Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

A Novel Approach for Stock Price Prediction Using Gradient Boosting Machine with Feature Engineering (GBM-wFE)

Rebwar M. Nabi
Technical College of Informatics
Sulaimani Polytechnic University
Sulaimani, Iraq
Rebwar.nabi@spu.edu.ia

Soran Ab. M. Saeed
VP for Scientific Affairs
Sulaimani Polytechnic University
Sulaimani, Iraq
Soran.saeed@spu.edu.iq

Habibollah Harron
University of Technology Malaysia
Johor, Malaysia
habib@utm.my

Article Info

Volume 5 – Issue 1 – June 2020

DOI:
10.24017/science.2020.1.3

Article history:

Received: 27 January 2020
Accepted: 10 March 2020

Keywords:

Stock Market Prediction, Feature Engineering
Feature Selection
Machine Learning
Predictive Analysis
Predictable Movement
Multiclass Classification

ABSTRACT

The prediction of stock prices has become an exciting area for researchers as well as academicians due to its economic impact and potential business profits. This study proposes a novel multiclass classification ensemble learning approach for predicting stock prices based on historical data using feature engineering. The proposed approach comprises four main steps, which are pre-processing, feature selection, feature engineering, and ensemble methods. We use 11 datasets from Nasdaq and S&P 500 to ensure the accuracy of the proposed approach. Furthermore, eight feature selection algorithms are studied and implemented. More importantly, a feature engineering concept is applied to construct two new features, which are appears to be very auspicious in terms of improving classification accuracy, and this is considered the first study to use feature engineering for multiclass classification using ensemble methods.

Finally, seven ensemble machine learning (ML) algorithms are used and compared to discover the ultimate collaboration prediction model. Besides, the best feature selection algorithm is proposed. This study proposes a novel multiclass classification approach called Gradient Boosting Machine with Feature Engineering (GBM-wFE) and Principal Component Analysis (PCA) as the feature selection. We find that GBM-wFE outperforms the previous studies and the overall prediction results are auspicious, as MAPE of 0.0406% is achieved, which is considered the best result compared to the available studies in the literature.

1. INTRODUCTION

1.1. Background

The stock market prediction has fascinated enormous considerations from scholars as well as the commercial industry. However, the question still leftover in terms of whether the historical price of the stock can be used to predict future prices. [1]. Efficient Market Hypothesis (EMH) and the random walk theory(RWT) are considered as the oldest study on stock market

prediction [1], [2]. Both EMH and RWT stated that it is challenging to predict the stock prices because they are mostly affected by the news instead of historical data. Consequently, the classification accuracy had reached to 50% only[3].

Contrariwise, numerous researches [4–14] stated the opposing claim provided by EMH and RWT. These researches propose that the stock price can be forecasted. Stock prices prediction(SPP) is crucial in the financial biosphere [7], [8], [12] as a practically precise forecast can produce unique business paybacks and verge in contradiction of bazaar risks. Though, it remainders hard to predict the stock price since the financial market is a multifaceted, evolutionary, and nonlinear lively system, which interrelates with political measures, economic circumstances, and traders' opportunities [12]. Nevertheless, understanding accurate prediction of stock prices in the quick term (one day, five days forward), intermediate term (ten days, 15 days forward), (20 days, 30 days forward), and extended term (Three-monthly) is regarded as one of the furthermost striking and evocative research topics in the investment ground and its submissions. The paybacks involved in imprecise forecasts have been inspiring encouraged investigators to advance novel and forward-thinking apparatuses and approaches. On the whole, there are two communal approaches to forecast the SPP such as, Fundamental Analysis (FA) and Technical Analysis (TA). The FA uses economic features to approximate the inherent values of securities, while the TA is based on historical prices of the stock.

So far, several researches have been applied to forecast the SPP. Regarding the techniques used to examine the stock, several of the are founded on statistical approaches although the majority are built by artificial intelligence (AI) and Machine Learning (ML) algorithms [8]. Frequently, the financial data is disordered, deafening, and nonlinear, which is hardly follow immovable pattern. Consequently, statistical methods, for example "moving average", "weighted moving average filtering potential smoothing", "regression analysis", "autoregressive moving average", "autoregressive integrated moving average", and "autoregressive moving average" do not achieve acceptable CA [5]. On the other hand, AI algorithms are capable to cater the arbitrary, disordered, and nonlinear data of the stock and have been extensively applied [5]. The Artificial Neural Network (ANN), Bayesian Analysis, K-Nearest, and Decision Tree are few examples of AI algorithms [15].

Therefore, this study aims to propose a novel multiclass classification approach to forecast the SPP using feature engineering. To the best of our knowledge, our study can be considered a pioneer in studying and implementing feature engineering for stock prediction utilizing ensemble methods. To support and prove this, we have searched and investigated international databases such as Science Direct[1], Elsevier[2], Scopus[3], IEEE Digital Library[4], Springer[5], and ACM[6]. Furthermore, several other platforms and databases were investigated, for example, Google Scholar, EBSCO Information Services, and DOAJ.

We will study and compare the current feature selection(FS) algorithms to discover the top-performing algorithm. Finally, we will compare available ensemble methods and other ML algorithms to find the best classifier.

The rest of the study is structured in this manner. In sector two, the background and related work are outlined. In section three, the project methodology is explained in detail. The results and discussion can be found in section four. The conclusion of the paper is explained in section five.

2. RELATED WORK

Based on the literature, numerous studies have been published. Jie Sun et al. [16] proposed the AdaBoost support vector machine combined with concept drift on weighting time (ADASVM-

[1] https://www.sciencedirect.com
[2] https://www.elsevier.com/en-xm
[3] https://www.scopus.com/home.uri
[4] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=2188
[5] https://www.springer.com/gp
[6] https://www.acm.org/

TW) to predict the financial distress. The results were promising, as they found that the proposed ADASVM-TW was outperforming the single SVM algorithm. The authors in [17] proposed using Random Forest as an ensemble method to predict the stock return value. The algorithm was used to minimize the prediction error by dealing with the problem as the classification model. Technical indicators such as Relative Strength Index (RSI) and Stochastic Oscillator (SO) were used to train the model as the input for multiple decision trees. Dong [18] discovered the dynABE, which is the dynamic Advisor-Based Ensemble for predicting the stock price by discovering the precise parts based on the companies of curiosity and differentiating the set of features into various advisors in the way that each advisor tackles a different area and follows the proposed ensemble procedure. Dong's approach achieved a misclassification error of 31.12%. The EALasso was proposed by [19] as the feature selection approach for multiclass and multiclass learning problems by keeping the oracle belongings of recognizing the truthful subdivision prototypical and partaking the ideal approximation accurateness.

Researchers in [20] recommended the mixture of an expert system consisted of a knowledge base(KB) and AI. The KB was applied to collect historical prices, numerous eminent technical pointers, counts and sentimentality notches of available news of the stock, movements in Google for the assumed stock ticker, and the number of exclusive visitors for Wikipedia pages. In the AI, numerous ensemble algorithms were implemented such as NN regression, SV regression, boosted RT, and RF regression. The MAPE $\leq 1.50\%$ is achieved.

Feature engineering is a massive subject and numerous approaches have been proposed, predominantly in the extent of involuntary feature learning. It is commonly known that data of the stock market covering daily prices, statements of earnings by distinct companies, and view articles from experts.

When constructing new features [21], it is advantageous when the result is interpretable. explainable features and approaches are extra reachable, which this produce better forecasting results. In addition, it is a virtuous concept to add complexity to advance the CA. The main aim of feature engineering is to reach to optimal features for the task. The stock market data is ready to be investigated by mathematical theories and applications. A mathematical model defines the relations, which forecast stock prices could be a method that maps a company's receiving history, historical prices, and trade to the forecasted stock price. Researchers in [22] implemented a NN method for stock forecast and the CA improved surprisingly.

Furthermore, several studies have been conducted that implemented feature engineering; however, none are related to stock prediction. Researchers in [23] used feature engineering fault diagnosis of induction motors. A semantic feature model in concurrent engineering was conducted by researchers in [24]. Another study used feature engineering for energy theft detection using gradient boosting and found useful combinations from the origin features [25]. The authors of [26] suggested a feature engineering approach for short period earthquake forecast using AETA dataset. Feature engineering for search advertising recognition was investigated by researchers in [27]. Researchers in [28] proposed feature engineering for stock prediction but only considered the binary classification. They found significant improvement in prediction performance.

Conversely, majority of the data related to the stock market and financial stress are the data imbalance and multiclass classification [9], [15]. Imbalanced data belongs to a dataset, that one or some of the values have a considerable bigger number of samples compared to others. Typical algorithms such as "logistic regression", "SVM", and DT are appropriate for stable datasets; however, when fronting imbalanced situations, these algorithms frequently deliver suboptimal CA results. The imbalance problem and multiclass classification attract many researchers to tackle these two issues. However, to date, the published approaches are not providing good accuracy in prediction. Ensemble learning techniques have been studied only recently [29], [30]. Therefore, this can be considered as a significant room in the area, as ensemble methods have been confirmed to be outweighed compared to other algorithms [8], [15], [29], [31]. However, the main issue in providing a good voting algorithm to fuse the weight of different classifiers and provide a correct aggregated decision is not optimal as it faces the local optima problem that is tackled by heuristic techniques, making the approach very limited and not algorithmic

for the general class of problems. Therefore, based on the above literature, it can be identified that there is significant room for improvement because of inaccuracy.

3. EXPERIMENT METHODOLOGY

3.1. Research Framework

In this study research framework, several phases are literature review and problem definition, dataset collection, Data pre-processing, feature Engineering, applying feature selection, finding the best ensemble classifier, proposing a novel multiclass classification approach GBM-wFE for stock prediction and evaluation analysis. Figure 2 illustrates the overall research framework in this study.

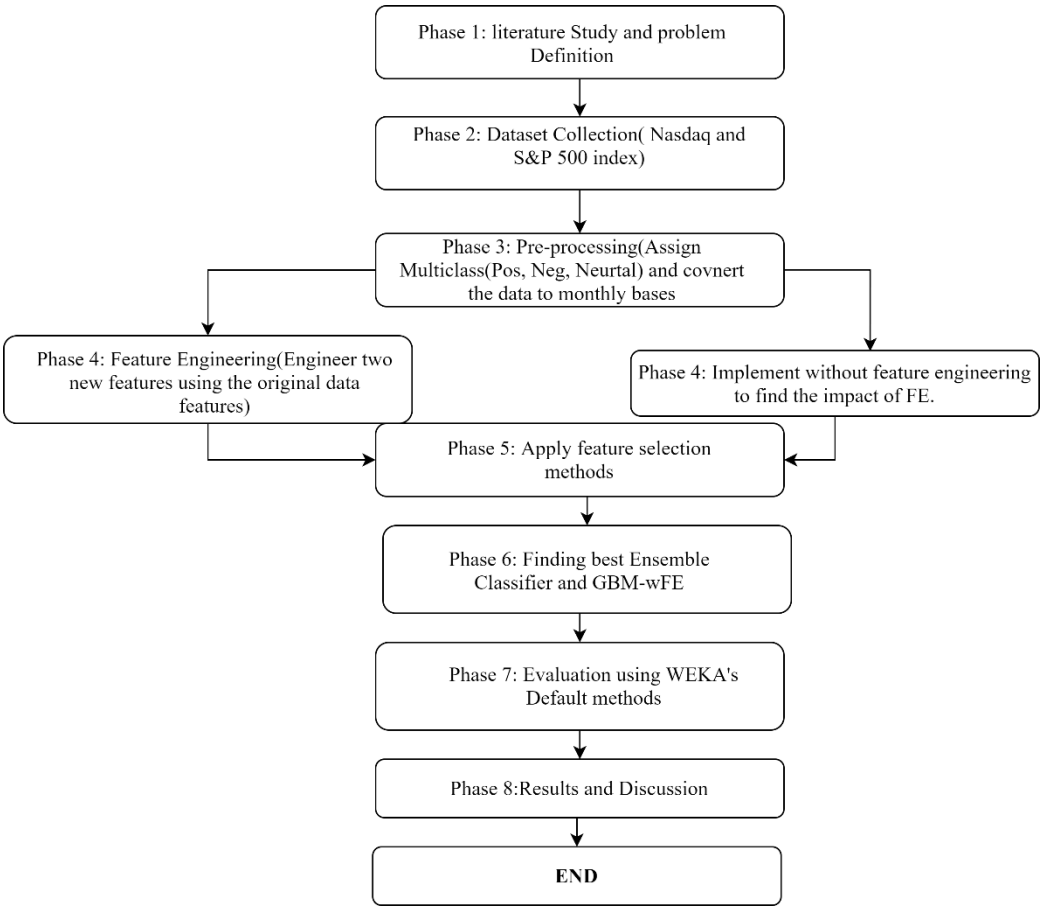

Figure 1: A research framework.

As we can see in figure 2, in the first phase, a thorough investigation and study in stock prediction is conducted to observe recent work, identify issues or problems arise and formulate a potential solution in solving the problem. In phase two, the datasets are downloaded from the international website such as the Nasdaq and S&P 500 index to evaluate our models and approaches. In this stage, the downloaded dataset is pre-processed to assign the classes for multiclass classes every month. Next, the feature engineering step is fitted in phase four, in which two new features are engineered to be added into the original dataset. Furthermore, feature selection is considered as phase five to apply different types of feature selection

algorithms that are available in WEKA. in Phase six, discovering the best ensemble method and proposing the GBM-wFE is studies. The evaluation process is conducted in phase seven and compares our contributions to the existing model. Finally, the contribution of this study can be found in phase eight which is the last stage of the research methodology.

3.2. Prediction System

For this project, the complete prediction system was developed in Java using the Waikato Environment for Knowledge Analysis (WEKA)'s Java library for ML. WEKA has a collection of ML algorithms for pre-processing, feature selection, and classification algorithms [32], [33]. We wrote and coded the system from the zero to run the experiments and evaluate the proposed approach, including feature engineering. The overall look of the prediction system is shown in the following figure.

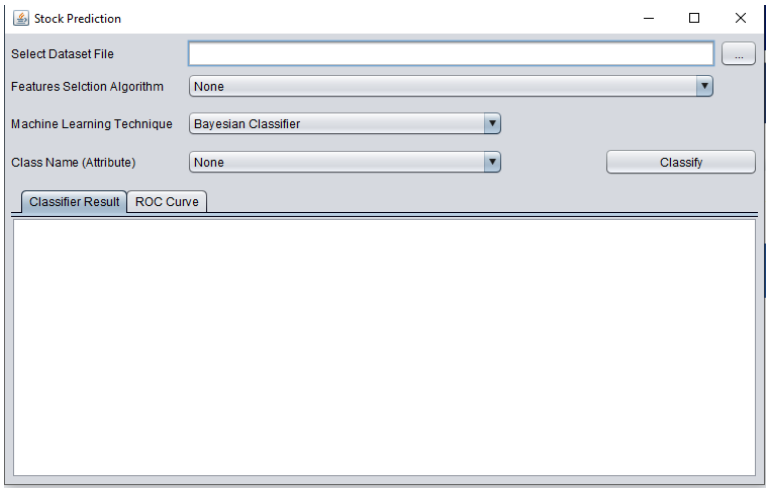

Figure 2: Overall look of the prediction system

NetBeans was used as an integrated development environment tool to develop the project. The external library was used as WEKA did not support a few algorithms.

3.3. Dataset Collection

The datasets were collected and obtained from the NASDAQ and S&P 500 index. In general, 25 years of historical data were downloaded for the CMCSA, CSCO, AAPL, SBUX, LRCX, MCHP, MSFT, NTAP, QCOM, SWSK and in S&P 500 the GSPC was chosen, which is the historical data of top 500 companies for the last 25 years. The duration of the data is Jan 1995 to Jan 2020. In total, we have extracted 3270 months records as the monthly based prediction. In general, 65% of the dataset is used for training and 35% is used for testing purposes.
The original downloaded data is daily bases and generally, each dataset has around 6294 records of the historical data. Here the explanation of stock market data based on duration such as daily, weekly, monthly, quarterly, yearly:

- Daily data is Single day Data.
- Weekly Data as 5 days Per Week – i.e. every day except Saturdays and Sundays.
- Monthly Data with Index – i.e. every month, with an index in December.
- Quarterly (Combined Months) – i.e. 4 issues per annum.
- Yearly Data means i.e. 2020.

In total, each dataset has six attributes:

1. Date: The current date of the stock movement.
2. A close price: The closing price of the stock.
3. Volume: The number of shares has been exchanged in a day.

4. open price: Open price of a stock.
5. high price: The highest price during a given day.
6. low price: The lowest price during a given day.

A sample of the downloaded data is shown in the below table.

Table 1: Sample of downloaded data

date	close	volume	open	high	low
2/25/2019	50.73	19,857,750	51.09	51.27	50.62
2/22/2019	51.17	28,022,740	50.73	51.17	50.64
2/21/2019	50.72	35,613,260	50.63	50.86	50.34
2/27/2009	4.91	27772220	4.88	5.08	4.88
2/26/2009	5.10	25730550	5.57	5.57	5.03
2/25/2009	5.35	20824750	5.38	5.51	5.14

3.4. Pre-Processing

It is widely known that pre-processing is regarded as a vital step in ML and data mining. Therefore, in our study, we suggest a new method to pre-process the data collected from Nasdaq. The stock movement to compare the predicted and real percentage change every month assigns the class to monthly data.

To find the monthly movement here, stock movement is the difference between the monthly close and open price:

Difference = close price (last date of the month) – open price (first date of the month)

The stock price movement in terms of percentage (%) is calculated as follows:

Percentage_Difference = Difference / open price (first date of the month)

For assigning the classification class in a multiclass classification case:

If Percentage_Difference >1, then the class is positive;
If Percentage_Difference <-1, then the class is negative;
Otherwise, the class is neutral.

The output dataset attributes are described below:

1. Month (based on that date)
2. Close (month-end date close price)
3. Volume (whole month daily volume addition)
4. Volume (whole month daily volume addition)
5. High (the highest price of the month)
6. Low (lowest price of the month)
7. Generated classes (multiclass).

Multiclass dataset for monthly based for every ten companies generated so in the output total of 20 datasets. Multiclass "MSFT3.csv". dataset of MSFT (multiclass), as shown in table two.

Table 2: generating the multiclass classification for the MSFT dataset.

M.	Close	V	Open	High	Low	Result
12	101.57	9.38E+08	113	113.42	93.96	negative
11	110.89	7.17E+08	107.05	112.24	99.3528	positive
10	106.81	9.2E+08	114.75	116.18	100.11	negative
9	114.37	4.7E+08	110.85	115.29	107.23	positive
8	112.33	4.54E+08	106.03	112.777	104.84	positive
7	106.08	5.6E+08	98.1	111.15	98	positive
6	98.61	5.96E+08	99.28	102.69	97.26	neutral
5	98.84	5.06E+08	93.21	99.99	92.45	positive
4	93.52	6.64E+08	90.47	97.9	87.51	positive
3	91.27	7.45E+08	93.99	97.24	87.08	negative
2	93.77	7.21E+08	94.79	96.07	83.83	Negative
1	95.01	5.68E+08	86.125	95.45	85.5	positive

To sum up, the total data after pre-processing is 3270 rows from 25 years as months, of which %65 are used for training and the other %35 are used for testing purposes.

3.5. Feature Engineering Implementation

As mentioned earlier, the study aimed to investigate and add new features to improve the classification accuracy. Two new features were added to the dataset to study the impact on improving accuracy. The first new feature is named High_Low_Difference (HL_Diff), which is defined as the difference of month's high and low price. The mean value of close open difference as daily bases was also constructed. The following mathematical equations were used to produce new features:

$$\mathrm{HL}_{\mathrm{Diff}} = High_{\mathrm{Max}} - Low_{Min} \qquad \text{eq (1)}$$

Where:

$$High_{\mathrm{Max}} = \text{Maximum high price of month}$$
$$Low_{\mathrm{Min}} = \text{Minimum low price of month}$$

$$Mean = \frac{\sum f\ (close - open)}{total\ days} \qquad \text{eq (2)}$$

Where:

$$\sum f = Sum\ of\ the \text{ difference of close and open price}$$

And:

$$total\ days = number\ of\ days\ in\ the\ duration$$

As can be seen in equation one, **HL_Diff** is calculated by the difference between the high and low month in total. It displays the entire month's supreme movement in the price. The **mean** is calculated based on the mean values of all differences between close and open prices. That shows the average movement in price.

The system automatically generates a new CSV file under the name "MSFT _3F.csv." We programmed the system to create multiple files with feature engineering and without feature engineering to compare results later. Table three demonstrates the newly constructed features after feature engineering.

Table 3: Features added dataset of MSFT for multiclass

Month	Close	Open	High	Low	HL_Diff	Mean	Res
12	101.57	113	113.42	93.96	19.46	-0.58211	negative
11	110.89	107.05	112.24	99.3528	12.8872	0.098095	positive
10	106.81	114.75	116.18	100.11	16.07	-0.62174	negative
9	114.37	110.85	115.29	107.23	8.06	0.115263	positive
8	112.33	106.03	112.777	104.84	7.937	0.219783	positive
7	106.08	98.1	111.15	98	13.15	-0.03738	positive
6	98.61	99.28	102.69	97.26	5.43	-0.19271	neutral
5	98.84	93.21	99.99	92.45	7.54	0.2677	positive
4	93.52	90.47	97.9	87.51	10.39	-0.22314	positive
3	91.27	93.99	97.24	87.08	10.16	-0.2769	negative
2	93.77	94.79	96.07	83.83	12.24	0.003684	negative
1	95.01	86.125	95.45	85.5	9.95	0.117619	positive

3.6. Implementation of Feature Selection Algorithm

In this study, to achieve one of the aims, multiple feature selection algorithms were used to find the best feature selection algorithm for a multiclass classification approach. It is worth mentioning that the WEKA's default configuration was implemented for all algorithms, which

means no parameter configuration was considered since it was not within the scope of this study. Generally, the following algorithm was considered:

.1. "Sequential Feature Selection (Best First) Search and CFS Subset Evaluation" (SEQ)
.2. "Genetic Search and CFS Subset" (GEN)
.3. "Ranker Search and Chi-Squared" (CHI)
.4. "Ranker Search and Recursive Feature Elimination"(REF)
.5. "Ranker Search and Correlation Coefficient"(CC)
.6. "Ranker Search and Info Gain Evaluation" (IG)
.7. "Ranker Search and ReliefF and it is Variant Evaluation" (RV)
.8. "Ranker Search and Principle Components Analysis Evaluation" (PCA)

To discover the best feature selection, we proposed an approach of which the flowchart is shown in figure three. The developed prediction system runs intensive experiments for all feature selection algorithms and produces the best-performing one to be considered in the overall approach which is proposed in the study.

3.7. Implementation of Ensemble Classifier Techniques

As the classifier, we implemented seven ensemble learning algorithms, all of which are used as a default configuration in the WEKA application programming interface library, which technically means we did not play with the parameters, base learners, and other parameters. The following algorithms were chosen for this study:

1. Bagging Classifier (BAG)

Bagging bags algorithm to decrease variance. Forecasts are produced by be an average of probability approximations, not by voting. One of the parameters is called the size of the bags as a proportion of the training dataset. Furthermore, another parameter is whether to compute the out-of-bag error, which tells the average error of the ensemble members [33], [34].

2. Stacking Classifier (SC)

In the SC the classifiers will be combined by using stacking for classification and regression problems. The base classifiers will be specified, the meta-learner, and the number of cross-validation folds.

3. Voting Ensemble Classifier (VE)

The baseline approach is provided by VE for combining classifiers. The default outline is to average their probability approximations or numeric forecasts for classification. Moreover, the other grouping outlines are obtainable, for example, using common voting for classification.

4. AdaBoost Classifier (ADA)

In the ADA the classic boosting is applied. It can be enhanced by specifying a threshold for weight pruning. ADA resamples if the base classifier cannot lever weighted occurrences (you can also force resampling) [33].

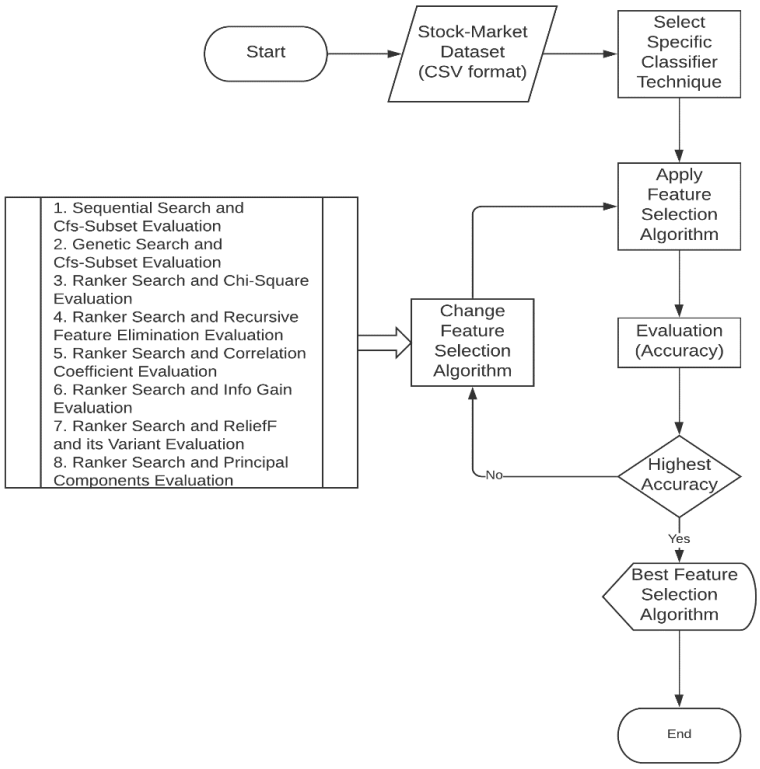

Figure 3: Finding the best feature selection algorithm flowchart

5. Gradient Boosting Machine (GBM)

Gradient Boosting Machine [35] (also referred to as slope boosted designs) sequentially fits brand-new versions to supply an extra exact price quote of a response variable in supervised knowing jobs such as regression and classification. GBM is a set of either regression or classification tree versions. Both are forward-learning set approaches that acquire anticipating outcomes using slowly boosted estimations. LogitBoost in WEKA provides similar results for GBM.

6. Multi-Boosting Classifier (MB)

Multi-Boosting is an ensemble strategy employed for improving the results of single classifiers, which is an extension of AdaBoost [36]. The input vectors are weighted and some of them have a higher chance to contribute to the new sets. Two types of weights are defined in the boosting strategy: the first for adjusting the contribution of data points (Bi), and the second for the integration of the single classifiers.

7. Random Forest (Random Subspace) (RF)

RF is a homogeneous ensemble prediction approach created by incorporating multiple decision trees as base learners. It is a bagging-based ensemble created using multiple decision trees and passing a subset of data to each of these base learners for training. The combiner provides the final results. Random Forest provides consistency to the model and thus provides a robust classifier. It tends to solve the over-fitting issue contained in Decision Trees [37].

8. Finding the Best Classifier (Finding the best from the eight above)

Another goal of the study was to discover the top-performing ensemble. The developed system runs an exhaustive comparison among the implemented algorithms to propose and select the

outperforming algorithm. The following flowchart demonstrates the proposed approach to achieve this aim.

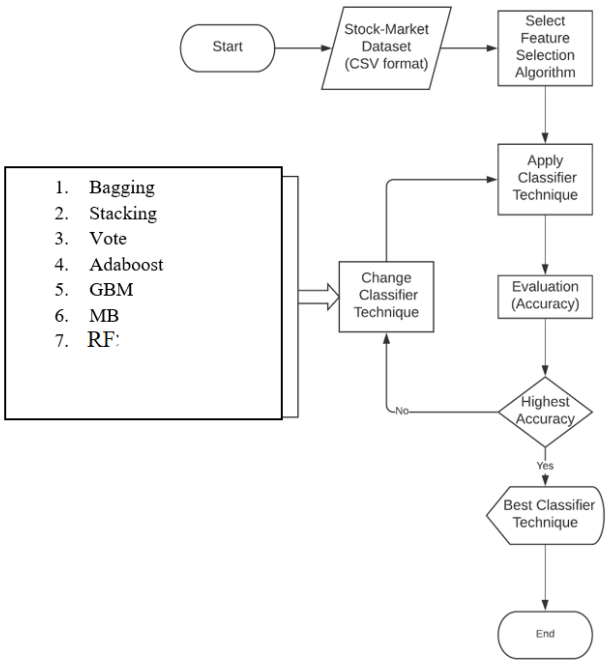

Figure 4: Find the ensemble classifier flowchart

3.8. Predicted Class (CSV File)

The developed system creates a file called "ClassPredict.csv", which contains the actual classes and predicted classes, so it has easily compared the file as input dataset class and output dataset class. The testing dataset has 29 months' records, so it contains 29 actual and predicted records. Here, for example, the testing dataset includes CMCSA company 29 months record's Predicted classes as output is shown in the below table.

Table 4: Input file dataset and predicted class file

Original Data CMCSA		Prediction Data			
Open	High	Open	High	Open	High
36.71	38.73	36.71	38.73	36.71	38.73
33.49	37.42	33.49	37.42	33.49	37.42
39.09	39.29	39.09	39.29	39.09	39.29

3.9. Evaluation Method

For a comparative study of the supervised learning algorithms for stock market prediction, we followed and used the WEKA library default evaluation methods [33]. To evaluate our works, we have used several evaluation metrics such as CA, Precision(PR), Recall(RE), F-Score(F1S), Mean Absolute Percentage Error (MAPE), Kappa Statistics(KAPPA) and Root Mean Squared Error (RMSE) and other available methods in WEKA accordingly to benchmark our proposed approach.

For reliable testing and results, we have divided our data into training and testing. Approximately, 65% used for training purposes, and 35% are used for testing. It is widely

known that WEKA provides various methods to evaluate classifiers such as Training Time (TT), CA, Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICI), Kappa, MAE, RMSE, Relative Absolute Error (RAE), and Root Relative Squared Error (RRSR), PR, RE, F1S and provides the Confusion Matrix (CM). Since we are using the default WEKA evaluation, we are not going to provide the equations and mathematical details behind them as it would lead to repetition. Details of all the evaluation methods can be found in [33], [38].

4. RESULTS AND DISCUSSION

4.1. Results

We conducted various experiments on all datasets of 11 companies. Besides, all the classification methods and FS algorithms on each dataset and a separate graph were generated for each experiment. To compare and evaluate feature engineering as adding new features to the dataset, each graph displays two dataset results. The first is a dataset with added features represented by a solid line and the second is a dataset without added features represented by a dotted line. An example of the generated graph is shown in the following figure.

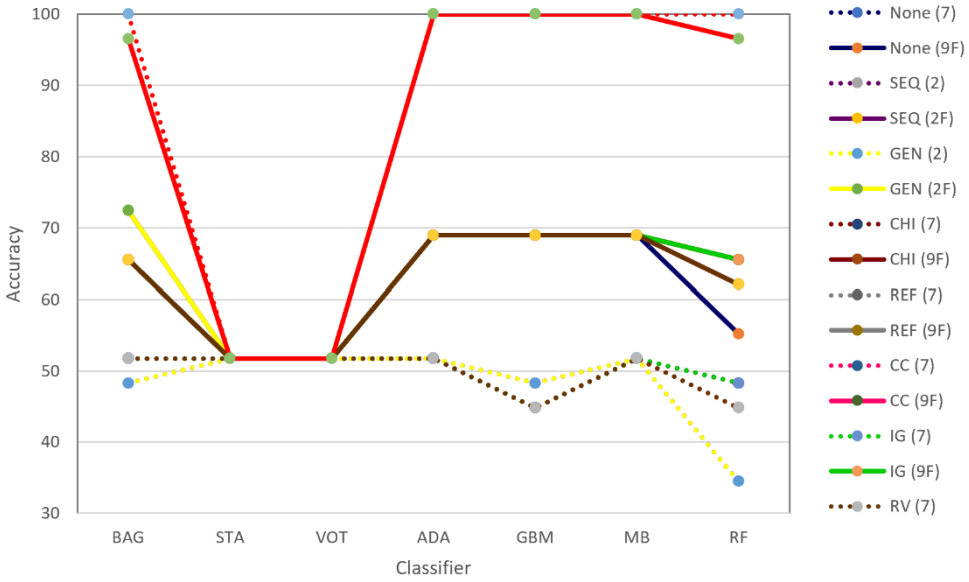

Figure 5: Performance comparison graph

The example of a performance comparison graph shown in figure three contains the following:

- The X-Axis signifies the algorithm sign – all seven classifiers
- Y-Axis represents prediction accuracy in percentage
- Various lines, the solid line is with FE (9F) and the dotted line is without FE (7F)
- Nine colors of the line with a dissimilar color for each FS method
- PCA (6F) (six features chosen and F for with feature).

Figure six illustrates the overall classification prediction on the LRCX company dataset. As can be seen in the majority of cases, the PCA outperforms all the other feature selection algorithms, in which, with few classifiers, the accuracy of 96.55% is achieved. Furthermore, various classifiers work better than the other classifiers when PCA is considered, for example, classifiers such as ADA and RF. Conversely, classifiers such as VOT and STA perform poorly, and in some cases, the accuracy of less than 65% is achieved.

When comparing the difference in feature engineering, it can be seen that in several experiments the accuracy is improved while sometimes it decreases. For instance, when GEN feature selection is considered, the feature engineering has improved the classification accuracy

significantly, in which for the BAG algorithm, the accuracy of approximately 50% is achieved, whereas, without feature engineering, less than 50% is achieved.

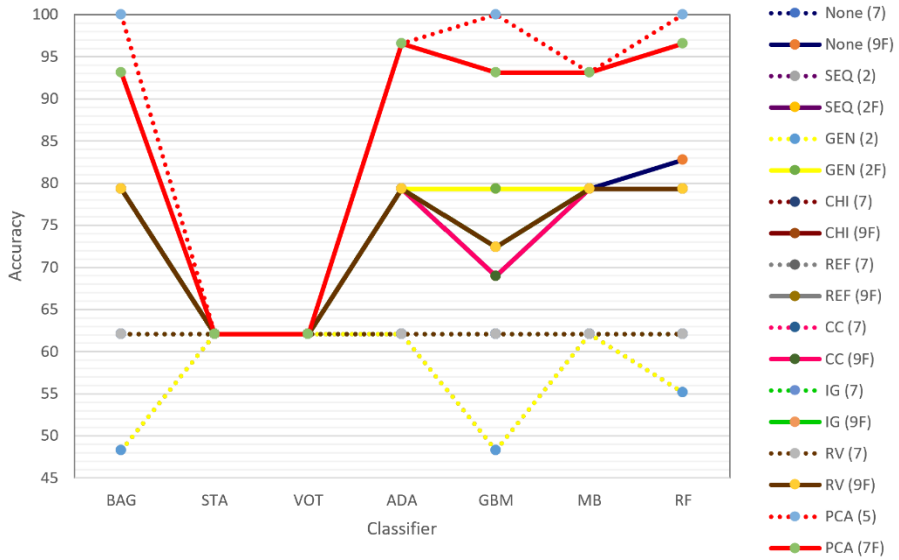

Figure 6: Classification result for LRCX company dataset

Furthermore, the overall CA result for the SWKS company dataset is shown in figure seven Similar to the LRCX company, the PCA algorithm with majority classifiers outclasses others, where for classifiers such as BAG, ADA, GBM, and RF, the accuracy of 100% is achieved. Conversely, few algorithms have low accuracies such as STA and VOT. Moreover, feature engineering has contributed significantly to improving the classification accuracy. For instance, the RF with GEN achieves approximately 67% accuracy when tested with feature engineering, whereas around 50% accuracy is achieved without feature engineering. To sum up, on the SWSK company dataset, PCA is also considered as the best feature selection algorithm, and the Genetic Algorithm comes second. Finally, feature engineering contributed significantly to improving the overall classification accuracy.

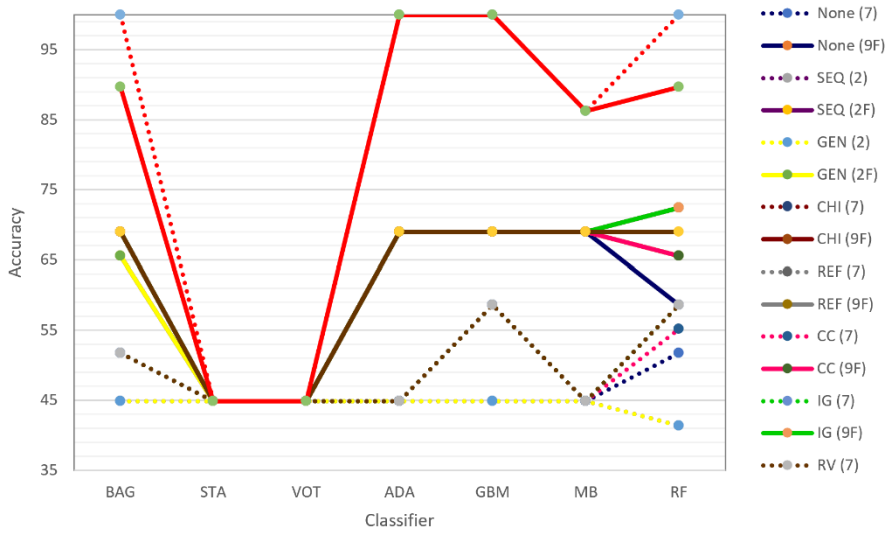

Figure 7: Classification result for SWSK dataset

The overall result prediction for the MCHP, MSFT, and NTAP datasets can be seen in figures eight, nine, and ten, respectively.

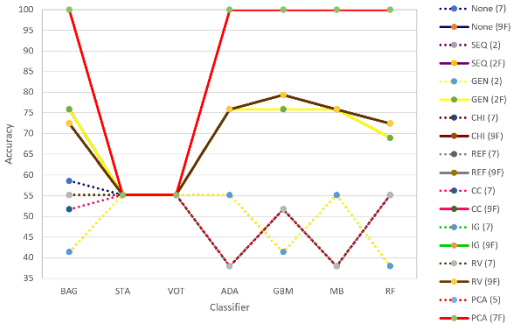

Figure 8: Classification result for MCHP dataset

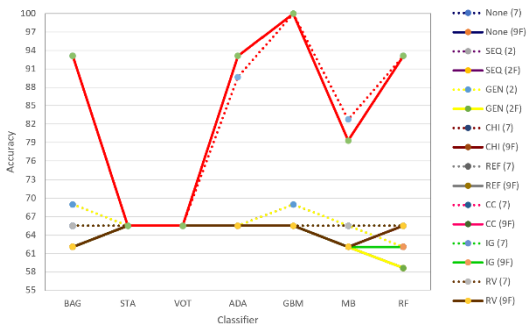

Figure 9: Classification result for MSFT dataset

,

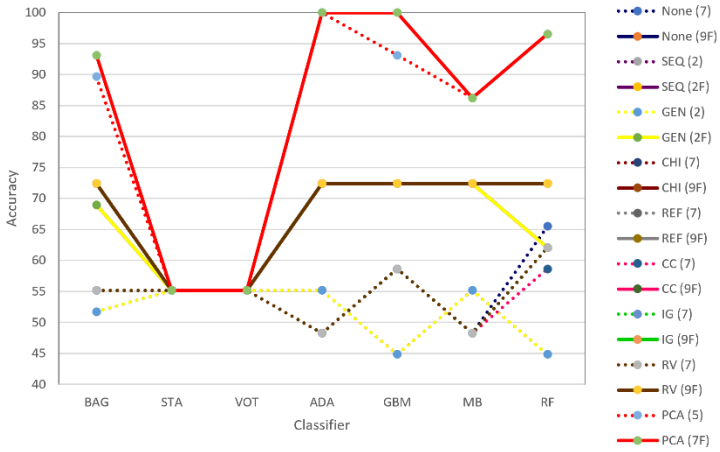

Figure 10: Classification result for NTAP dataset.

4.1.1. GBM-wFE Prediction Results

Table 4.6 describes the overall CA results of the projected approach (GBM-wFE) based on various evaluation metrics available in WEKA.

As it can be noticed that our proposed approach performance is outstanding, whereabouts the average CA is 99.28% on all datasets. To be more precise, our approach achieved 100% CA on AAPL, SBUX, MHCP, LRCX, MSFT, NTAP, QCOM, and GSPC. Furthermore, on CMCSA and CSCO the CA of 99.03% and 94.05% is achieved respectively.

However, the lowest CA percentage is noticed when the CSCO dataset is considered this could be because of the fewer number of records compared to other datasets. Furthermore, we have also calculated the F-Measures, precision, MAPE, RMSE, KAPPA statistics, and recall metrics so that we can use them during the benchmark and comparison with previous studies.

Our approach has achieved the MAPE of 0.19% and F-Measures of 0.99 out of one, which these results prove that the model enjoys significant success compare to the existing studies. In the next section, we will benchmark the GBM-wFE with literature.

Table 5:GBM-wFE prediction result.

Algorithm	Stock	Dataset	CA%	MAPE%	RMSE	KAPPA	F-Measures	Precision	Recall
GBM-wFE	Nasdaq	CMCSA	99.03	0.53	0.06	0.98	0.98	0.98	0.99
		AAPL	100.00	0.00	0.00	1.00	1.00	1.00	1.00
		CSCO	94.05	0.48	0.20	0.89	0.90	0.94	0.94
		SBUX	100.00	0.00	0.00	1.00	1.00	1.00	1.00
		LRCX	100.00	0.07	0.00	1.00	1.00	1.00	1.00
		MHCP	100.00	0.34	0.04	1.00	1.00	1.00	1.00
		MSFT	100.00	0.01	0.00	1.00	1.00	1.00	1.00
		NTAP	100.00	0.07	0.00	1.00	1.00	1.00	1.00
		QCOM	100.00	0.00	0.00	1.00	1.00	1.00	1.00
		SWKS	99.03	0.56	0.07	0.98	0.99	0.99	0.98
	S&P	GSPC	100.00	0.00	0.00	1.00	1.00	1.00	1.00
	Average		99.28	0.19	0.03	0.99	0.99	0.99	0.99

4.2. Discussion

4.2.1. Best Feature Selection

To find the best feature selection algorithm, we have run intensive experiments on the CMCA data using GBM for all eight feature selections algorithm we have considered in this study. Table 6 demonstrates the performance evaluation of all feature selection algorithm using CA evaluation metric.

Table 6:Find best feature selection using GBM.

N	Feature Selection Alg.	CA%
1	SEQ	70.19
2	GEN	70.19
3	CHI	31.73
4	REF	31.73
5	CC	54.80
6	IG	31.73
7	RV	54.80
8	PCA	99.03

As it can be observed in table five, the PCA is outperforming all the feature selection algorithms by achieving 99.03% of CA. whereas, the rest of the feature selection algorithms have almost

reached the same CA, which is relatively lower than PCA accuracy. Based on the achieved results, the PCA is going to be used to find the best classifier and the rest experiments.

4.2.2. The Best Ensemble Classifier

As mentioned earlier, another aim of this research was to discover the top-performing classifier. Therefore, to achieve this goal, we conducted intensive experiments on the 11 companies' datasets with FE and used PCA as the best FS algorithm with the seven ensemble classifiers considered in this study. Table seven shows the multiclass classification results with feature engineering.

Table 7: Multiclass classification with feature engineering using PCA

	CA% CMCSA	CA% CSCO	CA% AAPL	CA% LRCX	CA% SBUX	CA% MCHP	CA% MSFT	CA% NTAP	CA% QCOM	CA% SWKS	CA% GSPC	AVG%
Alg\DS	NASDAQ										S&P 500	
BAG	99.03	92.05	100	100	100	99.03	100	97.02	100	100	100	98.83
STA	49.03	54.45	45.19	47.11	52.88	52.88	58.65	60.39	55.76	49.03	52.88	52.57
VOT	49.03	54.45	45.19	47.11	52.88	52.88	58.65	60.39	55.76	49.03	52.88	52.57
ADA	99.03	93.06	100	96.15	100	99.03	100	97.02	100	99.03	100	98.48
GBM	99.03	94.05	100	100	100	100	100	100	100	99.03	100	99.28
MB	90.38	93.06	95.19	98.07	100	90.38	98.07	100	100	99.03	100	96.74
RF	100	93.06	100	100	100	99.03	100	98.01	100	100	100	99.10

As can be seen in seven, the overall prediction approach is tested on all datasets using the implemented ensemble classifiers, which have been considered in this study. GBM found to be outperforming all other ensemble classifiers by achieving the CA 99.28% on average on all datasets. Moreover, RF, BAG, and ADA were found to be very efficient as well, which on average, the accuracy of 99.10%, 98.83%, and 98.48% achieved respectively on all datasets. Conversely, STA and VOT were identified as the worst-performing ensembles by reaching the CA of 52.57% on average. Last but not least, the rest of the ensembles were ranked as the middle performance by achieving the CA of 98% and above on average approximately. Therefore, based on these results GBM will be chosen to develop the stock prediction approach along with feature engineering and PCA as feature selection.

4.2.3. Feature Engineering Benchmarking with WEKA

To explore the contribution and the significance of the proposed feature engineering approach, we conducted intensive experiments on several datasets. The CA result on all datasets with all feature selection algorithms is outstandingly improved. The average CA for SEQ and GEN is impressively boosted with feature engineering by achieving 69.23% whereas the average CA is only 44.55% without feature engineering. Correspondingly, the average CA is increased by 25.64% when the CHI, RFE, CC, IG, and RV considered. Besides, the average CA is also enhanced with PCA with feature engineering by 1.93%. Last but not least, on average, the CA with feature engineering is 55.12% while it is increased to 74.67% with feature engineering.

4.2.4. GBM-wFE Approach Benchmarking

The prediction performances of the GBM-wFE approach and benchmark will be demonstrated in this section. Tables eight, nine, ten, and 11 show the comparison result of MAPE, Kappa statistics, and RMSE evaluation criteria. The proposed GBM-wFE was found to be outperforming the available benchmark for stock prediction using ensemble methods.

Researchers in [40] proposed AdaBoost-LSTM (Long Short-Term Memory) and AdaBoost with a few other algorithms such as MLP, support vector regression (SVR), and ELM for financial time series forecasting using the stock index. In table eight, we compare our proposed model with their results.

Table 8: Benchmark comparison using MAPE with [40]

	Apparoch or Model	**MAPE**
Their Model	AdaBoost-MLP	1.023
	AdaBoost-SVR	0.841
	AdaBoost-ELM	0.782
	daBoost-LSTM	**0.413**
Our Model	GBM-wFE	**0.19**

As can be seen in table nine, our proposed ensemble method GBM-wFE is outperforming all the other proposed ensemble methods. GBM-wFE achieved a MAPE of 0.19% while their best model, which was AdaBoost-LSTM, achieved a MAPE of 0.413%.
In another benchmarking comparison with previous studies, table nine shows the result of the comparison of the proposed GBM-wFE model with the results of this study [41]. Researchers used an ensemble of Recurrent Neural Network (RNN) with LSTM on historical data.

Table 9: Benchmark comparison using RMSE with [41]

Benchmark	Ensemble Model	**Dataset**	**RMSE%**
Their Model	RNN-LSTM	POWERGRID	0.410
		SUBEX	0.413
		INDBANK	0.201
		GREENPLY	**0.011**
Our Model	GBM-wFE	SBUX	**0.0**
		CMCSA	**0.06**
		CSCO	**0.04**
		MSFT	**0**

As can be observed from the above table, our proposed GBM-wFE ensemble is outperforming the RNN-LSTM ensemble method. The GBM-wFE ensemble achieved an RMSE of 0.0% on the SBUX dataset, whereas in their best result, they achieved an RMSE of 0.011%. Accordingly, all the other results show our approach is performing better in all the datasets.
Furthermore, another study proposed an ensemble approach for stock price prediction using historical data [42], which is Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). The TOPSIS uses crow search based weighted voting classifier ensemble. The benchmarking and the result comparison are shown in table 10 using the S&P and NIFTY dataset. It is worth mentioning that researchers used and tested various classifiers as the base learner for their ensemble; however, we will choose their best accuracy result to compare it to our model.

Table 10: Result comparison with TOPSIS ensemble model

Benchmark	**Ensemble Model**	**Stock**	**CA%**
Their Models	TOPSIS-MV	S&P	82
	TOPSIS-WV		82.5
	TOPSIS-PSO-WV		84.5
	TOPSIS-DE-WV	NIFTY	81
	TOPSIS-CS-WV		84
Our Model	GBM-wFE	S&P 500	100
		NASDAQ	99.21

	Average	99.28

Our proposed ensemble is superior to the proposed TOPSIS ensemble with a different ensemble classifier. We can see in the table ten that the best CA achieved with TOPSIS-DE-WV is 81% on NIFTY and 82% on S&P 500 Stock, whereas the GBM-wFE achieved 100% on S&P 500 Stock and 99.21% on NASDAQ. Accordingly, GBM-wFE on average have achieved the CA of 99.28% but their average CA is not calculated. However, based on the table 4.13 it can be calculated, which is around 82.8%. So, we can conclude that our approach is superior to TOPSIS approach by 16.48%.

Finally, in the table 11, we also cite a few other studies to benchmark with the proposed GBM-wFE. We indicate and cite the research paper and its results. The following table demonstrate the superiority of our proposed GBM-wFE over the approaches used in studies available in the literature.

Table 11: Result comparison using ensemble models in the literature

Benchmark	Ensemble Model	Stock	RMSE	Kappa%
Literature Model	Adaboost [43]	NSE	-	0.4263
	Stacking [43]		-	0.5516
	FFNN [44]	YF	0.0201	Not Provided
Our Model	GBM-wFE	S&P 500	0.00	1
		NASDAQ	0.04	0.99

Table 11 elaborates the contribution and novelty of the proposed GBM-wFE ensemble model compared with the previous studies. Researchers in [43]proposed AdaBoost and the stacking ensemble method to predict the stock price movement and achieved a Kappa of 0.5516% as the best result; however, our model achieved a Kappa of (0.99) with the NASDAQ and Kappa of (1) with the S&P 500. Furthermore, our proposed model surpasses the proposed ensemble model in [44], in which an RMSE of 0.00 and 0.04 achieved for S&P 500 and NASDAQ respectively, whereas they had an RMSE of 0.02 in the best situation.

4.2.5. Summary

Based on the benchmarking and the comparison results in table's seven to 11, it can be concluded that this study has contributed significantly to the stock market prediction by proposing the novel multiclass classification using GBM-wFE.

This study proposed and proved that using feature engineering can significantly improve the accuracy of any ensemble model and can even improve the overall prediction model. It is worth mentioning that our study can be viewed as the first to consider feature engineering for multiclass classification when stock markets are used.

Moreover, this study proposed the GBM-wFE, which has been proven to outperform the ensemble methods used in studies in the literature, as the best MAPE, RMSE, CA, and Kappa statistics were achieved with better results.

5. CONCLUSION

The study aimed to propose a novel feature engineering approach for multiclass classification for stock prediction. It explored the best feature selection algorithms which are currently available on the WEKA library. It also aimed to find the best ensemble learning algorithm. Finally, it aimed to find the ultimate collaboration between feature engineering, feature selection, and ensemble classifiers.

This study collected Nasdaq and S&P 500 index listed stocks for the last 25 years as the dataset. The dataset included data of various companies, such as CMCSA, CSCO, AAPL, SBUX, LRCX, MCHP, MSFT, NTAP, QCOM, and SWKS. Monthly stock movement is predicted for

each month. We have implemented feature engineering to add two features to the dataset as 1. Mean value of Open and Close price difference and 2. The high low difference, which is part of feature engineering that improved the performance, shows in the results as increasing accuracy. The technology uses the interface of Java and WEKA to judge varied styles of feature selection and classifier over the given dataset. For the feature selection part, various algorithms were applied, which are CFS Subset, Chi-Squared, Recursive Feature Elimination, Correlation Coefficient, Info Gain, ReliefF and its Variant, PCA, Sequential Feature Selection (Best First), Genetic Search, and Ranker Search, for ML techniques applied different classifiers on datasets, such as Stacking, AdaBoost, GBM, Multi-Boosting, and Random Forest. We tested all the techniques using multiclass classification on stock market movement as positive, negative, and neutral.

In this project work, new features were added to the dataset and it was found that the accuracy of the prediction improved. The study proposed GBM-wFE which is found to be improved on average on all datasets and outperform the available studies in the literature as well. Furthermore, we recommended the best feature selection technique as PCA and the best ensemble classifier as GBM.

Numerous future works can be suggested. First, feature engineering can be extended by considering external factors such as growth domestic products (GDP) and calculating and engineering more features to be added to the feature set. Second, the daily price movement can be used instead of monthly movements. Third, the proposed approach can be tested and implemented on a large number of datasets, which can include 50 years instead of 25 years of data. Finally, some algorithms for feature engineering can be designed and proposed to be added to the WEKA library.

Acknowledgments

I would like to show my massive appreciation to my supervisor Prof. Dr. Soran for his continuous support. Furthermore, I would like to thank my both second supervisor, Prof. Hamido Fujita, without his support and contributions. I would not reach this level. Moreover, appreciation must also go to Prof. Habib bin Harron for his friendly and fantastic support. I would also like to thanks Sulaimani Polytechnic University for giving this fabulous opportunity to study Ph.D. Lastly. I should forget my lovely wife for her patient and support for the past three years.

REFERENCES

[1] E. F. Fama, "The Behavior of Stock-Market Prices," *J. Bus.*, vol. 38, no. 1, pp. 34–105, 1965.

[2] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll, "The Adjustment of Stock Prices to New Information," *Int. Econ. Rev. (Philadelphia).*, vol. 10, no. 1, pp. 1–21, 1969.

[3] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market."

[4] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, 2015.

[5] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction," *Expert Syst. Appl.*, vol. 80, pp. 340–355, Sep. 2017.

[6] T. A., "Improvement on Classification Models of Multiple Classes through Effectual Processes," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, 2015.

[7] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Syst. Appl.*, vol. 83, pp. 187–205, Oct. 2017.

[8] R. T. Farias Nazário, J. L. e Silva, V. A. Sobreiro, and H. Kimura, "A literature review of technical analysis on stock markets," *Q. Rev. Econ. Financ.*, vol. 66, pp. 115–126, 2017.

[9] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[10] L. Wang, Z. Wang, S. Zhao, and S. Tan, "Stock market trend prediction using dynamical Bayesian factor graph," *Expert Syst. Appl.*, vol. 42, no. 15, pp. 6267–6275, 2015.

[11] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," *J. Econ. Financ. Adm. Sci.*, vol. 21, no. 41, pp. 89–93, 2016.

[12] A. Nayak, M. M. M. Pai, and R. M. Pai, "Prediction Models for Indian Stock Market," *Procedia Comput. Sci.*, vol. 89, pp. 441–449, 2016.

[13] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Syst. Appl.*, vol. 79, pp. 153–163, Aug. 2017.

[14] Y. Zhao, J. Li, and L. Yu, "A deep learning ensemble approach for crude oil price forecasting," *Energy Econ.*, vol. 66, pp. 9–16, 2017.
[15] L. Zhou, Y. W. Si, and H. Fujita, "Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method," *Knowledge-Based Syst.*, vol. 128, pp. 93–101, 2017.
[16] J. Sun, H. Fujita, P. Chen, and H. Li, "Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble," *Knowledge-Based Syst.*, vol. 120, pp. 4–14, 2017.
[17] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," vol. 00, no. 00, pp. 1–20, 2016.
[18] Z. Dong, "Dynamic Advisor-Based Ensemble (dynABE): Case study in stock trend prediction of critical metal companies," 2019.
[19] S.-B. Chen, Y.-M. Zhang, C. H. Q. Ding, J. Zhang, and B. Luo, "Extended adaptive Lasso for multi-class and multi-label feature selection," *Knowledge-Based Syst.*, vol. 173, pp. 28–36, Jun. 2019.
[20] B. Weng, L. Lu, X. Wang, F. M. Megahed, and W. Martinez, "Predicting short-term stock prices using ensemble methods and online data sources," *Expert Syst. Appl.*, vol. 112, pp. 258–273, 2018.
[21] U. Khurana, D. Turaga, H. Samulowitz, and S. Parthasrathy, "Cognito: Automated feature engineering for supervised learning," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 1304–1307.
[22] W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," *Knowledge-Based Syst.*, vol. 164, pp. 163–173, 2019.
[23] P. S. Panigrahy, D. Santra, and P. Chattopadhyay, "Feature engineering in fault diagnosis of induction motor," in *2017 3rd International Conference on Condition Assessment Techniques in Electrical Systems, CATCON 2017 - Proceedings*, 2018, vol. 2018-Janua, pp. 306–310.
[24] Y. J. Liu, K. L. Lai, G. Dai, and M. M. F. Yuen, "A semantic feature model in concurrent engineering," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 3, pp. 659–665, Jul. 2010.
[25] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019.
[26] J. Huang, X. Wang, S. Yong, and Y. Feng, "A feature enginering framework for short-term earthquake prediction based on AETA data," in *Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019*, 2019, pp. 563–566.
[27] Y. Sun and G. Yang, "Feature engineering for search advertising recognition," in *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 2019, pp. 1859–1864.
[28] R. M. Nabi *et al.*, "Ultimate Prediction of Stock Market Price Movement," *J. Comput. Sci. 2019, Vol. 15, Page 1795*, vol. 15, no. 12, pp. 1795–1808, Dec. 2019.
[29] L. Zhou and H. Fujita, "Posterior probability based ensemble strategy using optimizing decision directed acyclic graph for multi-class classification," *Inf. Sci. (Ny).*, vol. 400–401, pp. 142–156, 2017.
[30] L. Zhou, Q. Wang, and H. Fujita, "One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies," *Inf. Fusion*, vol. 36, pp. 80–89, 2017.
[31] J. Sun, H. Fujita, P. Chen, and H. Li, "Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble," *Knowledge-Based Syst.*, vol. 120, pp. 4–14, 2017.
[32] H.-F. Yu *et al.*, "Feature engineering and classifier ensemble for KDD cup 2010," *JMLR Work. Conf. Proc.*, pp. 1–12, 2010.
[33] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."* 2016.
[34] L. Breiman, "Random Forests," 2001.
[35] Y. Freund, R. E. Schapire, and others, "Experiments with a new boosting algorithm," in *icml*, 1996, vol. 96, pp. 148–156.
[36] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[37] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learn.*, vol. 59, no. 1–2, pp. 161–205, 2005.
[38] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps - review and good into in ML and NN approaches and basics + Python samples --Each topic has two parts: the first part will cover the theoretical concepts and the second part will cover practical impleme*, vol. 19, no. 2. 2017.
[39] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.
[40] S. Sun, Y. Wei, and S. Wang, "AdaBoost-LSTM Ensemble Learning for Financial Time Series Forecasting," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.
[41] M. S. Hegde, G. Krishna, and R. Srinath, "An Ensemble Stock Predictor and Recommender System," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 1981–1985.
[42] R. Dash, S. Samal, R. Dash, and R. Rautray, "An integrated TOPSIS crow search based classifier ensemble: In application to stock index price movement prediction," *Appl. Soft Comput. J.*, vol. 85, p. 105784, Dec. 2019.

[43] S. A. Gyamerah, P. Ngare, and D. Ikpe, “On Stock Market Movement Prediction Via Stacking Ensemble Learning Method,” in *CIFEr 2019 - IEEE Conference on Computational Intelligence for Financial Engineering and Economics*, 2019, pp. 1–8.

[44] K. S. Gan, K. O. Chin, P. Anthony, and S. V. Chang, “Homogeneous ensemble feedforward neural network in CIMB stock price forecasting,” in *Proceedings - 2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2018*, 2019, pp. 111–116.