and Chen (2016) to ascertain the bull-flag patterns contained in historical stock patterns generated an unprecedented stock index return (TIR% and TIR) in forecasting the NASDAQ and TAIEX, indicating that the model can help stock analysts or stock investors to check stock patterns more carefully. This work seems to be promising and very thorough as they also consider the number of trades required for achieving such high TIR. However, their outperformance over other models seems to have two caveats. First, the proposed approach gives varying results depending upon the underlying index, which the authors address in the discussion section. Nonetheless, it would be valuable to see if their approach is able to generalize well for finding bull flag patterns across different stocks or indices and therefore achieve similar results as it did for TAIEX index. Second, it would have been interesting if the authors had shed more light as to why the 20-day holding variant was able to beat other models when compared to 5, 10, and 15-day holding variants.

Another interesting study that utilized a pattern recognition approach is the study by Arévalo et al. (2017) which suggest improvements to the trading strategy introduced by Cervelló-Royo et al. (2015). Arévalo et al. (2017) use the price ranges in which the detected bull/bear flag patterns lie in and test against the EMA (short and medium term) to successfully filter out patterns which would be worthwhile for making the trades in the first place. Most technical analysis studies do not take this into account. Arévalo et al. (2017) considered transaction costs and risk in order to determine the model's success. They apply robust and dynamic rules to make sure the take profit and stop loss levels for trade are not restricted to holding it for an x amount of days. Furthermore, they made sure that the take profit and stop loss levels are updated quarterly based on the best recent past performance of the trading rule. The authors use the OHLC dataset for DJIA and use intraday observations from 2000 to 2013 to test their trading strategy. The results indicate that their new and improved strategy performs better than the prior strategy (Cervelló-Royo et al. 2015) with higher profit and lesser risk which is commendable. The authors also test their strategy to make sure they pass the data snooping test (training on test data). Lastly, this study addresses all the parameters that Park and Irwin (2007) point out as being potential problems while determining profitability in technical analysis.

Most recently, Kim et al. (2018) propose a PMTS-based on dynamic time warping. The authors demonstrated how their PMTS performs with different parameters and found the most optimal set of parameters for trading resulting in higher Sharpe ratios. The strategy fetches above par annualized returns of about 9.58% returns with the most optimal parameter set. The authors claim that traders can use their PMTS to make more efficient trading decisions in order to help the markets to become more efficient. This claim seems to be a bit far-fetched because a lot of factors contribute to the market movements and their implicit assumption that the PMTS would always hold true for future market movements needs more evidence. Lastly, it would be interesting to see how their strategy would perform on normal stock trading instead of trading index futures and whether the later plays a role in the strategy's outcome.

4.3. Machine Learning

4.3.1. Supervised Learning

Machine Learning approaches, specifically supervised learning for stock prediction have shown great promises. Here we will discuss some recent and interesting results and highlighting their strengths or limitations. First, the study by Bernal et al. (2012) which evaluated an ESN network for stock prediction but do not give any reason behind comparing its performance against the Kalman Filter approach. Also, it is interesting to note that there is no increase in error as they forecast further into the future, which makes sense because the underlying dynamics of the stock were the same for the training and testing data. Thus, the paper does not consider the different dynamics that can prevail during the training and testing data periods respectively, which would render the approaches futile. A good way to work around the problem noted above would be to take a larger dataset which would cover a broader and varied set of stock market dynamics.

Another study by Milosevic (2016) applied a manual feature selection amongst financial ratios to understand identify optimal features. It is a wise approach because many ratios give little or no information regarding stock price movements. Furthermore, the author tests on 1700+ stocks and compared SVM, Naïve Bayes, and Logistic Regression with Random Forest; and found out that Random Forest outperforms the other approaches with a recall of 0.751. However, the methodology of classifying a stock as good or bad based on just one-year return being greater than 10% is not ideal. It should be compared with a benchmark, as done by Bernal et al. (2012) and Dutta et al. (2012).

Dey et al. (2016) explore a novel approach of Extreme Gradient Boosting (XGBoost) and compared it with SVM and ANN, in which XGBoost achieve superior results. However, the technique is tested on only two stocks and the model seems to be overfitting in the case where accuracy is 99% for Yahoo's stock prediction. Additional testing and transparency on overfitting would make a stronger case for their results. Di Persio and Honchar (2017) demonstrated why the LSTM variant is so popular and why it works better than the other RNN approaches. The authors further provided informative visualisations, which was able to explain how and when an RNN detects trends. This is a very interesting result. Another study that utilised LSTM is Roondiwala et al. (2017), however, the authors could not provide information about how they were able to prevent overfitting after training their LSTM network for 500 epochs.

Yang et al. (2017) used an ensemble of deep neural networks to predict Chinese stock markets. They utilized backpropagation and Adam training algorithms, however while training the network, the dataset was split into training and testing datasets only with no validation set, which is important for unbiased training of the network. The authors did not normalise the dataset (OHLC), which is also an important preprocessing step that can affect training algorithms as well as prediction results. Again, the ensemble model could not provide satisfactory predictions for an important feature i.e., closing prices of the indices. Nevertheless, the study highlights how an ensemble network performs better than its component networks for stock prediction. The model achieves an accuracy of 74.15% on high and low of Shanghai composite index and 73.95% and 72.34% respectively for the Shenzhen Stock Exchange component index. In terms of resources consumption, Bao et al. (2017) have observed that deep learning methods are time-consuming, and more attention should be paid to GPU-based and heterogeneous computing-based deep learning methods.

4.3.2. Unsupervised Learning

Unsupervised learning methods have been known to be equally strong for stock prediction. Powell et al. (2008) drew an effective comparison between SVM and K-means and explained which distance metric for K-means and which kernel function for SVM achieves best results. Additionally, for K-means it was observed that maximum accuracy is achieved with one feature, although they do not mention what other features were used, and which one feature worked out the best, post PCA.

Babu et al. (2012) proposed a clustering method called Hierarchical Agglomerative Clustering (HAC) and Reverse K-means clustering (HRK) to predict the short-term impact on stocks after the release of financial reports. Their experimental results show that the proposed technique outperforms K-means as well as the SVM in terms of accuracy, and the average profits for the HRK were 3.95% whereas SVM could only manage a profit of 1.46%. The paper nicely combines two clustering techniques to get the best features out of them and applied text analysis on financial reports to understand the impact of fundamental factors similar to Dutta et al. (2012).

Wu et al. (2014) proposed a model based on the AprioriAll algorithm (association rule learning) and K-means. This approach identifies patterns, initiates a buy position and then holds it until the end of the pattern. The results show that the proposed model outperforms the other related work (Wang and Chan 2007; Chen 2011) generating an average return of 2.22% compared to the 1.67% achieved by Wang and Chan (2007) and 1.5% achieved by Chen (2011) respectively. This is a particularly fascinating approach because not only does it outperform the other approaches, it does so with fewer trades and thus incurs lower trading costs. While comparing against mutual funds, the authors could not explain how they computed the annual returns from the 20-day return, because it seems to be outperforming the mutual funds by a huge margin. Another limitation is the assumption that the stock market patterns are repetitive, which may be true to some extent, but not always. Furthermore, if new patterns arise then the model needs to be retrained.

4.4. Sentiment Analysis

Sentiment analysis on social media is not an easy task because it is difficult to teach machines all the different contexts of how people express their emotions or opinions. For example, 'My flight has been delayed, Superb!', a human can quickly sense the sarcasm in this text, however a machine might identify this as a positive statement. For example, Bollen et al. (2011) used Twitter data to measure the correlations between the public comments and the changes in DJIA. The authors used an effective filtering technique to remove spam tweets, i.e., tweets are discarded if they contain the regular expression 'www.' or 'http://'. The 'calm' mood along with the historical prices as input to the SOFNN give 86.7% accuracy. However, the general assumption that the overall public mood affects the stock prices is naive because not all people who tweet invest in the stock markets (Pagolu et al. 2016). Hence a dataset of tweets pertaining to the general stock markets, specifically the DJIA, would be more indicative of the price changes in DJIA. The authors actually discussed these limitations and stated that by no means is the public mood is a good predictor of the changes in DJIA but that it may have some correlation. Mittal and Goel (2012) proposed their approach based on the work of Bollen et al. (2011). The difference between their work and Bollen et al. (2011) is that they chose a much larger dataset and observed that even 'happiness' is indicative of the DJIA prices. They also apply K-fold sequential cross-validation, which is apt for financial data and implemented a basic portfolio strategy which performs well. However, the accuracy achieved was lower than what Bollen et al. (2011) had reported.

Lee et al. (2014) proposed a text analysis model on the company's 8-K reports, which are more relevant compared to social media data. The authors demonstrated that text analysis helps to improve the prediction accuracy by 10%. According to Lee et al. (2014), the updates in 8-K reports have a short-term effect on stock prices, which makes sense because these reports contain company updates and a very few of them can affect the stock prices on a longer term.

Kalyanaraman et al. (2014) targeted a more reliable source of text data by creating their own dictionary to analyse sentiment polarity of sentences in news articles. The authors achieve an accuracy of 81.81% using linear regression with gradient descent optimisation. However, the accuracy is calculated over a long time period encompassing all selected news articles for each stock, and hence the results do not represent the true ability of the model to predict stock prices based on sentiment analysis of the news articles.

Cakra and Trisedya (2015) used five different algorithms for tweet classification, and the output is fed to a linear regression model for predicting Indonesian stock prices. A better alternative could be to

try an ensemble of machine learning algorithms to better classify tweets and then feed it to a linear regressor model or a non-linear algorithm.

Finally, Pagolu et al. (2016) directly assessed tweets related to Microsoft's stock and its products and tried to predict the price of the stock for the next day. They achieved an accuracy of 71.82% with a small dataset containing only around 3000 tweets. The authors also addressed the issue that all Twitter participants are not in the stock market business, and plan to use a stock market specific social media platform to gather opinions of investors in their future work. This approach has two shortcomings; one, such techniques can be derailed by malicious or biased tweets; and two, Twitter data is already public and hence, provides an efficient and faster approach to access and process the tweets than other newly devised platforms. If not processed fast enough, social media data would already have influenced the stock prices, making the prediction useless.

4.5. Hybrid

Hybrid approaches combine multiple different approaches, for example, Markowska-Kaczmar and Dziedzic (2008) show how they effectively use the PIP approach for reducing dimensionality and identifying trends before they feed the data to a feedforward NN for prediction. A big disadvantage of PIP is that it would not accurately discover sequences if the time series has a high amplitude between two adjacent points.

Tiwari et al. (2010) proposed an intriguing hybrid model combining statistical Hierarchical Hidden Markov model (HHMM) with decision trees to predict BSE SENSEX trend which yields an accuracy of 92.1% making a strong point as to why hybrid approaches are powerful. However, the paper does not give detailed and transparent results, such as figures or tables describing how and why 92% accuracy was achieved. One of the results specified the predicted SENSEX value to be around 1.2567×10^3 i.e., 1256.7 in 2011 which should have been 12,567 instead. Again, the authors used a dataset with just 52 instances (yearly) partly because features like dividends and earnings are not declared daily, however using a quarterly approach would have given them a decent sized dataset to validate the results.

Wang et al. (2012) proposed a satisfactory PHM consisting of ARIMA, ESM and BPNN which give a directional accuracy of 70%. However, the results clearly demonstrated that BPNN outperforms statistical models like ARIMA and ESM for a weekly prediction of stock prices. Creighton and Zulkernine (2017) extended the work of Wang et al. (2012) for daily stock price prediction. Their results show that PHM is not as good in predicting daily prices compared to weekly prices and fortify the reasoning that daily price prediction is more difficult as it is susceptible to noise and price fluctuations including other factors. Both papers applied a genetic algorithm for deciding the weights of the component models in the PHM.

Shen et al. (2012) proposed a very good hybrid approach for feature selection based on statistical methods like auto and cross correlation, and use SVM algorithm to predict stock prices, which boasted a 77.6% prediction accuracy on the DJIA. For validation they predicted prices for the long term and got up to 85% accuracy. Also, to test the model's generalisability, the authors predicted on two other indices, the NASDAQ and the S&P 500. SVM in general is a powerful algorithm for a variety of problems, combining that with other statistical techniques like auto and cross correlation only makes it better as proposed by Shen et al. (2012).

Yoshihara et al. (2014) proposed a novel approach by combining DBN with RNN-RBM to better predict long term changes in stock prices based on significant events. The authors validated the distribution of data in train, test, and validation sets. The input to the model was the news events which were converted to vectors using the bag of words model. The results show that the hybrid model outperforms SVM and DBN achieving the lowest test error rate of 39%. The authors noted that if there are events which do not have long-term effects on stock prices then the proposed model's performance is similar to that of the DBN. Use of more performance metrics like accuracy or returns based on a trading strategy would have given more insights into the performances of their approach.

Ding et al. (2015) proposed a novel neural tensor network for learning event embeddings and used a deep CNN to model the influences of long and short-term events on stock price movements. The results show that event embedding based representations work better than other approaches. The performance is strongly validated as it is compared with other state-of-the-art approaches. Compared to other relevant work, the paper achieves a 6% improvement in predicting the S&P 500 index. The authors also simulated their approach for trading and the results outperformed other approaches in terms of returns by a good margin.

Rather et al. (2015) combined linear and non-linear approaches to design a hybrid model for stock prediction. The results show that their hybrid model outperforms RNN and is better able to predict rapid fluctuations in Indian stock prices. The authors, however, did not mention the motivation behind using genetic algorithms for selecting weights. Furthermore, the work chooses a 50:50 train–test split which is a little unusual for stock market datasets; therefore, it would have been better if they could provide any insights on the same.

Finally, it is important to take non-stationarity into consideration while evaluating forecasting models. Most of the literature today does not take non-stationarity into account. Shah et al. (2018) is a good example of how an LSTM-RNN based model is able to predict really well on non-stationary data. The work of Shah et al. (2018) shows that the LSTM model not only gives great results for daily predictions, i.e., one-day ahead but also gives more than decent results for predicting over longer-term horizons, i.e., 7-day ahead predictions just using daily price as a feature. The authors intentionally utilised a larger training dataset (20 years price data), since that period encompasses multiple up and down cycles of the market. Hence, this would allow the LSTM model to learn better and therefore have a fighting chance of being applicable to future time periods with similar market movements.

5. Challenges and Open Problems

Stock market analysis and prediction continue to be an interesting and challenging problem. As more data are becoming available, we face new challenges in acquiring and processing the data to extract knowledge and analyse the effect on stock prices. These challenges include issues of live testing, algorithmic trading, self-defeating, long-term predictions, and sentiment analysis on company filings.

Regarding live testing, most of the literature on stock analysis and prediction claim that the proposed techniques can be used in real time to make profits in the stock market. It is a big claim to make because an algorithm may work fine on backtesting in controlled environments, but the main challenge is live testing, because a lot of factors like price variations, and uneventful news and noise exist. One such example is the Knight Capital Tragedy[1] where the company suffered a loss of 440 million. Hence, a viable research direction would be to understand how some of the popular stock analysis techniques work in live or simulated environments.

Algorithmic trading systems have changed the way stock markets function. Most of the trading volumes in equity futures are generated by algorithms and not by humans. While algorithmic trading gives benefits like reduced cost, reduced latency, and no dependence on sentiments, it also brings up challenges for retail investors who do not have the necessary technology to build such systems. Today, it is common to see events where panic selling is triggered due to these systems and hence the markets overreact. As a result, it becomes more difficult to evaluate market behaviour. With new algorithms continuing to flood the markets every day, comparison of the efficacy and accuracy of these algorithms pose yet another challenge.

An interesting aspect of this research area on stock market prediction is its self-defeating nature. In simple words, if an algorithm can use a novel approach to generate high profits, then sharing it in any way to the market participants will render the novel approach useless. Thus, state of the art

[1] https://dealbook.nytimes.com/2012/08/02/knight-capital-says-trading-mishap-cost-it-440-million/.

algorithms which are trading out there in the markets is proprietary and confidential. The research or methodology behind such algorithms is generally never published.

Researchers, analysts, and traders mostly focus on short term prediction of stock prices compared to longer term, i.e., weekly or monthly predictions based on historical data. Some good approaches to long term price prediction already exist such as the ARIMA. Stock markets are generally more predictable in the longer term. Several newer ANN approaches such as the LSTM and RNN are now being explored and compared against existing approaches in predicting long term dependencies in the data and the stock prices, which are equally valuable to the investors and data scientists.

Recently, due to the rising influence of social media on many aspects of our lives, a lot of attention is being given to sentiment analysis based on Twitter or news data. Social media data can be unreliable and difficult to process, and fake news is being posted on the web by multiple sources. A good alternative to these or additional resource would be the quarterly or annual reports filed by the companies (e.g., 10-Q and 10-K) for stock prediction to apply sentiment analysis. These documents, if decoded correctly, give a major insight into a company's status, which can help to understand the future trend of the stock.

6. Conclusions

Financial markets provide a unique platform for trading and investing, where trades can be executed from any device that can connect to the Internet. With the advent of stock markets, people have the opportunity to have multiple avenues to make their investment grow. Not only that, but it also gave rise to different types of funds like mutual funds, hedge funds and index funds for people and institutions to invest money according to their risk appetite. Governments of most countries invest a part of their healthcare, employment, or retirement funds into stock markets to achieve better returns for everyone. Online trading services have already revolutionised the way people buy and sell stocks. The financial markets have evolved rapidly into a strong and interconnected global marketplace. These advancements bring forth new opportunities and the data science techniques offer many advantages, but they also pose a whole set of new challenges. In this paper, we propose a taxonomy of computational approaches to stock market analysis and prediction, present a detailed literature study of the state-of-the-art algorithms and methods that are commonly applied to stock market prediction, and discuss some of the continuing challenges in this area that require more attention and provide opportunities for future development and research. Unlike traditional systems, stock market today are built using a combination of different technologies, such as machine learning, expert systems, and big data which communicate with one another to facilitate more informed decisions. At the same time, global user connectivity on the internet has rendered the stock market susceptible to customer sentiments, less stable due to developing news, and prone to malicious attacks. This is where further research can play an important role in paving the way how stock markets will be analysed and made more robust in the future. A promising research direction is to explore various algorithms to evaluate whether they are powerful enough to predict for the longer term, because markets act like weighing machines in the long run having less noise and more predictability. Hybrid approaches that combine statistical and machine learning techniques will probably prove to be more useful for stock prediction.

Author Contributions: Conceptualisation, D.S. and F.Z.; Methodology, D.S.; Literature survey, D.S., F.Z., and H.I.; Discussion, D.S. and H.I.; Writing—original draft preparation, D.S.; Writing—review and editing, F.Z., and H.I.; Supervision, F.Z.

Funding: This research received no external funding.

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

References

Abu-Mostafa, Yaser S., and Amir F. Atiya. 1996. Introduction to financial forecasting. *Applied Intelligence* 6: 205–13. [CrossRef]

Arévalo, Rubén, Jorge García, Francisco Guijarro, and Alfred Peris. 2017. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications* 81: 177–92. [CrossRef]

Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. 2014. Stock Price Prediction Using the Arima Model. Paper presented at the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, UK, March 26–28.

Babu, M. Suresh, N. Geethanjali, and B. Satyanarayana. 2012. Clustering Approach to Stock Market Prediction. *International Journal of Advanced Networking and Applications* 3: 1281.

Ballings, Michel, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. 2015. Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications* 42: 7046–56. [CrossRef]

Bao, Wei, Jun Yue, and Yulei Rao. 2017. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE* 12: e0180944. [CrossRef] [PubMed]

Bernal, Armando, Sam Fok, and Rohit Pidaparthi. 2012. *Financial Market Time Series Prediction with Recurrent Neural Networks*. State College: Citeseer.

Bhardwaj, Aditya, Yogendra Narayan, and Maitreyee Dutta. 2015. Sentiment analysis for Indian stock market prediction using Sensex and nifty. *Procedia Computer Science* 70: 85–91. [CrossRef]

Bhuriya, Dinesh, Girish Kausha, Ashish Sharma, and Upendra Singh. 2017. Stock Market Prediction Using a Linear Regression. Paper presented at the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, April 20–22; vol. 2.

Billah, Baki, Maxwell L. King, Ralph D. Snyder, and Anne B. Koehler. 2006. Exponential Smoothing Model Selection for Forecasting. *International Journal of Forecasting* 22: 239–47. [CrossRef]

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2: 1–8. [CrossRef]

Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. Hoboken: John Wiley & Sons.

Cakra, Yahya Eru, and Bayu Distiawan Trisedya. 2015. Stock Price Prediction Using Linear Regression Based on Sentiment Analysis. Paper presented at the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, October 10–11.

Cervelló-Royo, Roberto, Francisco Guijarro, and Karolina Michniuk. 2015. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications* 42: 5963–75. [CrossRef]

Chen, Tai-Liang. 2011. Forecasting the Taiwan Stock Market with a Stock Trend Recognition Model Based on the Characteristic Matrix of a Bull Market. *African Journal of Business Management* 5: 9947–60.

Chen, Tai-liang, and Feng-yu Chen. 2016. An intelligent pattern recognition model for supporting investment decisions in stock market. *Information Sciences* 346: 261–74. [CrossRef]

Cheng, Ching-Hsue, Tai-Liang Chen, and Liang-Ying Wei. 2010. A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Information Sciences* 180: 1610–29. [CrossRef]

Chong, Eunsuk, Chulwoo Han, and Frank C. Park. 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications* 83: 187–205. [CrossRef]

Creighton, Jonathan, and Farhana H. Zulkernine. 2017. Towards Building a Hybrid Model for Predicting Stock Indexes. Paper presented at the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, December 11–14.

De Faria, E. L., Marcelo P. Albuquerque, J. L. Gonzalez, J. T. P. Cavalcante, and Marcio P. Albuquerque. 2009. Predicting the Brazilian Stock Market through Neural Networks and Adaptive Exponential Smoothing Methods. *Expert Systems with Applications* 36: 12506–9. [CrossRef]

Devi, B. Uma, D. Sundar, and P. Alli. 2013. An Effective Time Series Analysis for Stock Trend Prediction Using Arima Model for Nifty Midcap-50. *International Journal of Data Mining & Knowledge Management Process* 3: 65.

Dey, Shubharthi, Yash Kumar, Snehanshu Saha, and Suryoday Basak. 2016. Forecasting to Classification: Predicting the Direction of Stock Market Price Using Xtreme Gradient Boosting. Working Paper. [CrossRef]

Di Persio, Luca, and Oleksandr Honchar. 2017. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets. *International Journal of Mathematics and Computers in simulation* 11: 7–13.

Diamond, Peter A. 2000. What Stock Market Returns to Expect for the Future. *Social Security Bulletin* 63: 38.

Ding, Xiao, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-Driven Stock Prediction. Paper presented at the 24th International Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina, July 25–31.

Dutta, Avijan, Gautam Bandopadhyay, and Suchismita Sengupta. 2012. Prediction of Stock Performance in Indian Stock Market Using Logistic Regression. *International Journal of Business and Information* 7: 105–36.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton: CRC Press.

Fama, Eugene F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance* 25: 383–417. [CrossRef]

Fama, Eugene F. 1995. Random walks in stock market prices. *Financial Analysts Journal* 51: 75–80. [CrossRef]

Fu, King Sun, and Tzay Y. Young. 1986. *Handbook of Pattern Recognition and Image Processing*. Cambridge: Academic Press.

Fu, Tak-chung, Fu-lai Chung, Robert Luk, and Chak-man Ng. 2005. Preventing Meaningless Stock Time Series Pattern Discovery by Changing Perceptually Important Point Detection. Paper presented at the International Conference on Fuzzy Systems and Knowledge Discovery, Changsha, China, August 27–29.

Gordon, Myron J. 1959. Dividends, Earnings, and Stock Prices. *The Review of Economics and Statistics* 41: 99–105. [CrossRef]

Gordon, Myron J., and Eli Shapiro. 1956. Capital Equipment Analysis: The Required Rate of Profit. *Management Science* 3: 102–10. [CrossRef]

Hiransha, M., E. A. Gopalakrishnan, Vijay Krishna Menon, and Soman Kp. 2018. NSE stock market prediction using deep-learning models. *Procedia Computer Science* 132: 1351–62.

Hossain, Mohammad Asiful, Rezaul Karim, Ruppa K. Thulasiram, Neil D. B. Bruce, and Yang Wang. 2018. Hybrid Deep Learning Model for Stock Price Prediction. Paper presented at the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, November 18–21.

Hu, Yong, Kang Liu, Xiangzhou Zhang, Lijun Su, E. W. T. Ngai, and Mei Liu. 2015. Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing* 36: 534–51. [CrossRef]

Imam, Shahed, Richard Barker, and Colin Clubb. 2008. The Use of Valuation Models by Uk Investment Analysts. *European Accounting Review* 17: 503–35. [CrossRef]

Kalyanaraman, Vaanchitha, Sarah Kazi, Rohan Tondulkar, and Sangeeta Oswal. 2014. Sentiment Analysis on News Articles for Stocks. Paper presented at the 2014 8th Asia Modelling Symposium (AMS), Taipei, Taiwan, September 23–25.

Kim, Sang, Hee Soo Lee, Hanjun Ko, Seung Hwan Jeong, Hyun Woo Byun, and Kyong Joo Oh. 2018. Pattern Matching Trading System Based on the Dynamic Time Warping Algorithm. *Sustainability* 10: 4641. [CrossRef]

Lee, Heeyoung, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the Importance of Text Analysis for Stock Price Prediction. Paper presented at the 9th International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26–31.

Leigh, William, Naval Modani, Russell Purvis, and Tom Roberts. 2002. Stock market trading rule discovery using technical charting heuristics. *Expert Systems with Applications* 23: 155–59. [CrossRef]

Leigh, William, Cheryl J. Frohlich, Steven Hornik, Russell L. Purvis, and Tom L. Roberts. 2008. Trading with a Stock Chart Heuristic. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38: 93–104. [CrossRef]

Lv, Dongdong, Shuhan Yuan, Meizi Li, and Yang Xiang. 2019. An Empirical Study of Machine Learning Algorithms for Stock Daily Trading Strategy. *Mathematical Problems in Engineering*. [CrossRef]

Markowska-Kaczmar, Urszula, and Maciej Dziedzic. 2008. Discovery of Technical Analysis Patterns. Paper presented at the International Multiconference on Computer Science and Information Technology, 2008, IMCSIT 2008, Wisia, Poland, October 20–22.

Milosevic, Nikola. 2016. Equity Forecast: Predicting Long Term Stock Price Movement Using Machine Learning. *arXiv*.

Mittal, Anshul, and Arpit Goel. 2012. Stock Prediction Using Twitter Sentiment Analysis. *Standford University, CS229*. Available online: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf (accessed on 3 March 2019).

Naseer, Mehwish, and Yasir bin Tariq. 2015. The efficient market hypothesis: A critical review of the literature. *IUP Journal of Financial Risk Management* 12: 48–63.

Nesbitt, Keith V., and Stephen Barrass. 2004. Finding trading patterns in stock market data. *IEEE Computer Graphics and Applications* 24: 45–55. [CrossRef] [PubMed]

Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment Analysis on Social Media for Stock Movement Prediction. *Expert Systems with Applications* 42: 9603–11. [CrossRef]

Pagolu, Venkata Sasank, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. Paper presented at the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, October 3–5.

Park, Cheol-Ho, and Scott H. Irwin. 2007. What do we know about the profitability of technical analysis? *Journal of Economic Surveys* 21: 786–826. [CrossRef]

Parracho, Paulo, Rui Neves, and Nuno Horta. 2010. Trading in Financial Markets Using Pattern Recognition Optimized by Genetic Algorithms. Paper presented at the 12th Annual Conference Companion on Genetic and Evolutionary Computation, Portland, OR, USA, July 7–11.

Patel, Jigar, Sahil Shah, Priyank Thakkar, and K. Kotecha. 2015. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications* 42: 2162–72. [CrossRef]

Peachavanish, Ratchata. 2016. Stock selection and trading based on cluster analysis of trend and momentum indicators. Paper presented at the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, March 16–18.

Phetchanchai, Chawalsak, Ali Selamat, Amjad Rehman, and Tanzila Saba. 2010. Index Financial Time Series Based on Zigzag-Perceptually Important Points. *Journal of Computer Science* 6: 1389–95.

Powell, Nicole, Simon Y. Foo, and Mark Weatherspoon. 2008. Supervised and Unsupervised Methods for Stock Trend Forecasting. Paper presented at the 40th Southeastern Symposium on System Theory (SSST 2008), New Orleans, LA, USA, March 16–18.

Rather, Akhter Mohiuddin, Arun Agarwal, and V. N. Sastry. 2015. Recurrent Neural Network and a Hybrid Model for Prediction of Stock Returns. *Expert Systems with Applications* 42: 3234–41. [CrossRef]

Roondiwala, Murtaza, Harshal Patel, and Shraddha Varma. 2017. Predicting Stock Prices Using Lstm. *International Journal of Science and Research (IJSR)* 6: 1754–56.

Schumaker, Robert P., and Hsinchun Chen. 2009. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The Azfin Text System. *ACM Transactions on Information Systems (TOIS)* 27: 12. [CrossRef]

Seng, Jia-Lang, and Hsiao-Fang Yang. 2017. The association between stock price volatility and financial news—A sentiment analysis approach. *Kybernetes* 46: 1341–65. [CrossRef]

Shah, Dev, Campbell Wesley, and Zulkernine Farhana. 2018. A Comparative Study of LSTM and DNN for Stock Market Forecasting. Paper presented at the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, December 10–13.

Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. 2012. *Stock Market Forecasting Using Machine Learning Algorithms*. Stanford: Department of Electrical Engineering, Stanford University, pp. 1–5.

Shiller, Robert J. 1980. *Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?* Cambridge: National Bureau of Economic Research.

Shiller, Robert C. 2000. Irrational Exuberance. *Philosophy & Public Policy Quarterly* 20: 18–23.

Tiwari, Shweta, Rekha Pandit, and Vineet Richhariya. 2010. Predicting Future Trends in Stock Market by Decision Tree Rough-Set Based Hybrid System with Hhmm. *International Journal of Electronics and Computer Science Engineering* 1: 1578–87.

Velay, Marc, and Fabrice Daniel. 2018. Stock Chart Pattern recognition with Deep Learning. *arXiv*.

Wang, Jar-Long, and Shu-Hui Chan. 2007. Stock Market Trading Rule Discovery Using Pattern Recognition and Technical Analysis. *Expert Systems with Applications* 33: 304–15. [CrossRef]

Wang, Ju-Jie, Jian-Zhou Wang, Zhe-George Zhang, and Shu-Po Guo. 2012. Stock Index Forecasting Based on a Hybrid Model. *Omega* 40: 758–66. [CrossRef]

Wu, Kuo-Ping, Yung-Piao Wu, and Hahn-Ming Lee. 2014. Stock Trend Prediction by Using K-Means and Aprioriall Algorithm for Sequential Chart Pattern Mining. *Journal of Information Science and Engineering* 30: 669–86.

Xu, Yumo, and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. Paper Presented at the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 15–20.

Yang, Bing, Zi-Jia Gong, and Wenqi Yang. 2017. Stock Market Index Prediction Using Deep Neural Network Ensemble. Paper Presented at the 2017 36th Chinese Control Conference (CCC), Dalian, China, July 26–28.

Yoshihara, Akira, Kazuki Fujikawa, Kazuhiro Seki, and Kuniaki Uehara. 2014. Predicting Stock Market Trends by Recurrent Deep Neural Networks. Paper presented at the Pacific RIM International Conference on Artificial Intelligence, Gold Coast, Australia, December 1–5.

Zhang, Jing, Shicheng Cui, Yan Xu, Qianmu Li, and Tao Li. 2018. A novel data-driven stock price trend prediction system. *Expert Systems with Applications* 97: 60–69. [CrossRef]

Zhong, Xiao, and David Enke. 2017. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications* 67: 126–39. [CrossRef]