

Intelligent Models for Stock Price Prediction: A Comprehensive Review

Kwabena Ansah, University of Ghana, Ghana

Ismail Wafaa Denwar, University of Ghana, Ghana

 <https://orcid.org/0000-0002-5777-6163>

Justice Kwame Appati, University of Ghana, Ghana*

 <https://orcid.org/0000-0003-2798-4524>

ABSTRACT

Prediction of the stock price is a crucial task as predicting it may lead to profits. Stock price prediction is a challenge owing to non-stationary and chaotic data. Thus, the projection becomes challenging among the investors and shareholders to invest the money to make profits. This paper is a review of stock price prediction, focusing on metrics, models, and datasets. It presents a detailed review of 30 research papers suggesting the methodologies, such as support vector machine, random forest, linear regression, recursive neural network, and long short-term movement based on the stock price prediction. Aside from predictions, the limitations and future works are discussed in the papers reviewed. The commonly used technique for achieving effective stock price prediction are the RF, LSTM, and SVM techniques. Despite the research efforts, the current stock price prediction technique has many limits. From this survey, it is observed that the stock market prediction is a complicated task, and other factors should be considered to accurately and efficiently predict the future.

KEYWORDS

Gated Recurrent Unit, Intelligent Models, Long Short Term Memory, Mean Square Error, Predictions, Stock Price

INTRODUCTION

The stock market is a platform or a mutual organization that provides a trader to buy or sell stock shares. They form one of the critical parts of a country's economy as it is an essential way for companies to raise capital (Billah, Waheed, & Hanifa, 2017). Businesses and corporations allowed to offer shares to the public are termed public listed companies, and they have a significant impact on the economies in which they operate (Pun & Shahi, 2018). In most modern economies, other business organizations heavily rely on the funds generated by these financial markets. Therefore, analyzing the behaviour and performance of these financial markets has become a crucial research field. These analyses may include but are not limited to predicting prices of securities such as stocks, bonds, foreign exchange rates, market indicators, and trading volumes (Samarawickrama & Fernando, 2018). The stock market attracts investors and investment institutions' attention due to its high returns (Yao, Luo, & Peng, 2018). Most investors' goal is to predict the stock market's associated risk to decide between buying

DOI: 10.4018/JITR.298616

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

or selling shares of stocks while seeking to maximize profit on investment. However, predicting the behaviour of stocks is difficult because the market is highly volatile and influenced by unmeasurable external factors, including the global economy, events, politics, and investor expectations (Oncharoen & Vateekul, 2018). Stock markets are considered the heart of the world's economy, in which billions of dollars are traded every day. The correct prediction of the future behaviour of markets would be extremely valuable in various areas (Hoseinzade & Haratizadeh, 2019). Traditionally, several conventional methods based on time series have been proposed to aid in predicting the stock market. Classical models like the Black-Scholes has also been used to model the stock market in predicting its volatility. Despite the many works done, accuracy still remains a challenge in this domain. Presently, these markets are known to have generated enormous data, which is of interest to the data science community. With the deep drive of intelligence, machine learning has played a useful role in the prediction of stock price leading to the proposal of several efficient algorithms as discussed under the section of this study named "Papers Reviewed". These learning algorithms learn from historical price data to predict future prices (Nelson, Pereira, & Oliveira, 2017). However, this historical data are expected to be clean as much as possible as a bit of tweak in the data can perpetuate massive differences in the outcome (Parmar et al., 2018). Taking into consideration the intricate nature of this domain and the diverse contribution made by the research community without it being properly synchronized, this study seeks to present a systematic review of what has happened in the past as a contribution to knowledge continuity. The study is organized as follow: we have the background that gives, in brief, an overview of machine learning, followed by research methods that explain the protocols adopted to carry out this study. Next, is papers reviewed where some selected papers are discussed. The discussed paper are analyzed in the result and discussion section, and finally is the conclusion and future works.

BACKGROUND

Machine Learning techniques have become prevalent today in the stock analysis due to their inherent capacity drawn from the enormous amount of data revealing stock price patterns (Kumar, Dogra, Utreja, & Yadav, 2018). According to Khadka (2019), machine learning can be classified into Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Under supervised learning, the machine is provided with labelled data, and a learning algorithm is allowed to generate a mapping function that can identify the expected output for a given unseen input (Sodhi et al., 2019; Linthicum, Schafer, & Ribeiro, 2019; Hao & Ho, 2019). On the contrary, with unsupervised learning (UL), the machine is provided with an unlabeled input dataset, and a learning algorithm generates a function to identify hidden structures in the given dataset patterns, similarities, and differences that exist among data without any prior knowledge (Sodhi et al., 2019; Deepika, Senthil, Rajan, & Surendar, 2017). The third category is reinforcement learning, where the objective is to develop systems that improve performance based on the environment's feedback. Steps that move an algorithm closer to its target are selected and propagated forward to the next iteration. The algorithm is reinforced and improves through repeated iterations until it reaches an optimal performance level or a stopping point of several iterations outlined in parameters (Linthicum et al., 2019). It is exposed to an environment where it takes decisions on a trial and error basis and learns from its actions and past experiences. For every correct decision, the machine receives reward feedback from the environment that acts as a reinforcement signal, and the information about the rewarded state-action pair is stored. A few of the most commonly used reinforcement learning algorithms are Q-Learning and Markov Decision Processes. Nearly all aspects of modern life are changed by machine learning. For instance, Netflix predicts movies users are interested in watching, and Google gives users insights based on their search histories. There is great optimism that these techniques can improve many sectors similarly (Beam & Kohane, 2018). The core utility of machine learning and big data analytics is recognising and extracting meaningful patterns from enormous raw input data. It results in higher levels of insights for decision-making and trend prediction. Therefore, extracting these insights and knowledge from

data is extremely important to many businesses since it enables them to gain competitive advantages (Mohammadi, Al-Fuqaha, Sorour, & Guizani, 2018). According to Sodhi, Awasthi, & Sharma (2019), five types of problems can be solved by machine learning techniques:

1. Classification: used to identify the category to which an object belongs. For example, is it spam? Or is it cancerous?
2. Regression: used to predict a continuous numeric-valued aspect associated with an object. For example, the probability that a user would click on an ad or stock price prediction.
3. Similarity/ Anomaly: used to retrieve similar objects or to find anomalies in behaviour. For example, searching for related images or detecting deception in user behaviour.
4. Ranking: used to sort relevant data according to a particular input. For example, Google Page Rank.
5. Sequence Prediction: used to predict the next element in a series of data. For example, predicting the next word in a sentence. However, similar objects can also be grouped into sets using clustering.

The following gives, in brief, some of the techniques and the rationale underpinning their development. For instance, the standards of Support Vector Machines (SVM) were first proposed in 1995 by Vapnik & Corinna Cortes (Vapnik, 1995) and first implemented by Vladimir N. Vapnik & A. Y. Chervonenkis in 1963. This machine is used for classification, regression, and outlier detection tasks. In a linear SVM approach, each input data is plotted as a point in n -dimensional space where n is input dimensions. Then the classification is performed by obtaining the hyper-plane differentiating the two classes (Rajput & Kaulwar, 2019). In the domain of stock markets, SVMs have been widely applied to build classifiers while embodying the Structural Risk Minimization principle (SRM) (Gandhmal & Kumar, 2019). Researchers used different techniques based on parameter optimization and Ensemble Classifiers (Labiad, Berrado, & Benabbou, 2016). The purpose of the support vector machine is to identify the maximum margin hyperplane. This is achieved by defining the decision boundary, which maximizes the separation between positive and negative examples. SVMs can learn from the high dimensional feature space. Mapping from lower to higher feature dimensional space is achieved using kernel function (Misra & Chaurasia, 2019; Pun & Shahi, 2018; Cakra, 2015).

The Random Forest (RF) technique, on the other hand, is an ensemble learning technique for both classification and regression problems. RF is a set or collection of decision trees that grow in randomly selected subspaces of feature space. It employs the Bagging approach to produce a randomly sampled set of training data for each tree and gives a prediction based on the majority voting (classification) or averaging (regression) (Labiad et al., 2016). Fundamentally, decision trees possess the characteristics of having very low bias and high variance; hence, slight noise in the data could lead the tree to diverge completely. This weakness is avoided in a random forest by training multiple decision trees on a different subspace of the feature space at the cost of a somewhat increased bias. None of the trees in the forest could see the entire training data. The data is recursively split into partitions while a particular node is built according to a specific attribute. The separating criterion choice is based on impurity measures such as Shannon Entropy or Gini impurity (Zhang et al., 2018).

The artificial neural network (ANN) is also a machine-learning algorithm inspired by human brain cells' biological structure. This model's core idea is to optimize given weights in a network structure with a given number of nodes as an input, hidden, and outcome layer. These weights are optimized according to the derivative of an error function through an iterative process. ANN has been used in financial applications such as credit card fraud detection, credit risk assessment, and among other optimization and classification problems. The financial market price forecasting is another field in which ANN has been used robustly (Omar, Daniel, Zineb, & Aida, 2018). The ANN models are implemented using Tensorflow, a popular deep learning framework provided by Google (Zhang et al., 2018) and they are widely used in classification, regression, and clustering tasks (Rasel, Sultana, & Hasan, 2017; Parmar et al., 2018).

Finally, Recurrent Neural Networks (RNN) is a subtype of neural networks that utilizes feedback connections. Several types of RNN models are employed in predicting financial stock prices (Samarawickrama & Fernando, 2018). Among the several types of RNN is the Long Short-Term Memory (LSTM) network, which introduces the concept of a memory cell and gate structure to associate memories and inputs remotely in time effectively. It also can grasp the long-term structure of data dynamically over time (Yao et al., 2018). Another type of RNN is the Gated Recurrent Units (GRU) which does not have an output gate (Samarawickrama & Fernando, 2018). LSTM is used to introduce a new structure called the memory cell controlled by three different gates: forget gate, input gate, and output gate. The forget gate decides which information of the previous cell state is remembered or forgotten. The input gate determines an input signal updates on which values of the cell state. Finally, the output gate allows the cell state to have or not have an effect on other neurons. This structure's importance is to model long-term dependencies in sequence data and prevent the vanishing gradient problem (Vargas, Dos Anjos, Bichara, & Evsukoff, 2018).

RESEARCH METHODS

A review protocol is set up at the planning phase of this systematic literature review (SLR) in this study. The review protocol has six aspects: research questions definition, search strategy design, study selection, quality assessment, data extraction, and data synthesis. In the first phase, a research question is formed based on the objective of this SLR. In the second phase, aiming at the research questions, a search strategy is designed to determine the studies relevant to the research questions; it involves determining search terms and selecting literature resources necessary for the subsequent search process. In the third phase, study selection criteria are defined to identify the relevant studies that address the research questions. In this stage, pilot study selection was employed to refine the selection criteria further. Next, the relevant papers undergo a quality assessment process in which we devised some quality checklists to facilitate the assessment. The remaining two phases involve data extraction and data synthesis, respectively. The data extraction form is initially planned in the data extraction stage and subsequently refined through pilot data extraction. Finally is the data synthesis stage. The subsequent sections present the review protocol's details.

Research Question

This review seeks to summarise and clarify the empirical evidence on stock price prediction models. Towards this aim, five research questions (RQs) were raised as follows in Table 1.

Table 1 Research Questions

Research Question (RQ)	Main Motivation
RQ1: Which year had the most publications on stock price prediction?	Identify the year with the most published papers on stock prediction
RQ2: What kind of datasets are the most used for stock prediction?	Identify whether predictive models are repeatable or not by examining the usage.
RQ3: What kind of methods (supervised and unsupervised learning) are the most used for stock prediction?	Identify trends for the prediction methods focus.
RQ4: What metrics are used the most for stock prediction?	Identify trends for prediction metrics focus.
RQ5: What are the common limitations and future works?	Identify the common limitations and future works on stock price prediction papers.

Search Strategy

The search strategy includes the search terms, literature resources, and search processes, as detailed in the following subsections (Wen, Li, Lin, Hu, & Huang, 2012).

Search Terms

The following steps were used to construct the search terms:

1. Derive key terms from the research questions
2. Identify synonyms and alternative spellings for key terms.
3. Search for keywords in relevant papers or books
4. Use the Boolean AND to link the key terms. The search queries are as follows:

((("All Metadata": "stock price prediction") AND "Full Text & Metadata": "machine learning") AND "Full Text & Metadata": "algorithms?"), content.ftsec:(+ "stock price prediction") AND (+ "machine learning" + "algorithm?") for IEEE and ACM respectively.

Literature Resources

Literature searches (advanced search) were conducted in December 2019 using two scientific databases. The papers included in the study were from the year 2015 to 2019. The initial results were as follows: IEEE Xplore returned 40 results (38 conference papers, 1 journal, 1 Magazine). The database covers big data, data science, computer science, data mining, and information science. ACM returned 11 results.

Search Process

The search process was divided into the following two phases:

1. Search the two electronic databases separately to form a set of candidate papers.
2. Skimmed through the CSV files after exporting from the two databases. 41 relevant papers were identified according to the search process.

Study Selection

The study included papers that describe research on the stock price prediction. The study excluded articles below the year 2015 and articles, which do not include experimental results. Articles concerning their years, datasets, metrics, techniques, evaluation criteria, results, limitations, and future works have been examined. The inclusion of papers was based on the study's similarity degree with the stock prediction research topic.

Study Quality Assessment

For quality, assessment questions are devised to assess the studies' rigorousness, credibility, and relevance. These questions are presented in Table 2. Three optional answers are provided for each question: "Yes", "Partly", or "No". These three answers are scored as follows: "Yes" = 3, "Partly" = 1.5, and "No" = 0. A given study's quality score is computed by summing up the answers' scores to the QA questions. From the assessment, some papers were discarded as those studies did not include relevant information. Out of the 41 papers, 30 papers were deemed relevant for the systematic review.

Data Extraction

The authors, journals, year, title, methods, datasets, findings, future works, and limitations were the parameters for data extraction. For each paper, the parameters were obtained to aid in answering the research questions. Articles that did not satisfy the parameters were deemed as weak papers.

Table 2 Quality Assessment

Number	Question
QA1	Are the aims or problems of the research clearly defined?
QA2	Are the methods employed for prediction conventional?
QA3	How common are datasets used?
QA4	Are conventional metrics for evaluating the models used?
QA5	Are the limitations and future work of the study analyzed explicitly?

Data Synthesis

The extracted data were analyzed and synthesized based on the evidence among the papers reviewed. The synthesis aids in answering the research questions. For example, this synthesis contains pieces of evidence in article A compared to paper B and other analysis papers. The data extracted in this review include quantitative data (for example, values of prediction) and qualitative data (for example, theories, strengths, and weaknesses of the prediction models).

Threats to Validity

The main threats to this review protocol's validity are analyzed from the following three aspects: study selection bias, publication bias, and possible inaccuracy in data extraction.

Study selection bias could be a threat as the search query is used to retrieve data based on the keywords used automatically; the query structure could be a threat as other relevant studies could be missed. A manual search was conducted on the databases with a less complex query since it will load more papers to avoid selection bias.

Publication bias is another threat, as authors want to claim their proposed models are more accurate than conventional methods. Only one research paper agreed that their proposed models were less traditional compared to other models. To avoid these, comparing only conventional models will be the right move. Authors must accurately report their findings based on the hybrid or proposed models. To minimize the threat of inaccuracy in data extraction, the parameters defined were adhered to strictly.

PAPERS REVIEWED

In this section, papers in the review journal articles and conference proceedings are discussed. The total number of papers in the review is 30 with their limitations, and future works where available are discussed as follows: In the study of Soni, Agarwal, Arora, & Gupta (2018), a stock price prediction model was built. The methods employed were Decision Tree, PSO, Black-Hole, Naïve Bayes. The proposed model was a "nature-inspired technique" that used a matrix with values 1 and 0 for prediction. The database for training and testing the models was the Nifty stock index dataset. The metric used for evaluating the models was mean accuracy. The proposed model had a higher mean accuracy of 96.10% than the rest of the models. Decision Tree had a mean accuracy of 81.16%, PSO 83.57%, Naïve Bayes 85.56%, and Black Hole 95.10%. The authors intend to expand the scope by including more models as the current study is limited to Decision Tree, PSO, Black-Hole, Naïve Bayes models. The study

of Nayak, Pai, & Pai (2016) also sought to predict stock market trends daily for the oil, mining, and banking sectors. The methods employed in the study were Support Vector Machine (SVM), Logistic Regression (LR), and Boosted Decision Trees (BTS) models. The database for training and testing the models was Yahoo Finance. It had an open price, close price, low price, high price, adjusted close price, and volume as features. The metric for evaluating the models was accuracy. In their findings, BTS performed better than LR and SVM. BTS had an accuracy *0.548, 0.76, 0.769* for banking, mining and oil respectively; for LR *0.654, 0.61, and 0.442* and, lastly for SVM *0.51, 0.59, and 0.442*. Again the study of Ouahilal, Mohajir, Chahhou, & El Mohajir (2017) optimized the stock price prediction with a novel Hybrid approach. The methods employed were Support Vector Regression (SVR) with the Hodrick-Prescott filter (Optimizing prediction). Christiano Fitzgerald (CF) filter and Band-Pass (BP) filter using the Fourier transformation were also used with SVR to compare results. The database for training and testing their models was Maroc Telecom (IAM) financial time series. The metric for training and testing the models was the Mean Average Percentage Error (MAPE). In their findings, the proposed model gives better results in terms of stock price predictions. SVR had a MAPE of *0.29, SVR+HP 0.11, SVR+CF 0.51, SVR+BP and 0.22*.

In Pun & Shahi (2018), stock prices prediction for the next day using the Nepal Stock Exchange was proposed. The methods employed in this study were Support Vector Regression (SVR) and Artificial Neural Networks (Back-Propagation Neural Network) models. Min-Max and Z-score are used to normalize the data, and the metrics for evaluating the models are mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), and Coefficient of Determination (R²). In their findings, SVR with min-max normalization performs better than BPNN in all sectors, except on Development bank, Finance, and Mutual Fund. Both models found low accuracy in Trade and Factory sectors. BPNN was also found to be better using z-score than using min-max in these sectors. However, SVR was observed more appropriately than BPNN in all sectors on average. The authors believe that the performance can be improved if the dataset size is increased in future research. Rasel, Sultana, & Hasan (2017) also sought to predict stock prices and trends using time series data of 1-day ahead market. The methods employed were Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The database for training and testing the models was the Wal-Mart Stores Inc. (WMT) dataset. The dataset attributes were high, open, low, and volume and the metrics for evaluating the model were the MAPE and RMSE. In their findings, ANN performed better than KNN and SVM. ANN had a MAPE of *0.75* and an RMSE of *0.60*. SVM had a MAPE of *2.75* and an RMSE of *1.90*, while KNN had a MAPE of *2.71* and RMSE of *2.28*. Their future work stated that different models and datasets of stock markets could be used to build a universal model. The study of Weng, Lu, Wang, Megahed, & Martinez (2018) also develop an expert financial system to predict historical data's short-term financial prices. The methods employed were four ensemble models; thus, Support Vector Regression, Boosted Regression Trees, Random Forest Regression, and Neural Network Regression Ensemble. The database for training and testing the models was the Citi Group stock price data set. The metric for evaluating the models was the MAPE, RMSE and MAE. In their findings, the Boosted Regression Tree performed better than Support Vector Regression, Random Forest Regression, and Neural Network Regression of the 19 stocks.

The study of Labiad, Berrado, & Benabbou (2016) sought to predict very short-term (10 minutes ahead) variations of the Moroccan stock market. The methods employed were Random Forest (RF), Gradient Boosted Trees (GBT), and Support Vector Machine (SVM). Intraday prices (tick-by-tick data) of Maroc Telecom (IAM) stocks are employed as an experimental database to evaluate the selected model's performances. The metric used for assessing the models was Mean Absolute Deviation (MAD). For technical analysis, Moving Average (MA), Rate of Change (ROC), Standard Deviation (SD), Psychological Line (PSL), Stochastic Oscillator, Relative Strength Index (RSI), and Observations/price variations (Up, Down) were considered. The findings show that RF and GBT are superior to SVM for their selected dataset. Usmani, Ebrahim, Adil & Raza (2019) also predicted the performance of the Karachi Stock Exchange (KSE) as merged into Pakistan Stock Exchange (PSX)

with a proposed Hybrid model. The methods employed were Support Vector Machine, Radial Basis Function (RBF), and ANN. The ANN was of two variants, including Single Layer Perceptron and Multi-layer Perceptron, to predict the stock price. The database used for training and testing the models was the KSE-100 index of the Pakistan Stock Market. The metrics used for validation were Auto-Regressive Integrated Moving Average (ARIMA) and Simple Moving Average (SMA). There were two variants of the proposed Hybrid model; Variant I gave about 72.8% accuracy while Variant II gave 95.7% accuracy on the training data set. The Hybrid model could not predict better than the results achieved by the MLP based sub-model alone on the test data set. The results suggested that the market's behaviour can be predicted using a more complex model implementing different machine learning techniques. Oncharoen & Vateekul (2018) study improved the stock market predictions of historical price data and technical indicators as input using a deep learning approach for their proposed model. The methods employed were Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) architectures for the proposed model. The databases used for training and testing the models were Intrinsic, Standard & Poor's 500 Index (S&P500), Dow Jones Industrial Average (DJIA), Reddit, and Reuters. The data were in numerical and textual format, and the metrics for evaluating the models was accuracy. In their findings, the proposed model gives a better prediction accuracy than the baseline models. Considering both numerical and textual information as inputs can improve prediction performance in a deep neural network. The authors anticipate introducing Recurrent Neural Networks and attention mechanisms into textual input may improve prediction accuracy.

Jiao & Jakubowicz (2018), in their study, sought to evaluate four classification algorithms' performance for stock movement direction. Random Forest, Gradient Boosted Trees, Artificial Neural Network, and Logistic regression were employed, while the S&P 500 index was the dataset for testing and training the model. The metrics for validating the models were standard cross-validation, sequential validation, and single validation. In their findings, it was challenging to predict stocks from the past. Recent information such as recently closed European and Asian indexes to predict S&P 500 can lead to a vast increase in predictability. Moreover, among various sectors, financial sector stocks are comparatively easier to predict than other sectors. Their study intends to use microeconomic (Interest rates, industrial production) data for further studies. The study of Ballings, Van Den Poel, Hespeels, & Gryp (2015) predicted stock prices by benchmarking ensemble methods against single classifier models. The ensemble methods were Random Forest, Adaboost, and Kernel Factory, while classifier models were Neural Networks, Logistic Regression, Support Vector Machines, and K-Nearest Neighbor. The database used for training and testing the models was the Amadeus Database, and the metric for validating the models was the area under the curve (AUC) and Cross-Validation. Findings indicate that Random Forest is the most favoured algorithm, followed by Support Vector Machines, Kernel Factory, Adaboost, Neural Networks, K-Nearest Neighbors, and Logistic Regression. Misra & Chaurasia (2019) sought to predict movement direction for the next day's high price for the S&P BSE Sensex index. Random Forest, Support Vector Machine, and Artificial Neural Network methods were adopted, with the S&P BSE Sensex index being the dataset. The metrics for evaluating the models were Precision, Recall, F-Score, and Accuracy. The technical indicators used for predicting movements were the Relative Strength Index (RSI), Accumulation/Distribution (AD), William%R, Stochastic%K, Momentum, and Commodity Channel Index (CCI). In their findings, the RF provided the best accuracy, which SVM and ANN follow. It further revealed that the combined models significantly improve over the single-pass model, supporting the assumption that conversion from continuous to discrete form indicators filters more noise. The study of Jeevan, Naresh, & Vijaya (2018) also predicted share price using various factors such as current market price, price-earnings ratio, base value, and some miscellaneous events. It is observed that changing stocks' market prices may not follow the same cycle based on various factors within the company. In their study, Long Short Term Memory (LSTM) and Recurrent Neural Networks (RNN) was employed, and the database for the training and testing was NSE data. The metric for the model evaluation was the sliding window. From their experimental results, RNN based architecture proved very efficient in predicting the stock

price by changing the configuration accordingly, using a backpropagation mechanism while gathering and grouping data to avoid overlapping data.

Vargas, Dos Anjos, Bichara, & Evsukoff (2018) also made a stock price prediction using technical indicators and financial news with a proposed model (SI-RCNN). This remains a challenging task because market behaviour is stochastic, volatile, and influenced by many factors, such as the global economy, politics, and investor expectations. The methods employed were Convolutional Neural Network (CNN) and Long Short-Term Memory, and the databases for training and testing the models were Reuters for financial news and CVX stock price from Yahoo Finance. Two sets of metrics were used, Set1: Stochastic %K, Stochastic %D, Momentum, Rate of Change, William's %R, Moving Average Convergence-Divergence, Relative Strength Index, Accumulation/Distribution (A/D) oscillator, and Disparity 5; Set 2: Exponential Moving Average, On Balance Volume and Bollinger Bands. In their findings, SI-RCNN architecture could make a reasonable profit (13.94% in 8 months) compared with a buy-and-hold strategy, which was 3.22% over the period. Besides, it showed that news titles and technical indicators as inputs gave a better forecast than using a single input such as LSTM for technical indicators alone. Though the study did not explicitly include future work, the authors recommend that some trading strategies be included, such as stop gain and stop loss and eliminate small variations to make the model focus only on events with a significant variation on prices. Zhang et al. (2018), in their work, also predicted the price trend for 30 days, given that the financial market is risky, chaotic, complex, dynamic, and full of uncertainties. Many factors, such as economic policy, breaking news, political events, and investors' sentiments, may cause asset market fluctuations. In their study, SVM, Neural Network, Naive Bayesian Classifier, and Random forest were used, and the database for training and testing the models was the Shanghai Stock Exchange (SSE) 50 index. The metric for evaluating the models was accuracy, and the technical indicators used for the prediction were Simple Moving Average(SMA), Exponential Moving Average(EMA), Average True Range(ATR), Average Directional Movement Index (ADMI), Commodity Channel Index(CCI), Price rate of change(ROC), Williams %R, Stochastic %K, Stochastic %D and Relative Strength Index (RSI). Their findings demonstrate that ANN performs better than the other three models and promising to find good patterns. Finally, the study of Ta, Liu, & Addis (2018) aimed at determining how machine-learning techniques could meet quantitative trading standards for stock movement prediction. In their study, Linear Regression and Support Vector Regression was utilized along with S&P 500 ETF-SPY of 10 years daily historical data using the Quandl API in testing their model. The metrics for evaluation were Mean square error, accuracy, and the Error rate was the metric. In their findings, the linear regression model performs better than support vector regression in the short-term prediction. However, the Support Vector Regression model tends to work better than linear regression in long-term prediction.

Summary Activities of Papers Reviewed

From Table 3, we present summaries of some papers reviewed in this study. One key observation made from the table is that machine learning knowledge has fairly dominated the stock market domain in the past decade. According to the table, all articles are trying to address the same problem that classical methods were trying to resolve: the improvement of accuracy in stock price prediction. Even though several works state their proposed methods to be more efficient over other benchmarked algorithms, it is not clear what defines best based on the accuracy metric as one will expect a rate of 80% or more, which is not the case except for Li, Bu, & Wu (2017). It is also clear from the table that some form of preprocessing techniques are leveraged, distinguishing one article from the other comparatively using the same dataset and metric.

RESULTS AND DISCUSSION

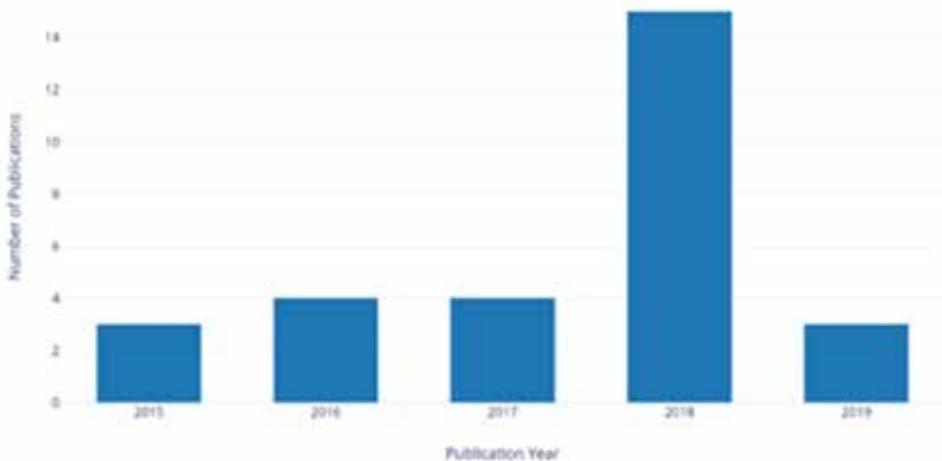
Table 3 Summary of Activities

Articles	ML Methods Used	Dataset	Metric	Performance
Tang & Chen (2018)	RNN, LSTM, CNN, FFNN, RNN + LSTM	Yahoo Finance Reddit World News Channel	Accuracy	LSTM 52.64% FFNN 50.33% CNN 51.38% RNN+LSTM 54.45%
Li, Bu, & Wu (2017)	LSTM, Naïve Bayes sentiment classifier	CSI300	Accuracy	87.86%
Cakra (2015)	Naïve Bayes (NB), Random Forest (RF)	Yahoo Finance CSV API (Indonesian companies) Twitter feeds	Accuracy	RF 60.39% NB 56.59%
Yao et al. (2018)	LSTM, Random Forest	CSI 300	Precision (AccRF), Recall (RecRF), Critical Error (CerRF)	LSTM 28.54%, 38.15%, 16.94% RF 20.96%, 20.42%, 23.67%.
Du, Liu, Chen, & Wang (2019)	LSTM	Apple Stocks	MSE, MAE	Multivariate 0.024 0.033 Univariate 0.035 0.155
Samarawickrama & Fernando (2018)	Feedforward Multi-Layer Perceptron (MLP), Simple Recurrent Neural Network (SRNN), Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM)	Colombo Stock Exchange	MAD, MAPE	MLP 1.7636 0.86% LSTM 2.2701 1.13% SRNN 0.72 – 6.57 0.67% - 5.60%
Hossain, Karim, Thulasiram, Bruce, & Wang (2019)	LSTM+GRU	S&P 500	MAE, MSE, MAPE	0.023 0.00098 4.13
Selvin, Vinayakumar, Gopalakrishnan, Menon, & Soman (2017)	LSTM; RNN, Concurrent Neural (CNN),	Infosys, TCS, and Cipla	Error Percentage (EP)	Infosys CNN 2.36 LSTM 4.18 RNN 3.90 Cipla CNN 3.63 LSTM 3.94 RNN 3.83 TCS CNN 8.96 LSTM 7.82 RNN 7.65

Count Analysis Based on Publication of Years (RQ1)

This subsection presents the analysis based on the publication years of the considered stock prediction techniques. Figure 1 illustrates the number of research papers published in the years from 2015 to 2019. From the 30 articles surveyed, the number of works that are 15 research papers was published in the year 2018. In 2015 and 2019, three research works were developed for stock price prediction. In 2016 and 2017, four research papers were published.

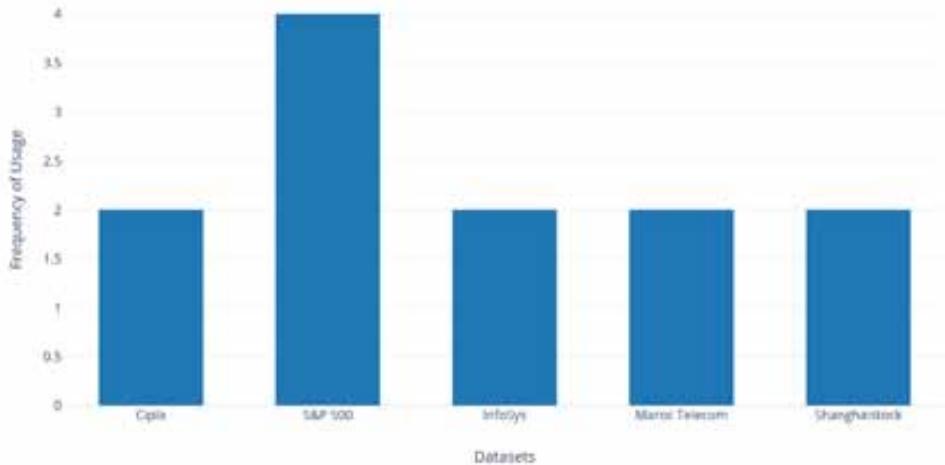
Figure 1. Analysis based on publication years



Count Analysis Based on Datasets (RQ2)

This section elaborates on the analysis carried out based on the datasets adopted in the research works. Various datasets employed for the effective stock market prediction are depicted in Figure 2. The frequently used datasets for the stock market prediction are Cipla, S & P 500, Infosys, Maroc Telecom, Shanghai stocks; these were used more than once in the papers. Other datasets considered are the Nifty stock index dataset, Larsen & Toubro (LT) and State Bank of India (SBIN) intraday price movement, Nepal Stock Exchange (NEPSE), Wal-Mart Stores Inc. (WMT) dataset, TCS, Citi Group stock, Yahoo Finance (specific dataset not disclosed), Apple Stock, Colombo Stock, CSI 300 constituent stocks, Amazon, Bosch, Bata, Eicher, Maroc Telecom (IAM) stocks, Pakistan Stock, Intrinio, DJIA, CSI300, Amadeus Database, S&P BSE Sensex index, Yahoo Finance CSV API, National Stock Exchange of India, CVX stock price, Brazilian, and Chilean currency exchange, Exchange (SSE) 50 index, DJIA stock.

Figure 2 Analysis based on datasets employed



Count Analysis Based on Prediction Methods(RQ3)

In this subsection, the analysis is carried based on the applied stock price prediction techniques. The techniques used for effective stock price prediction is depicted in Figure 3. Figure 3 shows that out of the 30 articles, 11 of the works employed the Random Forest (RF), 10 of the research papers used SVM, and 10 of the works are based on the LSTM. The RNN is employed in 3 out of the 30 papers. Artificial Neural Networks (ANN) are employed in 10 of the papers. Naive Bayes is used in 5, and 5 of the papers are based on Linear Regression. Five articles are based on K- Nearest Neighbor (KNN), and the remaining four are based on CNN. Thus, RF, SVM, and LSTM are the most employed techniques for stock price prediction. This analysis did not consider ensembles; instead, the number of times a method (separate) is used in a paper. It implies that 26% of the works used a supervised learning approach, 37% of the papers used neural networks, and 37% of the works used supervised learning and unsupervised learning.

Count Analysis Based on Performance Metrics (RQ4)

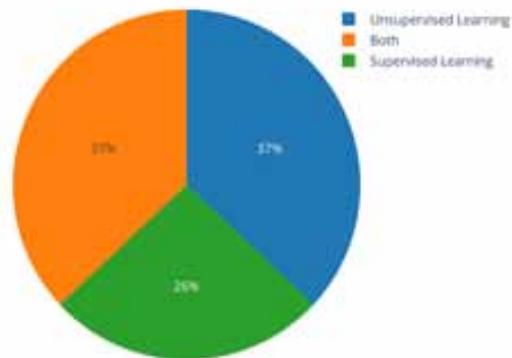
The commonly used performance metrics are MAPE, RMSE, accuracy, MSE, and MAE. Other metrics include Precision, F-measure, F1-Score, which are employed in research papers. Accuracy appeared in 9 out of the 30 articles, MSE appeared in 7, MAPE appeared in 5, RMSE, MAE, Recall, Precision appeared in 4 papers, F-measure appeared in 3, and other metrics appeared only once.

Count Analysis of Common Limitation and Future Works (RQ5)

Stock price prediction is a challenge owing to non-stationary and chaotic data. The limiting factors such as market sentiments, government policy decisions make stock price prediction a challenge.

Authors in various studies have suggested increasing the number of prediction models to use or improve reuse models for future works. Others intend to increase the size of the dataset.

Figure 3 Analysis based on prediction techniques



IMPLICATIONS FOR RESEARCH AND PRACTICE

The methods and the dataset for training and testing determine the accuracy of the stock prediction. Methods employed in the various studies have proven more accurate to other methods; the more data, the more precise the prediction using metrics to evaluate. This review has found that ensembles are rarely used, only 2 papers out of the 30 used ensembles. Ensembles improve machine learning results by combining several models. Using ensemble methods allows producing better predictions compared to a single model. Ensemble learning can help researchers handle both bias and variance — variance representing scattered results that are difficult to converge, and bias represents the error in targeting good results. Researchers are encouraged to explore the possibilities of using unconventional machine learning techniques to predict and monitor the conventional approach's performance. Researchers need to understand the metrics used for evaluating the models. Does a high value mean good, or does a low value mean bad? Metrics readings vary, and to avoid wrong interpretations, researchers must understand how they work. This study shows that neural networks (NN) have been employed in several studies; 11 out of the 30 papers used NN. The study of Gandhmal & Kumar (2019) stated that ANN was declared not a practical scheme for predicting the stock market as the neural models cannot tolerate high computational overhead due to large neurons in the hidden layer and appropriate weight adaption. NN, developed in, performed both the testing and the training slower; this affected the prediction performance. Moreover, overfitting, trapped in local minima, and black box technique are the drawbacks that can be handled using NN. The obtained results of NN based stock market prediction system devised were with low accuracy due to the influence of the misclassification of similar patterns, and the network parameters utilized were not optimized. The CNN-based stock prediction method's research issue is that the devised CNN with the deep learning framework was unsuitable for pervasive applications. CNN's recognition accuracy rate was comparatively weaker than the other state of the art prediction system for stock prediction. The devised decision support system did not use the practical knowledge and techniques to design a workable stock expert system in stock investment. ANN required a prolonged training process for developing an optimal model and suffered from a lack of explanation for determining the solution is generated. The NN based prediction system depends on the correlation value of the chosen feature.

Samarawickrama & Fernando (2018) have stated in their paper that the LSTM architecture has 3 gates, namely the input gate, output gate, and forget gate. LSTM is a solution to the vanishing gradient problem the simple recurrent network cannot solve. The study claimed the LSTM could preserve the error that can be propagated through time and layers. Similarly, Parmar et al. (2018) state that the stock market involves processing massive data, the gradients concerning the weight matrix may become very small and may degrade the learning rate, which corresponds to the Vanishing Gradient problem. LSTM prevents this from happening. The LSTM consists of a remembering cell, input gate, output gate, and forget gate. The cell remembers the value for long term propagation, and the gates regulate them. Hossain et al. (2018) reiterate that LSTM has a memory unit (remembering cell) to track the specific amount of training data. The GRU recurrent neural network is an LSTM, but what makes it different is the absence of the output gate; it is a gating mechanism in the neural network (Samarawickrama & Fernando, 2018). The difference between LSTM and GRU is, GRU combines the forget and the input gates into a single update gate, and it merges the cell state and the hidden state; GRU model is a simpler yet faster network than the standard LSTM models, although the primary purpose of using GRU is similar as LSTM (Hossain et al., 2018). As shown in this review, different machine learning techniques favour different projection contexts. Therefore, before deciding on the choice of machine learning models, researchers need to be aware of the contexts and understand the dynamics of the candidate machine learning models. The context and the methods employed have a direct and significant impact on the performance of the model.

CONCLUSION, LIMITATION AND FUTURE WORK

This study reviewed papers on stock price prediction to evaluate the progress and future research on stock prediction. The papers were evaluated with a specific focus on metrics, methods, and datasets and did not elaborate on the prediction models in detail. The aim was to classify studies concerning the metrics, methods, and datasets that have been used in stock prediction papers. The year with the highest number of publications was 2018. The techniques employed for the stock price prediction involves Support Vector Machine (SVM), Random Forest (RF), Linear Regression, Recursive Neural Network (RNN), Long Short-Term Movement (LSTM). Also, the issues for predicting the stock market are elaborated for suggesting useful future scope. The commonly used technique for achieving effective stock price prediction is RF, SVM, LSTM, and PSO (optimization). The study suggests the following changes in stock prediction research:

1. Increase the models based on machine learning techniques. As specified in this review, machine learning models have better features than statistical methods. Therefore, it would be useful to increase the number of models for machine learning.
2. Increase the usage of well-known datasets for stock prediction problems; this will enhance the prediction as datasets are used to train and test the models, making the patterns clear. The models improve over time.

This study considered papers from 2015 to 2019, and older papers were not considered. This study does not attempt to predict stock prices but compares findings (predictions) from other studies.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

- Ballings, M., Van Den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20), 7046–7056. doi:10.1016/j.eswa.2015.05.013
- Billah, M., Waheed, S., & Hanifa, A. (2017). Stock market prediction using an improved training algorithm of neural network. *ICECTE 2016 - 2nd International Conference on Electrical, Computer and Telecommunication Engineering*, (December), 8–10. doi:10.1109/ICECTE.2016.7879611
- Cakra, T. (2015). Stock Price Prediction using Linear Regression based on Sentiment Analysis. *International Journal of Scientific and Engineering Research*, 6(3), 1655–1659. <https://www.ijser.org/researchpaper/Stock-Price-Prediction-Using-Regression-Analysis.pdf>
- Du, J., Liu, Q., Chen, K., & Wang, J. (2019). Forecasting stock prices in two ways based on the LSTM neural network. *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 1083–1086. doi:10.1109/ITNEC.2019.8729026
- Hoseinzade, E., & Haratizadeh, S. (2019). PT. *Expert Systems with Applications*. Advance online publication. doi:10.1016/j.eswa.2019.03.029
- Hossain, M. A., Karim, R., Thulasiram, R., Bruce, N. D. B., & Wang, Y. (2019). Hybrid Deep Learning Model for Stock Price Prediction. *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, 1837–1844. doi:10.1109/SSCI.2018.8628641
- Jeevan, B., Naresh, E., & Vijaya, B. P. (2018). Share Price Prediction using Machine Learning Technique. *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, (1), 1–4.
- Jiao, Y., & Jakubowicz, J. (2018). Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 4705–4713. doi:10.1109/BigData.2017.8258518
- Kumar, I., Dogra, K., Utreja, C., & Yadav, P. (2018). A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 1003–1007. doi:10.1109/ICICCT.2018.8473214
- Labiad, B., Berrado, A., & Benabbou, L. (2016). Machine learning techniques for short term stock movements classification for Moroccan stock exchange. *SITA 2016 - 11th International Conference on Intelligent Systems: Theories and Applications*. doi:10.1109/SITA.2016.7772259
- Li, J., Bu, H., & Wu, J. (2017). Sentiment-aware stock market prediction: A deep learning method. *14th International Conference on Services Systems and Services Management, ICSSSM 2017 - Proceedings*. doi:10.1109/ICSSSM.2017.7996306
- Misra, P., & Chaurasia, S. (2019). Forecasting Direction of Stock Index Using Two Stage Hybridization of Machine Learning Models. *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 533–537. doi:10.1109/ICRITO.2018.8748530
- Nayak, A., Pai, M. M. M., & Pai, R. M. (2016). Prediction Models for Indian Stock Market. *Procedia Computer Science*, 89, 441–449. doi:10.1016/j.procs.2016.06.096
- Nelson, D. M. Q., Pereira, A. C. M., & Oliveira, R. A. De. (2017). *Stock Market's Price Movement Prediction With LSTM Neural Networks*. Academic Press.
- Oncharoen, P., & Vateekul, P. (2018). Deep Learning for Stock Market Prediction Using Event Embedding and Technical Indicators. *ICAICTA 2018 - 5th International Conference on Advanced Informatics: Concepts Theory and Applications*, 19–24. doi:10.1109/ICAICTA.2018.8541310
- Ouahilal, M., El Mohajir, M., Chahhou, M., & El Mohajir, B. E. (2017). Optimizing stock market price prediction using a hybrid approach based on HP filter and support vector regression. *Colloquium in Information Science and Technology, CIST, 0*, 290–294. doi:10.1109/CIST.2016.7805059
- Pun, T. B., & Shahi, T. B. (2018). Nepal Stock Exchange Prediction Using Support Vector Regression and Neural Networks. *Proceedings of 2018 2nd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2018*, 1–6. doi:10.1109/ICAIECC.2018.8479456

- Qian, Q., & Xiaoxia, W. (2019). Stock price reversal point prediction based on ICA and SVM. *ACM International Conference Proceeding Series*, (5), 101–104. doi:10.1145/3325730.3325760
- Rasel, R. I., Sultana, N., & Hasan, N. (2017). Financial Instability Analysis using ANN and Feature Selection Technique: Application to Stock Market Price Prediction. *2016 International Conference on Innovations in Science, Engineering and Technology, ICISSET 2016*. doi:10.1109/ICISSET.2016.7856515
- Samarawickrama, A. J. P., & Fernando, T. G. I. (2018). A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market. *2017 IEEE International Conference on Industrial and Information Systems, ICIIS 2017 - Proceedings*, 1–6. doi:10.1109/ICIINFS.2017.8300345
- Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, 1643–1647. doi:10.1109/ICACCI.2017.8126078
- Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to Machine Learning and Its Basic Application in Python. *SSRN Electronic Journal*, 1354–1375. 10.2139/ssrn.3323796
- Soni, D., Agarwal, S., Agarwal, T., Arora, P., & Gupta, K. (2018). Optimized Prediction Model for Stock Market Trend Analysis. *2018 11th International Conference on Contemporary Computing, IC3 2018*, 2–4. doi:10.1109/IC3.2018.8530457
- Ta, V. D., Liu, C. M., & Addis, D. (2018). Prediction and portfolio optimization in quantitative trading using machine learning techniques. *ACM International Conference Proceeding Series*, 98–105. doi:10.1145/3287921.3287963
- Tang, J., & Chen, X. (2018). Stock market prediction based on historic prices and news titles. *ACM International Conference Proceeding Series*, 29–34. doi:10.1145/3231884.3231887
- Usmani, M., Ebrahim, M., Adil, S. H., & Raza, K. (2019). Predicting Market Performance with Hybrid Model. *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology, ICEEST 2018*, 1–4. doi:10.1109/ICEEST.2018.8643327
- Vargas, M. R., Dos Anjos, C. E. M., Bichara, G. L. G., & Evsukoff, A. G. (2018). Deep Learning for Stock Market Prediction Using Technical Indicators and Financial News Articles. *Proceedings of the International Joint Conference on Neural Networks*, 1–8. doi:10.1109/IJCNN.2018.8489208
- Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, 54(1), 41–59. doi:10.1016/j.infsof.2011.09.002
- Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, 112, 258–273. doi:10.1016/j.eswa.2018.06.016
- Yao, S., Luo, L., & Peng, H. (2018). High-frequency stock trend forecast using LSTM model. *13th International Conference on Computer Science and Education, ICCSE 2018*, 293–296. doi:10.1109/ICCSE.2018.8468703
- Zhang, C., Wang, Y., Ji, Z., Zhao, X., Zhang, J., & Yang, Y. (2018). Predicting Chinese stock market price trend using machine learning approach. *ACM International Conference Proceeding Series*, 6–9. doi:10.1145/3207677.3277966

Smail Wafaa Denwar is an IT Officer at Ghana Electrometer Limited, an electricity manufacturing company supplying the Electricity Company of Ghana (ECG) with electricity meters; he has been working there since 2017. He obtained a Bachelor of Science in Information Technology degree at Valley View University, Accra, Ghana, from 2013 to 2017. Mr. Ismail proceeded to the University of Ghana, which led to an award of a Master of Science in Computer Science degree from 2019 to 2020, running concurrently with his job at Electrometer. He is currently assisting in the publication of research papers in other journals. His research interest is in areas of Neural Networks (Text Generation, Text Classification, Time-series forecast), Machine Learning (Classification problems such as intrusion detection based on network traffic among others), Databases, and Image Processing (particularly tumor detections). Mr. Ismail has a strong passion for research and is looking forward to doing a Ph.D. in the future; he has a passion for finding answers to solving complex problems.

Justice Kwame Appati is a lecturer in the School of Physical and Mathematical Science (SPMS) and the Department of Computer Science. He began his teaching career at Kwame Nkrumah University of Science and Technology in Kumasi as a graduate assistant and then later moved to University of Ghana in 2017 as a lecturer. Dr Appati earned a PhD, in Applied Mathematics from Kwame Nkrumah University Science and Technology in 2016. He also graduated in 2010 and 2013 with a BSc. Mathematics and MPhil Applied Mathematics from the same institution. His current research includes the automatic detection and classification of human intestinal worm and his most recent publication is "Multi-criteria ranking of voice transmission carriers of a telecommunication company using PROMETHEE, 2018". Academically, Dr. Appati is a self-disciplined and motivated person who is passionate with his job and continually challenges his students and set high expectation for them. He has also singly and jointly supervised undergraduate and postgraduate students from Kwame Nkrumah University of Science and Technology (KNUST), National Institute of Mathematical Sciences (NIMS), African Institute of Mathematical Sciences (AIMS) and University of Ghana Currently, Dr. Appati handle course like Design and Analysis of Algorithm, Artificial Intelligence, Formal Methods and Computer Vision. He looks forward to working with everyone interested in his field of study more especially, Intelligence and Data Science.