International Journal of Advanced Technology and Engineering Exploration, Vol 8(83)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.874387

Research Article

Stock movement prediction using hybrid normalization technique and artificial neural network

Binita Kumari[1]* and Tripti Swarnkar[2]
Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Odisha, India[1]
Department of Computer Application, Siksha 'O' Anusandhan Deemed to be University, Odisha, India[2]

Received: 08-July-2021; Revised: 19-October-2021; Accepted: 22-October-2021

Abstract

Prediction of stock market indices has pinched considerable debate due to its brunt on economic development. Prediction of appropriate stock market indices is important in order to curtail the risk related with it in order to decide on effective investment schemes. Thus, selection of a proper forecasting model is highly appreciated. The objective of this paper is to efficiently normalize data in order to obtain accurate forecasting of stock movement and compare the results. A new technique called the hybrid normalization methodology for the efficient forecasting of stock movement has been implemented. This study discusses three normalization approaches along with our proposed normalization technique and their effect on the forecasting performance. In our work, we implemented Support Vector Machine (SVM), Artificial neural Network (ANN) and K-Nearest Neighbor (KNN) for stock trend forecasting as of their risk management capabilities. This article deals primarily with the normalization of input data for the estimation of stock movement. Simulation was performed on six stock indices (BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index) from different parts of the world market. The comparative study indicates that the hybrid normalization process is comparable to other normalization techniques. The results of the hybrid normalization process with ANN are found to be prominent as compared to other classifiers. The approximate accuracy obtained by using the hybrid normalization technique were 71%, 60% and 71% in the combination of KNN, SVM and ANN respectively.

Keywords

Artificial neural network, Hybrid normalization, KNN, SVM, ANN, Stock market indices.

1.Introduction

The forecasting of stock market progress is seen as a daunting work as the financial sector is very complex, dynamic and non-linear efficient framework [1]. Over the past decade, multi-fold surveys have been executed on mining of financial time series details, along with data mining procedures and conventional statistical methodologies. Numerous studies have focused on the usage of different classifiers [2–5] in the domain of the forecasting financial market. The stock forecasting method consists of a variety of steps, such as collecting information, developing organized information, standardizing information and classifying or predicting. A part of the stock market indices forecasting system delivers the criterion(s) that shows the outcome in a variety of units and scales to a standard and corresponding numeral scope. This method, called normalization, will have a profound impact on the result of the calculation [6].

As addressed by the researchers, in the conduct of a standardization method the data shall be further contacted by the noise company within the dataset. The effect of standardization in the identification of Ribonucleic Acid (RNA)-seq disease has been studied too. Accordingly, it is clear from the literature that the normalization method chosen to perform a data mining task can have an effect on the accuracy of the performance and it can get affected by the underlying outlier too. Therefore, proper preprocessing of the input features are needed. We also observed that the study of impact of normalization methods in financial market domain is unexplored. Thus, this article intends to test the

*Author for correspondence

predictive performance using different normalization techniques for six different stock indices. With an objective to improve the stock movement prediction accuracy, we propose a new hybrid normalization technique and compare the classification results of Support Vector Machine (SVM), Artificial neural Network (ANN) and K-Nearest Neighbor (KNN). Thus, the main contributions of our study are – (i) assess the effect of normalization technique on stock movement accuracy. (ii) propose a new hybrid normalization technique. (iii) evaluate the performance based on various metrics for six different stock indices. The rest of the paper is organised as follows: in section 2, we describe the related work. In section 3, the methods and materials used in our work are discussed. Section 4 provides detail about the proposed prediction model. Section 5 deals with the results and discussion followed by conclusion and future work in section 6.

2.Literature review

Looking at market dynamics and critical patterns, stock traders and anyone wanting to select the right stock or, conceivably, the favorite time to search for or sell stocks is very attractive [7]. Many full-scale financing related components, such as political events, organization methods, general monetary circumstances, etc., affect the stock costs [8]. Political agreements and administrative initiatives can have a significant impact on stocks. Soft computation approaches are widely used for financial market problems and are useful techniques for predicting non-linear behavior [9].

The SVM as well as ANN has been utilized for stock forecasting by various researchers. Be that as it may, even subsequent to building such a significant number of dynamic models. ANN incorporates couple of deterrents inside the learning method which impacts the result as appeared in [10]. As a consequence, a handful of researchers like advanced systems that rely on a powerful statistical basis, such as SVM [11]. The SVM technique, which is a supervised learning approach, uses classification and regression problems. SVM demonstrates high efficiency by reducing systemic risk as shown by the authors. Various researches use SVM to interpret time series information [12, 5]. The SVM is a popular machine learning system and has been used for non-linear predictions due to its eye-catching decisions and its high degree of execution in various issues.

In Vanstone and Finnie [5] tried to use the neural framework for estimating, but saw the SVM as superior to the multi-layer neural system framework for predicting monetary time. ANN has been used in many domains and one of them is in stock predictions. There are several studies which has been used ANN to model stock prices [13–17]. There are several researchers using neural networks to predict stock market volatility [18, 5]. The results obtained through the use of ANNs are superior to those obtained through the use of linear and logical regression models [19, 7]. Research using ANNs to forecast financial outcomes [20] yielded results with a 3% average failure rate. Even during the financial crisis, a multi-layer perceptron model [13] with macroeconomic indicators used to forecast Istanbul Stock Exchange produced a signal with a 73.7% correctness ratio, demonstrating the skill of ANNs in prediction. ANN outperform the adaptive exponential smoothing approach in predicting market movement, according to a report [20]. Many economists and financial analysts have advocated for the presence of financial market nonlinearity and uncertainty [21]. For stock index price movement prediction, a crow search-based weighted voting classifier ensemble with TOPSIS is suggested in [22]. Stock movements were predicted by looking at the causation between firms rather than the relevance between companies [23]. Using a bidirectional Gated Recurrent Unit (GRU) network based on Reinforcement Learning (RL) and an attention mechanism, a unique stock price movement prediction network has been developed in [24].

Normalization is a necessary part of every technique where protocols for managing information is implemented. In this regard, a study of the implementation of normalization procedures in various areas has been carried out. A massive amount of analysis of the information while not placing any burden on the quality of the results have been discussed along with the pre-requisite [25–27]. For the oversampling of unbalanced datasets, a pre-processing method known as Synthetic Minority Oversampling Technique- Edited Nearest Neighbor (SMOTE-ENN) has been used in [28]. As addressed by the researchers in [26, 27] the conduct of a standardization method shall be further contacted by the noise company within the data set. Authors attempted and valued the effect of standardization on the identification of RNA-seq disease [29]. Fourteen standard learning methods were assessed for the development of a powerful selection miniature in order to select the most appropriate standardization technique [30].

From literature, it is observed that the normalization method chosen to perform a data mining task can have an effect on the accuracy of the performance. We also observed that the study of impact of normalization methods in financial market domain is unexplored. Therefore, proper preprocessing of the input features is needed. In our paper, we're going to take a closer look at the value of stock forecast standardization. We also propose a hybrid normalization approach to improve the predictability accuracy.

3.Methods

3.1Datasets

To confirm the effect of the normalization of input data on forecasting results, five different Sensex datasets from different countries namely from India, United States of America, Hong Kong, China and Tokyo have been considered. BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index are selected as experimental data sets in this analysis. The analysis uses the details from 3/11/2015 to 27/11/2020. The accumulated data involves each day's high, open, low, adjusted close, volume, date and closing rate. They are used as informational indicators. Data was obtained from Yahoo finance. The purpose of our paper is to set the guidance of each day's record of the Sensex movement. A big issue with any dataset is that there are no up / down class marks in it. Therefore, as described in [31], we use the Δc (here ST) attribute that indicates the variation in closing price. Stock Trend (ST) has been used as an identifier for a class. '1' and '0' convey that the coming day index is increased or reduced than the current index. Miniatures for forecasting are produced and the output is used to determine the production.

The following rule is used to calculate these "0" or "1" values from previous closing prices:
If Closingprice(i)>Closingprice(i−1), then ST(i)= 1 i.e., increase
If Closingprice(i) < Closingprice(i−1), then ST(i) = 0 i.e., decrease
where Closingprice(i) = closing price of i^{th} row and ST(i)= index movement of i^{th} row

Data was obtained from Yahoo finance (https://in.finance.yahoo.com/). *Table 1* provides the details for the datasets used. Each stock index consists of 1235 number of total collected samples and seven features (each day's high, open, low, adjusted close, volume, date and closing rate).

Table 1 Dataset description

Dataset	Duration	No. of collected samples	No. of features
BSESN	30/11/2015 to 27/11/2020	1235	7
NIFTY50	30/11/2015 to 27/11/2020	1235	7
NASDAQ	30/11/2015 to 27/11/2020	1235	7
HANG SENG	30/11/2015 to 27/11/2020	1235	7
NIKKEI225	30/11/2015 to 27/11/2020	1235	7
SSE composite index	30/11/2015 to 27/11/2020	1235	7

3.2Normalization

Normalization is a pre data administration phase where we scale the input data to a small scope. Essentially, the standardization of data is needed to manage characteristics of different units and sizes with the ultimate aim of achieving better results. Except when pre-processed, valuables with different ranges or different precision obtain various driving values. More grounded drivers will blur important attributes. If a mining technique has a random sampling aspect, standardizing the specimen size will aid ensure that all the sources are evaluated uniformly. It also ensures the minimization in data-availability bias. Input data standardization performs an essential aspect in the process of predicting stock. Then again, on the off chance that the mining calculation has an irregular examining segment, at that point normalizing for test size may benefit guaranteeing that all origins are dealt with similarly, and that information accessibility predisposition (and its relating deception of the information universe) is decreased. Standardization of info information assumes a significant job in the stock expectation process. The following three normalization methods have been used to analyze their effect on stock estimates.

1. Min-max
2. Z score
3. Robust

3.2.1Min-max

A widely used method of data normalisation is min-max normalisation. The smallest value of each feature is converted to a 0, the maximal value is converted to a 1, and all other values are converted to a decimal amid 0 and 1 according to the formula given in *Table 2*. Min-max normalisation has one notable drawback: it struggles to deal with outliers.

3.2.2Z-score

Z-score normalisation is a data normalisation approach that avoids outliers. The mean and standard deviation of the data A are used to normalise the values in this procedure. The formula is shown in Equation 1.

$$A_i = \frac{A_i-\mu}{\sigma} \qquad (1)$$

Here A_i means the ith extent of the data set, μ is the mean and σ is the standard deviation.

Here, is the feature's mean value, and is the feature's standard deviation. A value will be normalised to 0 if it is precisely equal to the mean of all the values of the feature. It will be a negative number if it is lower than the mean, and a positive number if it is over the mean. The standard deviation of the original characteristic determines the magnitude of the negative and positive integers. The normalised values will be nearer to 0 if the unnormalized data have a significant standard deviation.

3.2.3Robust

In the presence of outliers, one method for standardising input variables is to disregard the outliers when computing the mean and standard deviation, then scale the variable using the derived values. The median (50th percentile), as well as the 25th and 75th percentiles, can be used to accomplish this. The median of each variable is then subtracted, and the Interquartile Range (IQR), which is the difference between the 75th and 25th percentiles, is divided. The formula for robust normalization technique is given in *Table 2*.

The formulas for the three normalization techniques utilized in our paper for comparison with our proposed technique are shown in *Table 2*.

Table 2 Normalization methods used

S. no.	Normalization techniques	Formula
1	Z-score	$A_i = \frac{A_i - \mu}{\sigma}$
2	Robust	$A_i = \frac{A_i - \text{median}}{75 \text{ percentile} - 25 \text{ percentile}}$
3	Min-max	$A_i = \frac{A_i - \min A_i}{\text{ma x}\, A_i - \min A_i}$

A_i = the ith extent of the dataset
μ = the feature's mean value
σ = the feature's standard deviation

3.3Technical indicators

For stock market indices, the input features usually used are- date, opening value, closing cost, maximum cost, lowest cost, adjusted close along with total volume. Numerous studies have exhibited that technical indicators are useful for the forecasting of stock [32, 6]. By applying the opening value equation, the lowest cost, the highest cost as well as the trading volume of the data, the calculation of the technical indicators can be done. Some of the technical indicators which are commonly adopted are shown in *Table 3*.

Table 3 Some widely used technical indicators

Technical indicators
Relative index
Stochastic slow
Stochastic indicator
Disparity 5
20-day bias
Rate of Change (ROC)
Momentum
Relative Strength Index (RSI)
Psychological line
Disparity 10
Moving Average Oscillators (MAO)
Commodity Channel Index (CCI)

3.4Support vector machines

As indicated by the researchers in [12, 25], SVMs are administered learning models which analyze data and classify patterns for the usage in classification and regression studies. It pursues by creating hyper planes in a multidimensional space that segregates specific class mark instances. It is capable of managing a variety of variables like continuous as well as categorical. SVM are efficient in high-dimensional spaces, provided that the total count of dimensions is greater as compared to the number of samples. They are versatile and memory proof. While employing SVM to predictions, the main point to be discussed is the kernel functions choice. Many researchers addressed the kernel features selection for financial prediction. In our study, we implemented the Gaussian Radial Basis Function (RBF) due to its capability of handling the nonlinear data.

3.5Artificial neural network

The structure and function of biological neural networks was used to design ANN architecture. ANN is made up of neurons that are organized in layers, much like neurons in the brain. The feed forward neural network is a common neural network that has three layers: an input layer that receives external data for pattern recognition, an output layer that solves the problem, and a hidden layer that connects the other layers. Acyclic arcs link neighboring neurons in the input as well as output layers. The ANN learns datasets using a training algorithm that adjusts neuron weights based on the error rate between goal and actual performance. In general, ANN learns datasets by using back propagation algorithm as a training algorithm. ANN's general structure is represented in *Figure 1*.

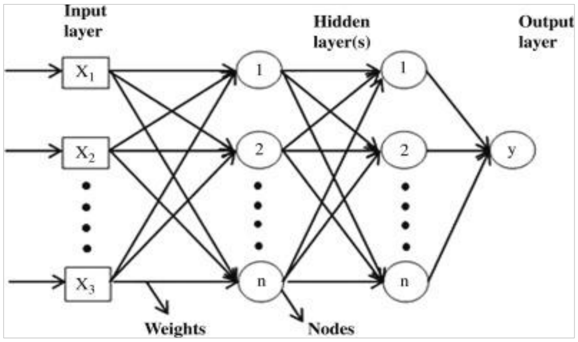

Figure 1 A general structure for ANN

3.6K-Nearest neighbor

KNN is a well-known machine-learning algorithm that has been applied to large number of data mining projects. The concept is that a large amount of training data is used, with each data point being defined by a collection of variables. Each point is conceptually sketched in a high-dimensional space, with each axis corresponding to a separate variable. On having a new (test) data point, we crave to identify the K nearest neighbors who are the utmost “similar” to it. The square root of the total count of points in the training data set (N) is commonly used as the value for K. (If N is 400, K equals 20.)

3.7Proposed methodology

This section describes the specifics of our proposed hybrid normalization methodology which integrates the various scores obtained from three different normalization methods.

A dataset may have outliers. Due to their impact on various normalization techniques, outlier presence does not guarantee balanced feature scales. We are familiar that normalization stands to be a scaling-up process to scale input info within a slightly defined scale. Therefore, when variables with differing ranges or differing precision have varying driving values, they can impact the ultimate result. In order to achieve a balanced feature scale, we propose the hybrid normalization technique which handles the outliers and also gives a balanced feature scale.

Hybrid Normalization is the technique of integrating the various scores obtained from different normalization methods. We have obtained the values by taking the log of average of the values obtained by each individual normalization technique. The value is calculated as below (Equation 2):

$$\mathrm{A}' = \mathrm{Log}((\textstyle\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{A_i}))/\mathrm{N}) \qquad (2)$$

Where A_i is value obtained from each normalization technique and N is the numbers in total for normalization techniques used.

For our proposed algorithm we consider a dataset. For all the features from 1 to n, minimum value and maximum value are found. Then we normalize all the values using min max formula as given in *Table 2*. The scaled values are stored in an array (M). Then again considering the original feature set, the values are scaled using z-score technique whose formula is given in *Table 2*. The scaled values obtained are stored in another array (Z). Once again, the scaling process is repeated for the original whole dataset using robust normalization technique considering the formula as shown in *Table 2* and the scaled values are stored in another array (R). In the next step, the average value for each corresponding array element is calculated as $[M(A_i)+Z(A_i)+R(A_i)]/3$. In order to dampen the effect of negative values, log is applied to the new scaled value. Thus, each element is now scaled as $\log\{[M(A_i)+Z(A_i)+R(A_i)]/3\}$. This final scaled value is taken as input to the model.

Algorithm1 depicts the detail steps for our proposed normalization technique.

Algorithm 1: Steps for hybrid normalization methodology

```
Input: The dataset A
Output: The normalized dataset A’
For features = 1 to n
    min $A_i$ = the minimum value
    $ma\,x\,A_i$ = the maximum value
    For each value $A_i$
        $M(A_i) = \frac{A_i - \min A_i}{ma\,x\,A_i - \min A_i}$
    end for
end for
$\mu$ = mean of dataset A
$\sigma$ = standard deviation of dataset A
For features = 1 to n
    For each value $A_i$
        $Z(A_i) = \frac{A_i - \mu}{\sigma}$
    End for
End for
For features = 1 to n
    median = median of the feature
    Interquartile range = 75 percentile − 25 percentile
```

```
$R(A_i) = \frac{A_i - \text{median}}{75\ \text{percentile} - 25\ \text{percentile}}$
End for
For features = 1 to n
        $A_i = \log\{[M(A_i)+Z(A_i)+R(A_i)]/3\}$
End for
```

Figure 2 shows a general structure of our approach. We first collected data for six different countries. Then we generated a synthesized dataset by including another set of 83 technical indicators and ST. We then normalized the datasets using our proposed hybrid normalization methodology. We tend to check the classification accuracy for different datasets using different classifiers namely KNN, SVM and ANN. We did the comparative analysis of our proposed method with the existing methods.

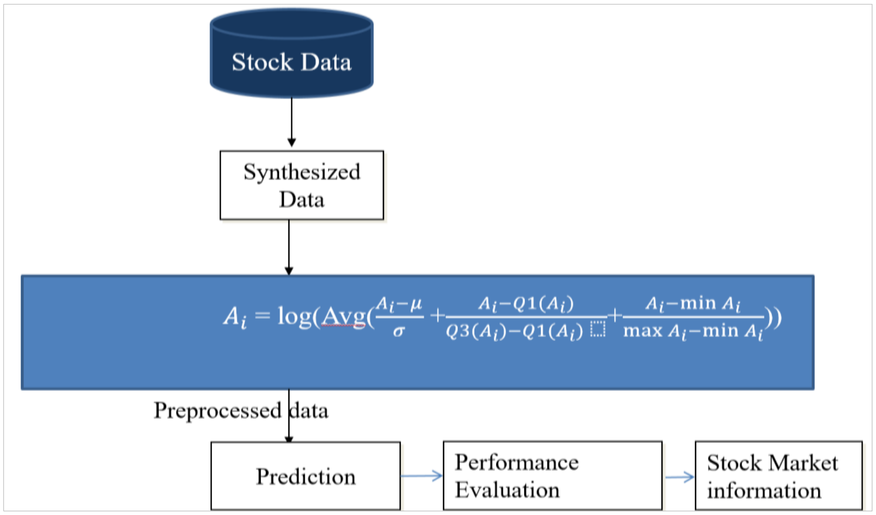
Stock Data
Synthesized Data
$A_i = \log(\text{Avg}(\frac{A_i-\mu}{\sigma} + \frac{A_i - Q1(A_i)}{Q3(A_i)-Q1(A_i)} + \frac{A_i - \min A_i}{\max A_i - \min A_i}))$
Preprocessed data
Prediction
Performance Evaluation
Stock Market information

Figure 2 A general structure for our proposed methodology

4.Results

The major goal of this research is to improve the stock movement prediction performance for different stock indices by using hybrid normalization method.

Data were obtained from Yahoo Finance for six Sensex data sets, which are BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index. In our next step, the task was to predict the heading of the daily stock value record as '1' or '0' depicting decrease or an increase in the closing cost. In extension to the opening cost, the closing cost, the lowest cost, the highest cost, the total trading volume, 83 suitable technical indicators were treated as initial feature pools. According to the researchers in [33, 34], the technical indicators are feasible means for presenting the true market condition in the financial time series prediction. We can be further instructive than using mere prices [34]. Thus 83 technical indicators which are used mostly all over the world have been generated using Python packages. Some of the technical indicators used in our analysis are given in *Table 2*. Thus, we generated a synthesized dataset consisting of initial stock values, 83 technical indicators and stock trend. The 83 normalised technical indicator values are passed as input and their appropriate class labels, 0 or 1, are provided as output to separate classifiers throughout the training phase. The ranges of the 83 technical indicators are different. We must first combine the values of these indicators into a single range before employing them in classifier models. The data is normalized using 4 different normalization techniques namely min-max, z-score, robust and our proposed hybrid normalization techniques. As recorded by the researchers in [9, 35] along with the literature review, we have got those the different normalization techniques as set out in *Table 2* (z score and min max) are commonly used in a variety of fields, such as medicine, industry, finance, business etc. On the basis of a literature survey, usage of 70% of the data points has been done as training details. The rest 30% of data points are used as test material. The classifiers are trained and validated using training data, and the model's performance is then assessed using test data. This research uses 3 different classifier models to forecast stock index movements: ANN, KNN and SVM. A general structure for our experimental setup has been shown in *Figure 3*.

Figure 4, *Figure 5*, *Figure 6*, *Figure 7*, *Figure 8* and *Figure 9* show the closing price trends for the collected datasets.

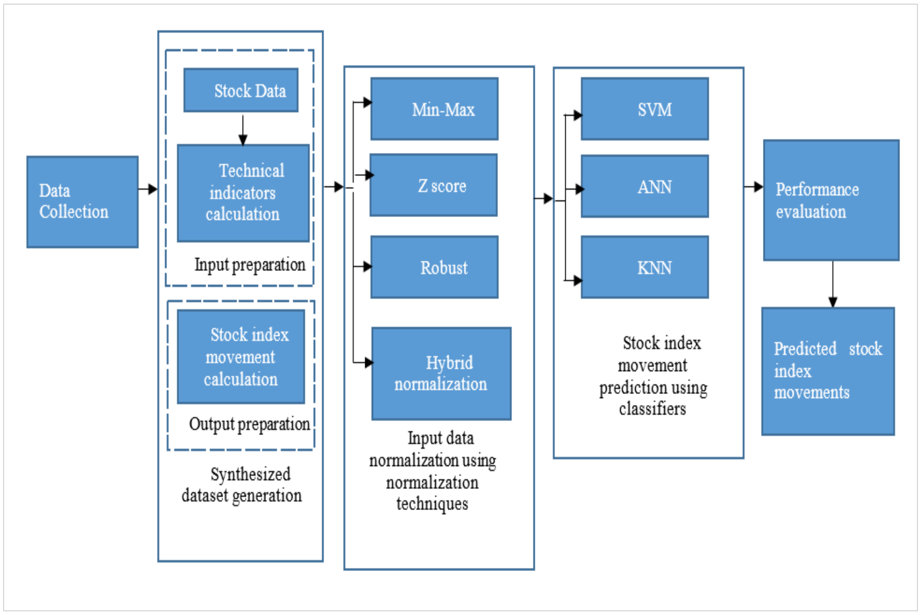
Data Collection
Stock Data
Technical indicators calculation
Input preparation
Stock index movement calculation
Output preparation
Synthesized dataset generation
Min-Max
Z score
Robust
Hybrid normalization
Input data normalization using normalization techniques
SVM
ANN
KNN
Stock index movement prediction using classifiers
Performance evaluation
Predicted stock index movements

Figure 3 A general structure for our experimental setup

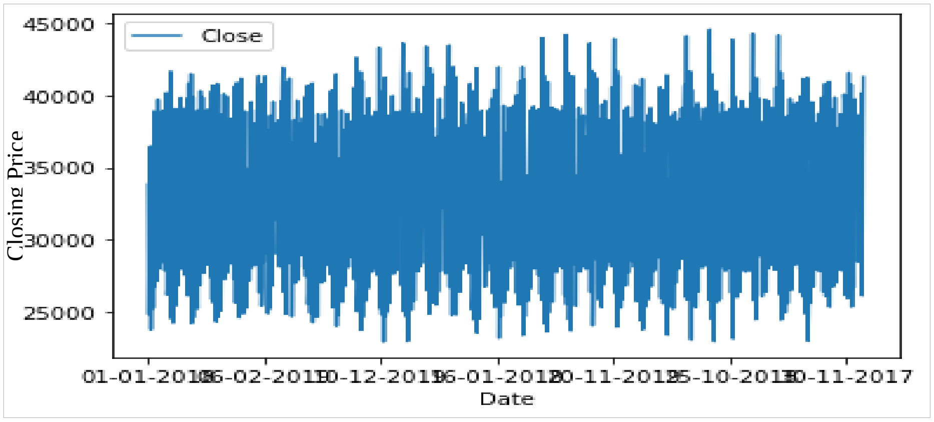

Figure 4 Closing price trend for BSESN

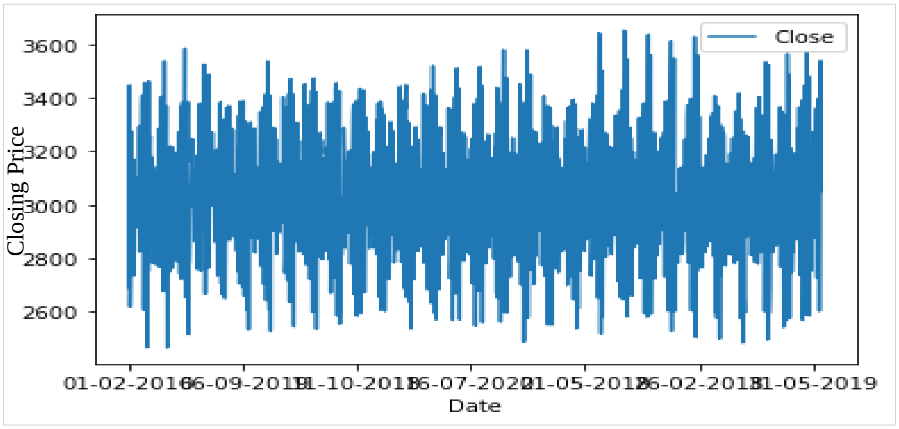

Figure 5 Closing price trend for SSE

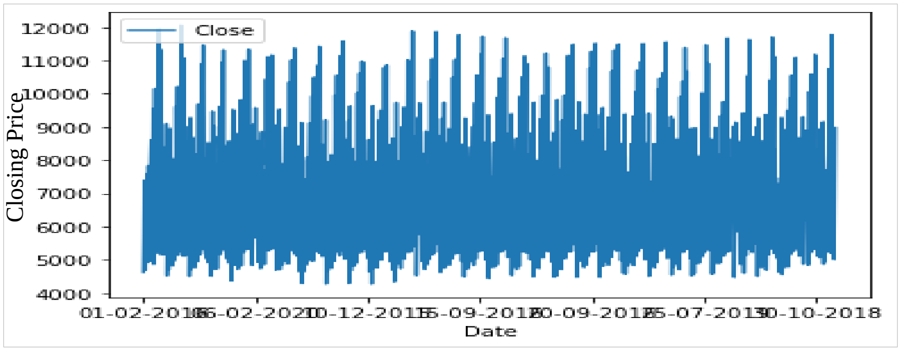

Figure 6 Closing price trend for NASDAQ

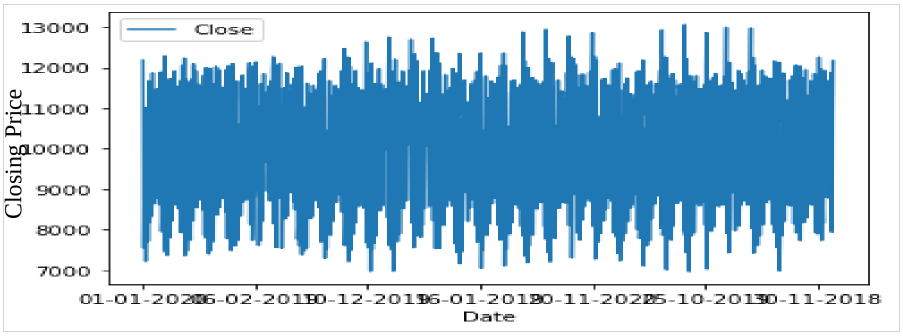

Figure 7 Closing price trend for NIFTY50

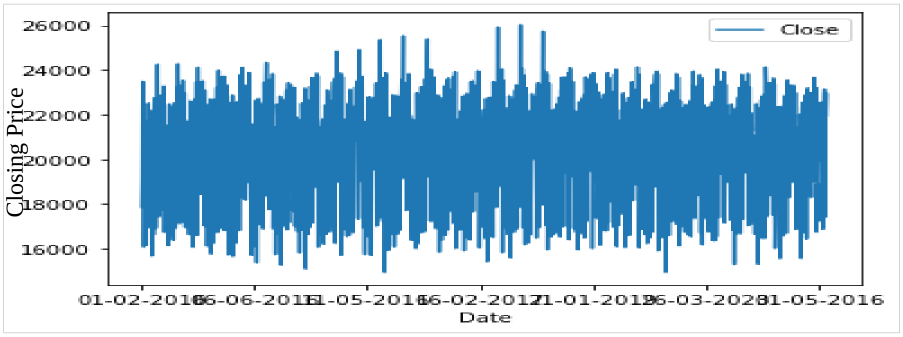

Figure 8 Closing price trend for NIKKEI225

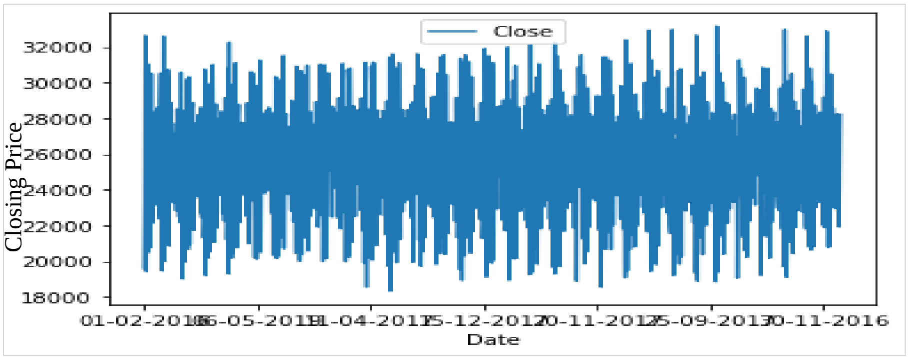

Figure 9 Closing price trend for HANGSENG

Table 4, *Table 5* and *Table 6* shows the value of different performance measures for KNN, SVM and ANN respectively for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index. The top-performing normalization technique is highlighted in the table for each criterion. *Table 4* to *Table 6* summarize the value of each performance evaluation measure for all the 3 classifiers using different normalization techniques for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index datasets respectively. For all criterion, the best doing normalization technique is highlighted and underlined in the table.

The efficiency results of KNN using Z-score, robust, min-max and our proposed hybrid normalization for the prediction of 2 class names, mainly up or down for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index are listed in *Table 4*. We have considered the value of K=5.

The efficiency results of SVM using Z-score, robust, min-max and our proposed hybrid normalization for prediction of 2 class names, mainly up or down for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index are listed in *Table 5*. We have used RBF kernel and grid search for parameter tuning.

The efficiency results of ANN using Z-score, robust, min-max and our proposed hybrid normalization for prediction of 2 class names, mainly up or down for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index are listed in *Table 6*. We have used 100 epochs.

Table 4 Accuracy results for different normalization techniques and Hybrid Normalization technique using KNN

Dataset	**Normalization technique**	**KNN**			
		Accuracy	**Precision**	**Recall**	**F1-score**
BSESN	MM	0.66	0.68	0.65	0.66
	ZS	0.69	0.72	0.65	0.68
	RS	0.70	**0.82**	0.65	0.72
	HN	**0.71**	**0.82**	**0.71**	**0.76**
NASDAQ	MM	0.69	0.72	0.65	0.68
	ZS	0.69	0.72	0.65	0.68
	RS	**0.71**	**0.76**	**0.70**	**0.72**
	HN	**0.71**	0.74	0.67	0.70
NIFTY50	MM	0.69	0.72	0.65	0.68
	ZS	0.71	0.76	0.70	0.72
	RS	0.70	0.74	0.61	0.66
	HN	**0.72**	**0.79**	**0.75**	**0.76**
NIKKEI25	MM	0.71	0.76	0.70	0.72
	ZS	0.71	0.76	0.70	0.72
	RS	0.70	0.74	0.68	0.70
	HN	**0.73**	**0.80**	**0.71**	**0.75**
HANGSENG	MM	0.71	0.76	0.70	0.72
	ZS	0.70	0.69	0.48	0.56
	RS	0.68	0.70	0.67	0.68
	HN	**0.76**	**0.78**	**0.76**	**0.76**

SSE composite index	MM	**0.70**	**0.74**	0.61	0.66
	ZS	0.69	0.72	0.65	0.68
	RS	0.68	0.70	0.67	0.68
	HN	**0.70**	**0.74**	**0.72**	**0.72**

Table 5 Accuracy results for different normalization techniques and Hybrid Normalization technique using SVM

Dataset	**Normalization Technique**	**SVM**			
		Accuracy	**Precision**	**Recall**	**F1-score**
BSESN	MM	0.61	0.62	0.59	0.60
	ZS	0.66	0.68	0.65	0.66
	RS	0.59	0.61	0.60	0.60
	HN	**0.67**	**0.70**	**0.69**	**0.69**
NASDAQ	MM	0.59	0.61	0.60	0.60
	ZS	0.62	0.58	0.65	0.61
	RS	0.63	0.68	0.50	0.57
	HN	**0.65**	**0.71**	**0.69**	**0.69**
NIFTY50	MM	0.62	0.68	0.65	0.66
	ZS	0.65	0.52	0.59	0.55
	RS	0.63	0.68	0.60	0.63
	HN	**0.66**	**0.74**	**0.72**	**0.72**
NIKKEI25	MM	0.59	0.61	0.60	0.60
	ZS	0.64	0.58	0.60	0.58
	RS	0.64	0.60	0.51	0.55
	HN	**0.65**	**0.71**	**0.68**	**0.69**
HANGSENG	MM	0.58	0.61	0.56	0.58
	ZS	0.63	0.68	0.60	0.63
	RS	0.63	0.58	0.50	0.53
	HN	**0.64**	**0.69**	**0.68**	**0.68**
SSE Composite Index	MM	0.58	0.61	0.56	0.58
	ZS	0.62	0.51	0.50	0.50
	RS	0.61	0.60	**0.64**	0.61
	HN	**0.63**	**0.65**	0.63	**0.63**

Table 6 Accuracy results for different normalization techniques and Hybrid Normalization technique using ANN

Dataset	**Normalization Technique**	**ANN**			
		Accuracy	**Precision**	**Recall**	**F1-score**
BSESN	MM	0.69	0.72	0.65	0.68
	ZS	0.70	0.68	0.61	0.64
	RS	**0.85**	0.82	0.74	0.77
	HN	**0.85**	**0.88**	**0.87**	**0.87**
NASDAQ	MM	0.69	0.72	0.65	0.68
	ZS	0.82	0.80	**0.89**	0.84
	RS	0.81	**0.88**	0.80	0.83
	HN	**0.84**	**0.88**	0.86	**0.86**
NIFTY50	MM	0.71	0.76	0.69	0.72
	ZS	0.81	0.85	0.71	0.77
	RS	0.83	0.89	0.78	0.83
	HN	**0.85**	**0.89**	**0.86**	**0.87**
NIKKEI25	MM	0.67	0.71	0.65	0.67
	ZS	0.83	0.80	**0.88**	0.83
	RS	**0.89**	0.86	0.78	0.81
	HN	0.87	**0.89**	0.87	**0.87**
HANGSENG	MM	0.68	0.70	0.67	0.68
	ZS	0.86	0.83	0.85	0.83
	RS	0.83	0.81	0.78	0.79
	HN	**0.87**	**0.88**	**0.87**	**0.87**
SSE Composite	MM	0.64	0.68	0.60	0.63

Index	ZS	0.86	0.83	0.75	0.78
	RS	0.85	0.82	0.74	0.77
	HN	**0.87**	**0.87**	**0.86**	**0.86**

Figures 10, *11* and *12* demonstrate and contrast the findings achieved for the various methods in *Table 4, 5* and *6*. *Figures 10*, *11* and *12* show the graphical representation of the accuracy results for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index for different normalization techniques and hybrid normalization technique for stock movement prediction using KNN, SVM and ANN.

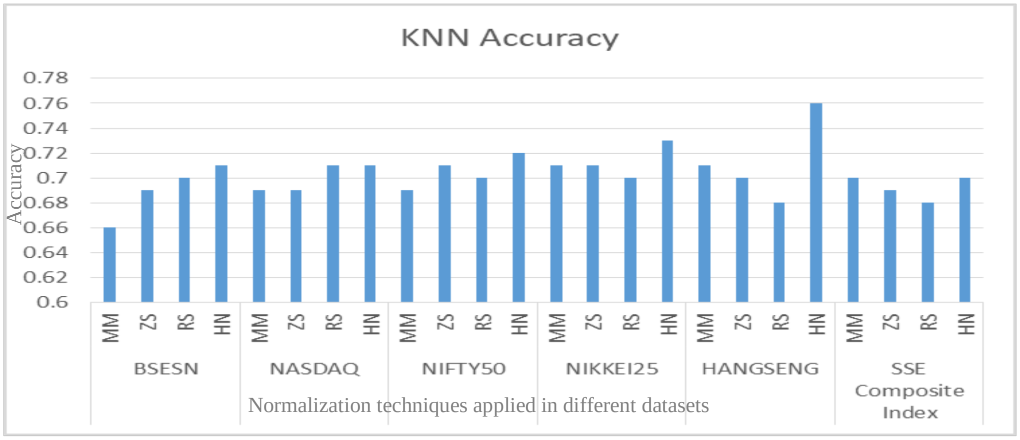

Figure 10 Accuracy results for different normalization techniques and Hybrid Normalization technique using KNN

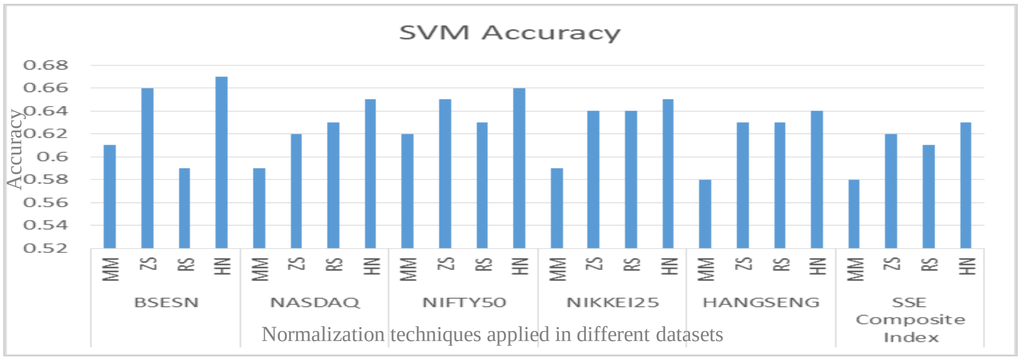

Figure 11 Accuracy results for different normalization techniques and Hybrid Normalization technique using SVM

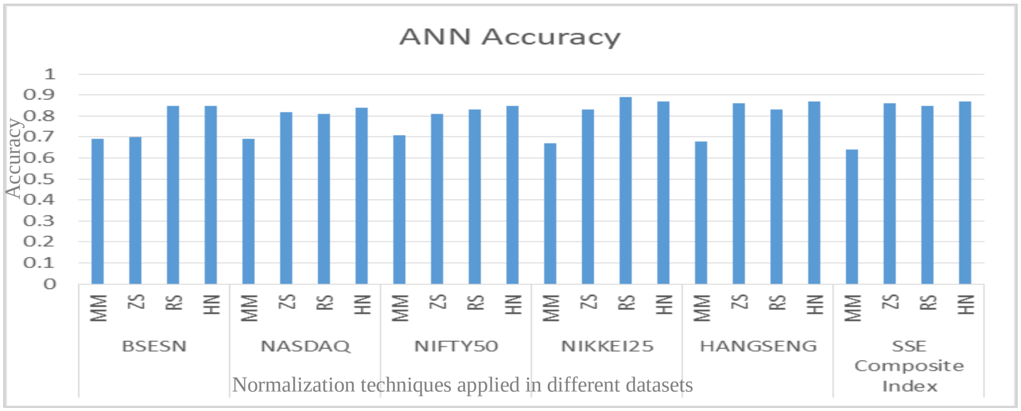

Figure 12 Accuracy results for different normalization techniques and Hybrid Normalization technique using ANN

5.Discussion

We first collected data for six different countries. Then we generated a synthesized dataset by including another set of 83 technical indicators and ST. We then normalized the datasets with 4 different normalization techniques including our proposed methodology as

1. Min-max
2. Z score
3. Robust
4. Hybrid Normalization

We tend to check the classification accuracy for different datasets using different classifiers namely KNN, SVM and ANN. We did the comparative analysis of our proposed method with the existing normalization methods.

We effectively track the forecasting efficiency and effect of normalization techniques between Z score, Robust and Min-max along with SVM, ANN and KNN considering the same training data set as well as testing data set for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index respectively. The assessment of the miniature was carried out using four performance evaluation criteria, which are calculated using the following formulas.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

$$\text{F1 Score} = \frac{\text{2xPrecisionxRecall}}{\text{Precision+Recall}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

From *Table 4* to *6*, we observe that the predictive accuracy for different stock indices varies when min max, z score and robust normalization techniques are applied. This variation is due to the inline feature of the dataset itself. A dataset may have outliers. Due to their impact on various normalization techniques, outlier presence does not guarantee balanced feature scales. We are familiar that normalization stands to be a scaling-up process to scale input info within a slightly defined scale. Therefore, when variables with differing ranges or differing precision have varying driving values, they can impact the ultimate result. The value ranges of the input dataset including stock data and 83 technical indicators are different. We must first combine the values of these input values into a single range before employing them in classifier models. Thus, it can be concluded that the accuracy of the prediction depends on the normalization method used for the input data since all the normalization techniques do not scale the features into same values. Therefore, when variables with differing ranges or differing precision have varying driving values, they can impact the ultimate result. Thus, the final prediction output of using the same normalization technique to various types of data sets together with the same classification methodology could be diverse. Likewise, the utilization of various forms of normalization strategies to an individual data set can often have varying results due to the properties of the bottom-line data set. Thus, to overcome the difference in effect due to different normalization technique on the same dataset, we have proposed the Hybrid Normalization method where we considered the feature handling capability characteristic of different normalization techniques together for a dataset.

We also observe that the predictive accuracy of hybrid normalization is better or same as compared to min max, z score and robust normalization techniques for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index. In order to achieve a balanced feature scale, we use the Hybrid Normalization technique which handles the outliers and also gives a balanced feature scale Same input information may scale to different values for different normalization technique due to the underlying characteristics of the dataset and the normalization technique itself. Thus, by applying Hybrid Normalization, we are considering the diversity in the scaled input information. By taking the average of the scaled values for the different normalization techniques, we are generating the final scaled value which is considered as the input to the model. Thus, we are trying to reduce the diversified range of scaled outputs which have been obtained using different normalization methods. As a result, there is an increase in the prediction accuracy.

It is also observed that the predictive output of SVM, ANN and KNN differs when specific input data normalization methods are enforced for a dataset. This variation is due to the inline feature of the dataset itself and the various normalization techniques used to scale the input data.

Additionally, we also observe from *Tables 4*, *5* and *6* that the classification accuracy of ANN is better as compared to SVM and KNN. We also observe that

ANN gives better accuracy for all the datasets when the Hybridized Normalization is implemented for the normalization of the input features. This is because of the better risk handling capability of ANN as compared to SVM and KNN. Thus, we may conclude that the predictive accuracy of ANN is better for Hybrid Normalization as compared to other normalization methods.

Figures 10, 11 and *12* show the graphical representation of the accuracy results for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index for different normalization techniques and Hybrid Normalization technique for stock movement prediction using KNN, SVM and ANN. It is further noticed that the predictive performance of SVM, ANN and KNN established using Hybrid Normalization is better in comparison to Min-max, Z score and Robust normalization techniques. As seen in *Table 4, 5* and *6*, it can be argued that the utilization of various types of standardization methods to an individual data set may have varying results due to the properties of the underlying data set. We may therefore assume that the efficiency of the prediction depends on the normalization method used for input data, together with more criterions. But Hybrid Normalization takes into account the characteristics of the various normalization methods and thus gives a better scaled input value. Thus, our study initially shows that applying the same normalization technique to various datasets can provide divergent performance rates. Likewise, the application of various forms of normalization strategies to a lone dataset can often give varying results due to the properties of the bottom-line dataset. Therefore, the precision of the prediction outcomes varies for different standardization methods. Divergent normalization methods can offer varying prediction efficiency outcomes for the same machine learning technique and data set. Therefore, the effects of the fault frequency can also vary for varying datasets. A potential solution to this problem is a Hybrid Normalization process.

5.1Limitations

The current study's design is limited, as is the case with the majority of investigations. This study has two key limitations that could be addressed in future research. First is the limited access to data. Second is the lack of previous research studies on the normalization effect. We had collected data from yahoo finance. We were able to collect the historical data for six stock indices only. As a result, we could not generalize our findings. This could be a possible direction for future scope. Thus, our proposed methodology can be tested with various other datasets. From literature, it is observed that the normalization method chosen to perform a data mining task can have an effect on the accuracy of the performance. But we observed that the study of the impact of normalization methods is too limited. We also observed that the study of impact of normalization methods in financial market domain is unexplored. This could be a possible direction for future scope. Thus, our proposed methodology can be tested with various other domains. A complete list of abbreviations is shown in *Appendix I*.

6.Conclusion and future work

The normalization method which is utilized for the input data normalization significantly impacts the accuracy result of the machine learning processes. Applying totally different strategies for normalizing the input data files for SVM running, offers diverse estimations of precision measures. The selection of standardization technique ought to depend on the features being anticipated, and on the rule of loss minimization. Each application of classifier needs creating choices regarding parameter tuning etc. It has occurred from our study that the choice of technique of normalization of input data will considerably influence the accuracy results given by classifiers. When varying types of normalization methods are applied on same data-set using the same machine learning method, the outcome may vary. A possible solution to this problem is Hybrid Normalization. We applied the Hybrid Normalization to for BSESN, NIFTY50, NASDAQ, HANG SENG, NIKKEI225 and SSE composite index. The results show that Hybrid Normalization technique makes a considerable impact on the classifier accuracy. From our study we observe that different normalization technique would work differently for different datasets and domains or in all conditions. But in most of the cases hybrid normalization is easy to use and gives enhanced result in comparison to other normalization techniques. In our study we have considered the six stock indexes from various countries like India, China, Tokyo, Hong Kong and Unites States of America from around the globe. We conclude from observations that ANN gives better accuracy results when implemented with hybridized normalization technique. For future work, we suggest to explore other classifiers for their accuracy result and behavior. The proposed normalization technique can also be explored for other domains and more datasets.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Abu-mostafa YS, Atiya AF. Introduction to financial forecasting. Applied Intelligence. 1996; 6(3):205-13.

[2] Huang Z, Chen H, Hsu CJ, Chen WH, Wu S. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems. 2004; 37(4):543-58.

[3] Fernández-lozano C, Canto C, Gestal M, Andrade-garda JM, Rabuñal JR, Dorado J, et al. Hybrid model based on genetic algorithms and SVM applied to variable selection within fruit juice classification. The Scientific World Journal. 2013.

[4] Yeh TL. Capital structure and cost efficiency in the Taiwanese banking industry. The Service Industries Journal. 2011; 31(2):237-49.

[5] Vanstone B, Finnie G. An empirical methodology for developing stockmarket trading systems using artificial neural networks. Expert Systems with Applications. 2009; 36(3):6668-80.

[6] Jain S, Shukla S, Wadhvani R. Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications. 2018; 106:252-62.

[7] Khan ZH, Alin TS, Hussain MA. Price prediction of share market using artificial neural network. International Journal of Computer Applications. 2011; 22(2):42-7.

[8] Xie B, Passonneau R, Wu L, Creamer GG. Semantic frames to predict stock price movement. In proceedings of the annual meeting of the association for computational linguistics 2013 (pp. 873-83).

[9] Chen WH, Shih JY, Wu S. Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets. International Journal of Electronic Finance. 2006; 1(1):49-67.

[10] García S, Fernández A, Luengo J, Herrera F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Information Sciences. 2010; 180(10):2044-64.

[11] Hsu MW, Lessmann S, Sung MC, Ma T, Johnson JE. Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Systems with Applications. 2016; 61:215-34.

[12] Żbikowski K. Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. Expert Systems with Applications. 2015; 42(4):1797-805.

[13] Tay FE, Cao L. Application of support vector machines in financial time series forecasting. OMEGA. 2001; 29(4):309-17.

[14] Sáez JA, Galar M, Luengo J, Herrera F. Tackling the problem of classification with noisy data using multiple classifier systems: analysis of the performance and robustness. Information Sciences. 2013; 247:1-20.

[15] Leigh W, Modani N, Hightower R. A computational implementation of stock charting: abrupt volume increase as signal for movement in New York stock exchange composite index. Decision Support Systems. 2004; 37(4):515-30.

[16] Racine J. On the nonlinear predictability of stock returns using financial and economic variables. Journal of Business & Economic Statistics. 2001; 19(3):380-2.

[17] Mitra SK. Optimal combination of trading rules using neural networks. International Business Research. 2009; 2(1):86-99.

[18] Kumari B, Swarnkar T. Importance of data standardization methods on stock indices prediction accuracy. Advanced Computing and Intelligent Engineering. Springer, Berlin. 2020.

[19] Kim KJ. Financial time series forecasting using support vector machines. Neurocomputing. 2003; 55(1-2):307-19.

[20] Huang W, Nakamori Y, Wang SY. Forecasting stock market movement direction with support vector machine. Computers & Operations Research. 2005; 32(10):2513-22.

[21] Barak S, Arjmand A, Ortobelli S. Fusion of multiple diverse predictors in stock market. Information Fusion. 2017; 36:90-102.

[22] Dash R, Samal S, Dash R, Rautray R. An integrated TOPSIS crow search based classifier ensemble: in application to stock index price movement prediction. Applied Soft Computing. 2019.

[23] Nam K, Seong N. Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. Decision Support Systems. 2019; 117:100-12.

[24] Xu H, Chai L, Luo Z, Li S. Stock movement prediction via gated recurrent unit network based on reinforcement learning with incorporated attention mechanisms. Neurocomputing. 2021; 467(7):214-28.

[25] Neely CJ, Rapach DE, Tu J, Zhou G. Forecasting the equity risk premium: the role of technical indicators. Management Science. 2014; 60(7):1772-91.

[26] Chen Y, Hao Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications. 2017; 80:340-55.

[27] Jha P, Mohan N, Laha AK, Dutta G. Artificial neural network models for forcasting stock price index in Bombay stock exchange. IIMA Institutional Repository. 2010.

[28] Altay E, Satman MH. Stock market forecasting: artificial neural network and linear regression comparison in an emerging market. Journal of Financial Management & Analysis. 2005; 18(2):18-33.

[29] Han H, Men K. How does normalization impact RNA-seq disease diagnosis?. Journal of Biomedical Informatics. 2018; 85:80-92.

[30] Garcia LP, De CAC, Lorena AC. Effect of label noise in the complexity of classification problems. Neurocomputing. 2015; 160:108-19.

[31] Sahin U, Ozbayoglu AM. TN-RSI: trend-normalized RSI indicator for stock trading systems with evolutionary computation. Procedia Computer Science. 2014; 36:240-5.

[32] Cao LJ, Tay FE. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on Neural Networks. 2003; 14(6):1506-18.

[33] Senol D, Ozturan M. Stock price direction prediction using artificial neural network approach: the case of Turkey. Journal of Artificial Intelligence. 2009; 1(2):70-7.

[34] Kaastra I, Boyd M. Designing a neural network for forecasting financial and economic time series. Neurocomputing. 1996; 10(3):215-36.

[35] Vanstone BJ, Finnie GR. Combining technical analysis and neural networks in the Australian stockmarket. In international conference on artificial intelligence and soft computing 2006 (pp. 125-30).

Binita Kumari is an Assistant Professor in the Department of Computer Science and Engineering, Faculty of Engineering and Technology, SOA University, Bhubaneswar, India. She received her BTech from the Utkal University, MTech from the SOA (Deemed to be University), India, in 2001 and 2011, respectively. She is currently a Doctorate student at the Chair for Computer Science and Engineering at the SOA (Deemed to be University), India. Her current research interests include Deep Learning, Data Mining and Machine Learning.
Email: binitarath@hotmail.com

Tripti Swarnkar is a Professor and a HOD in the Department of Computer Application, Faculty of Engineering and Technology, SOA (Deemed to be Universit)y, Bhubaneswar, India. She received her PhD in Engineering from the IIT Kharagpur, India in 2015. She is guiding PhD’s in the field of Machine Learning, Deep Learning, Biomedical Engineering and Bioinformatics. She has more than 25 Master theses in the area of Machine Learning, Biomedical Applications and Bioinformatics. She has published many research papers in international journals and conference proceedings. She is an author of two books and few invited book chapters, published by leading international publishers.
Email: triptiswarnakar@soa.ac.in

Appendix I

S. No.	Abbreviation	Description
1	ANN	Artificial Neural Network
2	BSESN	Bombay Stock Exchange Sensitive Index
3	CCI	Commodity Channel Index
4	IQR	Inter Quartile Range
5	KNN	K Nearest Neighbour
6	MAO	Moving Average Oscillators
7	NASDAQ	National Association of Securities Dealers Automated Quotations
8	NIFTY50	National Stock Exchange Fifty
9	RBF	Radial Basis Function
10	ROC	Rate of Change
11	RSI	Relative Strength Index
12	SMOTE-ENN	Synthetic Minority Oversampling Technique- Edited Nearest Neighbor
13	SSE composite index	Shanghai Stock Exchange Composite Index
14	SVM	Support Vector Machine