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.

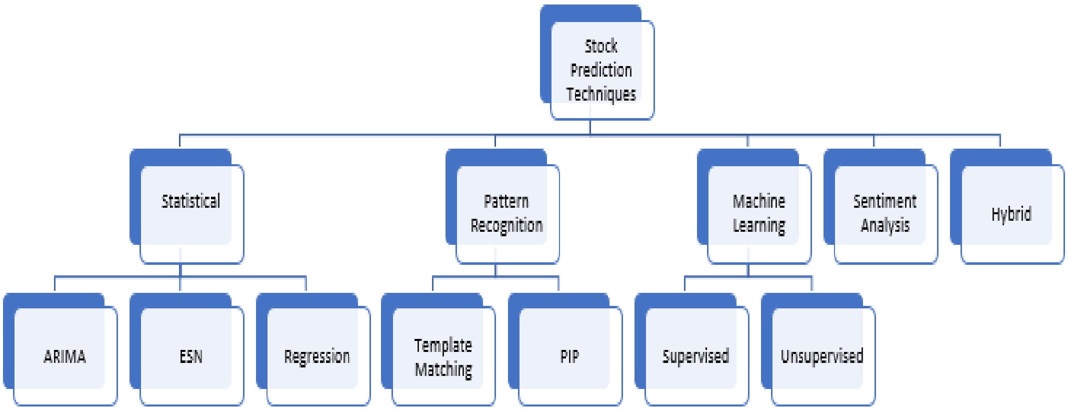

Figure 1. Taxonomy of stock prediction techniques.

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.