Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Hybrid Model to Predict Stock Movements Using Novel Features and Fully Modified HP Filter

SIDRA IJAZ[1], KHALID IQBAL[1], SADAF YASMIN[1], AHTHASHAM SAJID[2], SYED ATTIQUE SHAH[2], AKHTAR JAMIL[3], NIGHAT USMAN[4], AND DIRK DRAHEIM[5]

[1]COMSATS University Islamabad (Attock Campus), Attock, 43600, Pakistan
[2]Department of Computer Science, Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Quetta, 87300, Pakistan
[3]Department of Computer Engineering, Istanbul Sabahattin Zaim University, 34303 Istanbul, Turkey
[4]Department of Computer Science, Bahria University (Lahore Campus), Lahore, 54600, Pakistan
[5]Information Systems Group, Tallinn University of Technology, Akadeemia Tee 15a, 12618 Tallinn, Estonia

Corresponding author: Khalid Iqbal (e-mail: khalidiqbal@cuiatk.edu.pk).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT In the financial world, predicting stock market prices with high accuracy is considered a challenging task. Forecasting market prices is an interesting area for investors and traders. Successful predictions lead to high financial revenues and prevent investors from market risks. In addition to the historical stock data, the social media also facilitates forecasting of stock price movement and has become a vital platform of information sharing. Therefore, combining these two vital information can complement in obtaining higher accuracy for prediction. Recently, machine learning-based methods have proven to be more effective for forecasting stock market prices, yet there is still room for improving their accuracy. In this paper, a novel hybrid model has been proposed for stock prediction that focuses on improving the prediction accuracy. The proposed method consists of three main components: 1) filtering technique 2) feature extraction 3) machine learning-based prediction. The filtering technique included a fully modified Hodrick–Prescott filter that helped smooth the historical data. Novel feature extraction step included both technical (for financial data) and content-based (for tweets data) features which were derived from different data sources. Finally, we investigated both deep learning (DL) and traditional machine learning (ML) approaches for prediction. For ML, we selected support vector regression (SVR), Auto-Regressive Integrated Moving Averages (Auto ARIMA), Random Forest (RF) while as part of DL we employed Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). We sensibly choose twitter as an online social network to analyze news by extracting sentiment features and historical data from Yahoo finance. Twitter content and financial time series makes the hybrid model a powerful forecasting tool. Several experiments were performed to evaluate the performance of our proposed method. The main aim of this evaluation effort is to validate the improvement in the accuracy of forecasting the stock market. The obtained results revealed that the proposed provides competitive results with state-of-the-art approaches.

INDEX TERMS Stock Prediction, neural networks, Support vector machine, Fully Modified Hyper-Prescott Filter, Random forest, Auto ARIMA, LSTM, GRU

I. INTRODUCTION

IN 24-hours of trading in a stock market, the closing price for every actual stock is measured as the final price at which it is traded during regular market hours on a given day. The closing price is taken as the most accurate estimation of the stock by stakeholders until resumption of next day trading. Usually, the close price is foreseen and convenient for traders, investors and market to estimate fluctuation in stock market prices and therefore, it is considered as a standard benchmark to track the stock's performance on a daily basis. Predicting market closing price is the most significant applications of the stock market. It helps investors

in recognizing the current situation of the stock market and reveal the upcoming stock market behaviour and help buyers and sellers to understand when and which stock should be purchased for their investment's growth. By utilizing prediction applications companies can save millions of dollars and can prevent their selves from big losses. Accurate predictions not only describe the current stock market situation but also keep trader's investors and stakeholders, alert for future opportunities and threats on the bases of ongoing trends.

Many techniques can be applied to predict the correct stock closing price by using various data sets such as Social media and Macroeconomics. Social media platform widens up the scope of social interaction with peoples where they exchange multimedia information initiate online discussion on any social topics. Moreover, various news agencies report the stock market prices on a timely basis on their social media accounts, forming an effective data source. Macroeconomic contain several components that correspond to short-term irregularities seasonal variations, long term trend movements, and medium-term business cycles. Mostly financial time series analysis deal with medium-term business cycles, and a long-term trend's movement. The essential movements are mostly hidden in the original macroeconomics. Effective data set selection is an important aspect for the prediction system as irregular and seasonal fluctuations in the data can largely affect the accuracy of the results [1], [2]. Irregularities that affect the information in the record are often more difficult to read directly from the original macroeconomics series. Accurate consideration analysis of data and noise filtering is an important factor to improve the accuracy of stock price forecasting [3].

Accurate machine learning forecasting leads towards market revenue [4], [5]. In [6], Kim forecasts future direction of stock price indexes by utilizing SVM forecast model. In [7] authors suggest that SVM outperforms various neural network methods in financial time series forecasting. In recent times, researchers apply different hybrid techniques to predict stock closing price [8], [9]. Recently, Meryem Ouahilal *et al.* [10] developed a machine learning and filtering techniques based hybrid approach integrating Support Vector Regression and Hodrick–Prescott filter to enhance the prediction of the stock price.

Artificial neural network (ANN) has widely been used for stock prediction due to its ability to approximate nonlinear relationship between data. In [45] authors used ANN for prediction of the KSE-100 Index for the data approximately over 3 years. The authors in [46] employed artificial neural networks for stock price estimation. Similarly, Khalid Alkhatib and colleagues used the k-NN algorithm to estimate the stock prices of six companies in the Jordan market [47]. Auto ARIMA is a very popular statistical method, which is widely used for stock price prediction. For instance, S. Wadi, M. Almasarweh and A. Alsaraireh applied the auto ARIMA model to the closing prices obtained from the Amman Stock Exchange (ASE) from 2010 to 2018 [48]. The results proved the efficiency of the proposed method for stock prediction. In [49], the authors also employed the autoregressive integrated moving average (ARIMA) model for accurate prediction of stock data on the Amman stock exchange in the Jordan market.

Recently, the deep learning-based methods have been proposed for stock prediction due to its high accuracy. Particular, the Recurrent Neural Networks (RNNs), are specially designed for time series data analysis. For example, Xiao et al. used a deep convolutional neural network for event-based stock market prediction [50]. Bengio et al. used the LSTM model for stock price estimation [51]. For stock forecasts, Chen, Zhou and Dai (2015) [52] exploited a LSTM stock market data analysis in China. Likewise, Wei Bao and authors have created a three-stage deep learning framework by combining LSTM and autoencoders (SAEs) models for stock price estimation [53]. They created the LSTM model for the next day's closing price estimate.

The motivation for this paper is to address the problem of predicting stock closing price with a hybrid approach to further improve the accuracy of the results. We proposed a novel hybrid model which combines both technical features and content features. The model follows the architecture proposed by [1], however, additional novel features were included in the model to further improve the accuracy of the classifiers.. The technical features were obtained from the stock historical data while the content features were obtained from Tweets data. In addition, noise filtering approaches were also used to reduce the impact of noise on the prediction accuracy. Finally, we employed machine learning techniques for stock prediction. We compared three traditional machine learning techniques (SVR, RF, and Auto ARIMA) and two deep learning-based approaches (LSTM and GRU). The obtained results indicated the effectiveness of the proposed hybrid model for prediction of stock close price.

Initially, we examine the outcomes and effectiveness of parameters with the aim to discover an optimal technique for enhancing the accuracy of the closing price. Several studies [11], [12] highlight the presence of high noise issue in financial data. Thereby, forecasting directly with (SVR) perhaps delicate to noise and leads towards overfitting. To overcome these limitations novel hybrid model is proposed in this paper. We presented our model for better prediction of financial market price via learning time series and textual data. After getting smoothing and noise-free time series, for our proposed solution, we have used predicting algorithm on growth component g_t of market data and content features to predict the stock closing price. We evaluate the performance of various deep learning and traditional machine learning approaches for prediction by applying them one by one in our hybrid model. Aim of using this approach is to improve the prediction of existing approaches such as SVR and RNN for increasing the accuracy of predicting closing price. Fully Modified HP filter helps in removing noise and increase the smoothness of our financial data set. The proposed framework of our model is displayed in Figure 1.

The major contributions of our work are listed below:

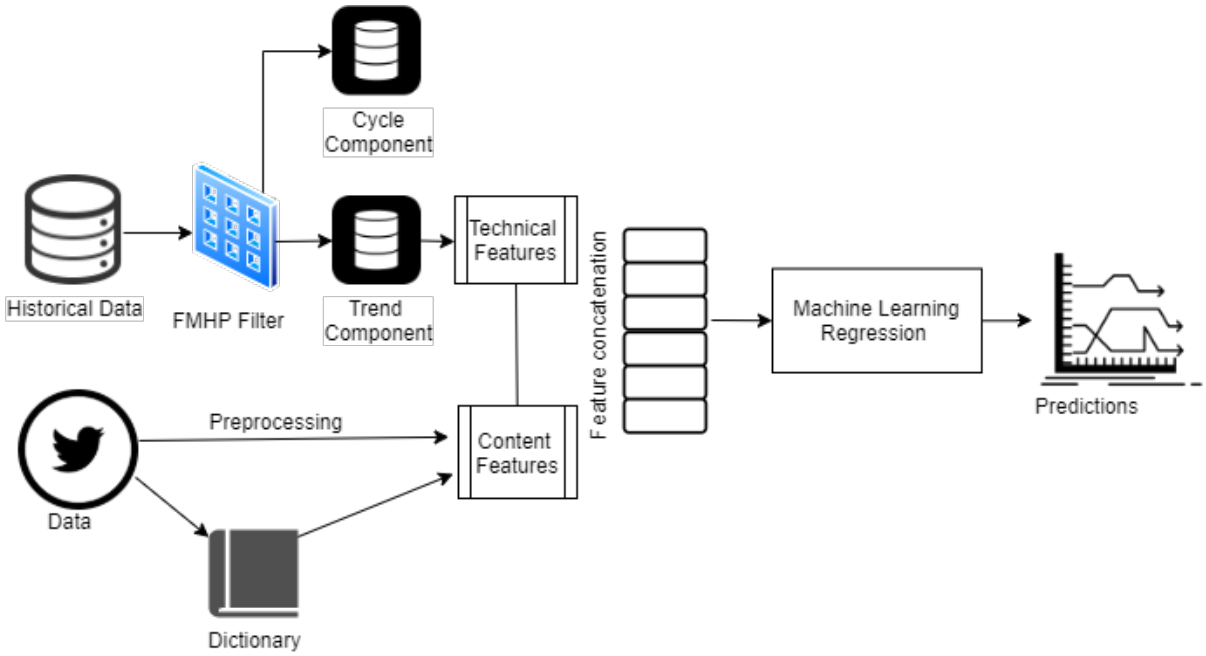

FIGURE 1. A hybrid SVR model to predict stock movement using novel features and FMHP filter.

- We evaluated the importance of aggregation and incorporation of new sentiment features extracted from online social media data sources related to the information on stock prices.
- We thoroughly examined various new technical feature's filter such as a fully modified HP filter to improve the smoothness of historical data and to improve the performance of the prediction model.
- We validated the efficiency of the proposed features and noise filtering approaches using five different machine learning techniques (three traditional- and two deep learning-based methods). Our experiments show that our model can achieve comparatively better prediction performance than already existing methods in literature.

The rest of the paper is organized as follows. Section 2 reviews the existing work about stock movement prediction, machine learning and filtering techniques involved. Section 3 explains the research data sets collection. Section 4 illustrates the overview of the proposed hybrid solution. Section 5 presents the details about the prediction models used. Section 6 describes the experimental results and comparisons. Section 6 provides the discussion and finally, Section 7 summarizes the whole paper.

II. EXISTING WORK

A. STOCK MARKET PREDICTION

In the financial world predicting stock movement is an important objective, because successful predictions lead towards high financial revenue and prevent investors from market risks. In USA empirical study on stock market prediction shows a significant result on the rate of stock returns. Deep learning methods have more probability to predict stock movements using historical data set [15]. Existing researches made many successful attempts by applying Machine learning algorithms filtering methods and feature engineering techniques. Some other uses a hybrid model to predict stock price movement. New techniques and methods encouraged many other researchers to explore more knowledge in predicting the financial market [16].

B. PREDICTIVE ANALYTIC TECHNIQUES AND MACHINE LEARNING

Predict result generation through a certain systematic process is called Predictive modelling. If results are in categorical form than their outcome is known as classification and if they are in a numerical form then they called regression. Where clustering, association rules and descriptive modelling are used among observations as an interesting association. Predictive analytic such as determine machine learning regression algorithms and other statistical formulas help researchers in predicting stock future price. The objective is to make predictions about business's activities happened in the future [17], [18].

Regression analysis tested by predicting expected values through calculating observed values. This represents a kind of function related to data mining which forecasts a number. Commonly regressions data set is subdivided into two parts, one for constructing model and other for testing the model. Regression helps to forecast trends for business planning environmental modelling advertising financial forecasting and time series forecasting. Various families of regression algorithms are used to conduct prediction analysis and measuring error rates [19]. In early work traditional neural network to forecast stock index. Neural Network has the capability

to forecast over time stock prices. Researchers mostly used Support vector Machine because of its strong ability of classification. Other used SVM for higher frequencies trading for short term predictions and price fluctuations [20]. Furthermore, through combine SVM and Genetic algorithm (GA) a hybrid system of machine learning is also quite effective [21]. Moreover, social media data sources assist in a great deal in predicting future asset returns and simplifying forecasting of stock volatility [22]. Some researcher incorporates the Neural Network as a part of the hybrid model, while others suggested two stages approach e.g. Artificial Neural Network (ANN), Support Vector Regression and (combines Genetic algorithm (GA) and Recurrent Neural Networks (RNN)) to forecast stock prices and get better outcomes above all economic approaches.

Recently, deep learning is getting research attention for economic market predictions, many researchers implemented deep constitutional neural network (CNN) to the processed embedding event that is collected and selected from different news websites to forecast stock price indexes [23], [24]. Decision tree regression is applied for classification purposes by decomposing data set into small subsets and used them as time to time which give output in the form of a tree structure [25]. Standard deviation reduction and multiple linear regression are applied to find the association between two variables one is the input variable, and other is a response variable in a vector space [26], [27]. Linear regression has columns corresponding to regression variables such as $x1, x2, \ldots, xm$, columns as interaction term of order, columns of one's who defined intercept [28], [29]. Many other researchers used SVM and KNN algorithms with different kernel techniques and various solutions made predictions on the bases of sentiments and financial news posted on social media data sets [30], [31]. Social media platforms spread the information fast and can provide useful data for informing the impact of stocks rise and fall with people behaviour [32]. Some of the researchers use various indicators with social media to find the impact on the economy [33]. Accuracy in financial data increases the performance while predicting stock prices that's why a lot of researcher uses filtering techniques [34]–[36].

C. FILTERING TECHNIQUES

Time-series analyzing usually contains two phases: 1): Contains a model which is able to represent time series. 2): Second using that model to predict future values. Time series are the observations of linear sequence on a specific variable. Commonly observations are selected on regular bases like days, months, years etc. But their sampling couldn't be regular. Suppose a time series consist of regular patterns. Values of time series become a function of earlier time series values. X in equation 1 and 2 is taken as a targeted value which we are trying to predict where X_t indicates a value of X at time t, intervals, aim is to develop a model [10].

$$X_t = f(X_{t_1}, X_{t_2}, X_{t_3}, \ldots, X) + e_t \tag{1}$$

$X_t - 1$ represent the value of X for previous observations, $X_t - 2$ is a value of two observations, etc.Where e_t denotes noise present in data which couldn't follow predictable patterns and known as random shock.The values of those variables which occurs past from current observations are called lag values. And if financial time series follows repeating patterns, then the value of X_t is extremely much correlated with the $X_t - cycle$ component value where cycle shows the number of current observations in a regular cycle. Like, whole month observations to an annual cycle are model by following:

$$X_t = f(X_{t_{12}}) \tag{2}$$

The objective of constructing a financial series model is the same as aim for predicting other types of models e.g. finding error among predicting values of targeted values and observed values. The financial series analysis becomes one of the basic forecast analytic need for several businesses. As most of the elements of the data are observed data elements such as products sales, stocks prices and others. For the strategic view, mostly managers and decision takers will continuously need predicted trends and seasonal patterns for different types of elements. All forecasting time series methods are categories into two methods: Qualitative and Quantitative methods. There is always a risk of containing some noise which influences whole information of time series data set.

Because of daily fluctuates it is difficult to understand and analyze the change in trend. Forecasting long trend macroeconomic, financial stocks exposes the performance of entire investment atmosphere to some extent. Noise filtering is important for excellent trend analyzation. This work evaluates two noise filtering techniques HP and FMHP filter, compared their effectiveness by filtering time series data. Noise impacts the entire information present in the original data directly. Before applying the predictive model, it is better to analyze and understand the trend and improving the accuracy parameter of a stock price.

Business cycle approximation is essential for macroeconomics research. Hodrick-Prescott (HP) filter, is the most famous tool for extracting cycle from macroeconomic time series [37]. But it has certain issues which include a fixed value of λ across time series and the endpoints bias (EPB). To minimize the first issue, McDermott in 1997 [38] proposed Modified HP filter (MHP). After that, Bloechl in 2014 [39] introduced a loss function minimization approach to encounter the EPB issue while keeping lambda fixed same as in HP filter. Hanif *et al.* [40] merge the endogenous lambda method of McDermott proposed in 1997 [38], with the loss function minimization method introduce by Bloechl in 2014 [39] to examine End Point Bias in Hyper Prescott filter, while intuitively changing the weighting scheme used in the latter. Hanif *et al.* [40] suggests an endogenous weighting scheme associated with endogenous smoothing parameter which resolves the EPB issue of HP filter and called it as fully modified HP filter. FMHP filter outperforms many of

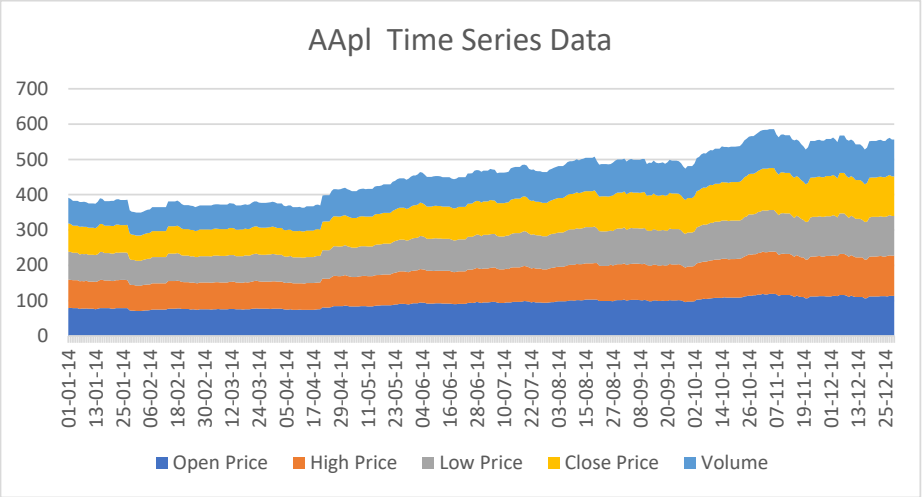

FIGURE 2. AAPL Financial Time Series

conventional filters in power comparison study as well as in real observed data (multivariate and univariate) analytic for large countries.

III. RESEARCH DATA SETS

Two data sets were used in this work (i.e., stock market historical data and Twitter textual data) from the same time period, useful to get more accurate outcomes while predicting stock close price.

A. HISTORICAL TIME SERIES DATA SET

The historical data was obtained from Yahoo Finance Stock Index. Yahoo finance is part of Yahoo network that provides financial news, international market data including various stock quotes, released media, financial reports, commentaries and other original content. In the stock market, the closing price is an important standard benchmark of any specific trading day. Closing price is considered the most knowledgeable estimate of security until trading commences over next trading day. Closing price of any financial time series provides a valuable marker for traders to assess the fluctuation. Closing price is used to measure sentiments of the market for trading day in order to compare current price of one trading day with previous day closing price. These are the reason behind choosing close price as forecasting target for original financial time series data. Dataset is collected on daily bases from Yahoo finance[1] during a period from (2014 to 2015) for the Search Apple Inc. Pvt. Limited (AAPL). Our data contains 6 attributes with 366 samples, which contains Close price, High price, Volume, Open price, Low price, and Date as shown in Figure 2. Our aim is to forecast the future closing price for various amount of time.

B. TWITTER

Forecasting stock movement through social media is a well-known area of interest. Now a day's social media as became one of the most famous platforms for representing public opinion and sentiments about the current situation. Previous researchers forecast different stock markets extracting data from Facebook, Sina Weibo and other social media forums. However, twitter is a famous platform and become the researcher's attention because of its large number of followers and public comments. Public sentiments becomes an emerging ground of research. Early research help in gathering public opinion and determines cumulative public mood through twitter analyzation. We use Twitter data[2] from (2014 to 2015). Ups and downs in market are correlated through public sentiments and opinions which are expressed by peoples in tweets. Twitter has a large number of users, who upload useful content on daily basis. Aim of selecting twitter

[1]https://finance.yahoo.com/quote/YHOO/history?ltr=1
[2]https://github.com/yumoxu/stocknet-dataset

is to analyze, how accurately change in AAPL stocks price will be predicted. Novel Sentiment features are explained in next section.

IV. METHODOLOGY

This section describes the proposed hybrid model for stock prediction. First, we describe the preprocessing step followed by a compact description of each technique. Finally, we describe the application of the methods for stock prediction.

A. PREPROCESSING

1) Filtering Historical data using FMHP Filter

Hanif *et al.* [40] proposed new endogenous lambda method by signifying some intuitive modifications in weighting scheme along with endogenous smoothing parameter to resolve EPB issue of HP filter which is known as fully modified (FMHP) filter. FMHP first estimates λ endogenously then estimates gt (growth component), $k\lambda$ by using leave-out approach of McDermott, $\lambda = 1$ as starting value. Working of Fully Modified HP filter explain in [16]. The main changes applied in Hodric-prescott filer are:

- Use of linear or nonlinear increase of penalization which minimizes cumulative loss in at terminal points.
- g_t denotes the growth component of y_t where y_t=g_t+c_t, c_t is the cyclical component of y_t.
- Fixed the value of $k = 20$
- Endogenous weights (for end observations) i.e. endogenous α.

Figure 3 shows the extraction of trend and cyclical component after applying Fully Modified HP Filter on time series data.

2) Twitter Data Preprocessing

After collection, the twitter data is first pre-processed for analysing. Steps involved in this phase are elaborated in Figure 4. First, we arrange per day tweets, gaps are filled by inserting missing tweets with previous day tweets. Then, all the data is converted into lowercase. After that, we remove all numbers, punctuation's, stop words, URLs, and white spacing before analyzing the data.

3) Technical Features

Our regression model focuses on the closing price [10]. The goal here is to fit the following association via performing regressions analysis. Fully Modified Hodrick–Prescott Filter are used to filter the noise and separately normalize data value of each attribute. A Fully Modified Hodrick–Prescott Filter decomposes financial time series and convert using common frequencies in numerous series. It helps in smoothness of data. Where,

$$Close_{t+1} = f(Open_t,\ Close_t, Hight_t, Low_t, Volume_t)$$

Technical feature will be measured on the bases of historical data of stock market where t is current trading day and previous trading day taken as $t-1$. First five features of individually trading day will be directly obtained from dataset [10]. Other features will be calculated using the historical data shown in Table 1.

4) Data Analyzing using Domain-Specific Dictionary

Many studies discovered that taking a better decision on any social, economic, and political issues involves examining social news website and other interrelated information. In this aspect predicting stock markets fluctuation involves partly analyzing the association between social-economic news reports and the patterns of the stock market price. So far, in literature opinion mining, many of the studies include a sentiment dictionary to provoke sentiment values from a huge number of documents. A sentiment dictionary contains a pair of selected words and their sentiment values. In our work, the number of frequencies ni of each keyword is present in twitter data according to per day tweets. We calculate standard and mean sentiments using domain specific dictionary[3]. After calculating standard and mean sentiments with respect to each word, we than calculate average value of each day i mean and standard sentiments. Average is taken as a representative of the whole data list which represented by a single value. Here we use arithmetic mean to calculate the average mean and standard sentiment of the day because in statistic arithmetic mean is used as a center tendency of the data set. Average value of each day i mean and standard sentiment is taken from Twitter data [41]. Here, arithmetic mean of an observed data set is the sum of all numeral values of each and every observation divided by the sum of all numbers n containing observations shown in Table 2. Suppose, in $x_{1+}x_{2+}x_3, ..., x_n$ dataset, A denotes the arithmetic mean and represented as equation 3 [54], [55]:

$$A = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_{1+}x_{2+}x_3, ..., x_n}{n} \tag{3}$$

V. PREDICTION MODELS

A. SUPPORT VECTOR REGRESSION

Modern theory was developed by Vapnik and Chervonenkis over the last three years. Where for the first-time support vector machines (SVM) was used to solve regression issues and called it a support vector regression (SVR). The fundamental idea behind SVR is to make non-linear original dataset X (shown in Equation 4) into high dimensional feature space so that linear regression is applied [10]. Suppose there's a set of X given, then $xi \in X = Rn$ is an input vector, $yi \in Y = R$ of the matching out value. Where i is the total number of datasets, Support Vector Regression (SVR) function is,

$$S = (x_1, y_1), (X_2, Y_2).....x_i, y_i) \in (X * Y) \tag{4}$$

and,

$$f(x) = w.\phi(x) + b \tag{5}$$

[3]https://nlp.stanford.edu/projects/socialsent/

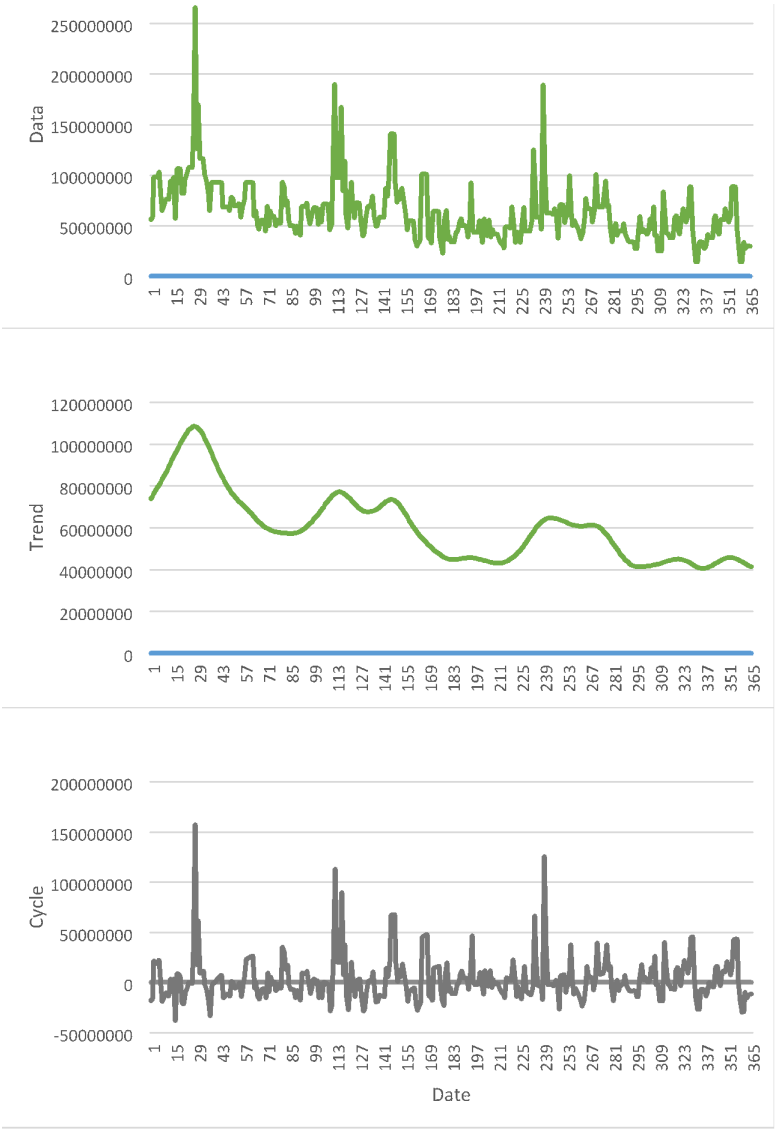

FIGURE 3. FMHP Filter Analysis

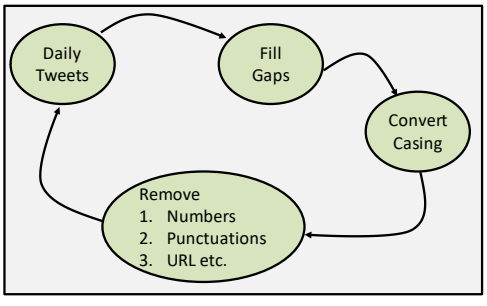

FIGURE 4. Preprocessing of tweets data

$w(x)$ in Equation 5 [10] has a function of nonlinear mapping, where w is weight vector, b used as bias value, evaluated by minimizing the risk function.

$$R(w) = \frac{1}{2}||W||^2 + C\sum_{i=1}^{i} L_e((y_i), f(x)) \tag{6}$$

$\frac{1}{2}||W||^2$ in Equation 6 [10]is a flatness function and C as penalty parameter, describes trade-off between training error and generalized performance, let $L_e((y_i), f(x))$ is known as insensitive loss function and describes as,

$$L_e((y_i), f(x)) = \begin{cases} |(y_i), f(x)| - \varepsilon, |(y_i), f(x)| \geq \varepsilon \\ o, |(y_i), f(x)| < \varepsilon \end{cases} \tag{7}$$

In Equation 7 [10] $|y_i - f(x_i)|$, is defined as predicting value of an error and $\in$ is defined as a loss function, when error for estimation taken as an account, by using two positive slack variables which ζ and ζ^* represents the gap between original values with corresponding to other boundaries values. Where

TABLE 1. TECHNICAL FEATURES

TECHNICAL FEATURES	DESCRIPTION
O_t	Opening Price of the market on time t.
C_t	Final price on which market is traded on time t.
H_t	Highest Price of a market s traded during day t.
L_t	The lowest price of a market is traded during day t.
V_t	Shares traded in the market during day t.
$V_t - V_(t-1)$	Volume change [1]
$\frac{V_t - V_(t-1)}{V_(t-1)}$	Volume limit [1]
$\frac{H_t - L_(t-1)}{C_(t-1)}$	Amplitude [1]
$\frac{C_t - O_(t-1)}{C_(t-1)}$	Difference [1]
$R_(t,f) = \frac{LN(C_t,f)}{C_t-1,f} \times 100\%$	Return of firm f at time t [Novel]
$ROP = O_t - O_(t-1)$	(ReturnOpenPrice) open-to-open [Novel]
$RCP = C_t - C_t - 1$	(ReturnClosePrice), close-to-close [Novel]
$DROP = \frac{O_t - O_(t-1)}{O_(t-1)}$	Change in ReturnOpenPrice [Novel]
$DROP = \frac{C_t - C_(t-1)}{C_(t-1)}$	Change in ReturnClosePrice [Novel]
$VPT = VPT_(t-1) + V \times \frac{C_t - C_(t-1)}{C_(t-1)}$	VPT (Volume per total) is measured when the volume is multiplied by change price and is calculated as running price total from prior period [Novel]

TABLE 2. Sentiment Features with Formulas

Content Features	**Formula**	**Description**
Average Mean Sentiment	$Avg = \frac{(x1+x2+\ldots+xn)}{n}$	x denotes the number of keywords and n the total number of keywords present in each day.
Average Standard Sentiment	$Avg = \frac{(x1+x2+\ldots+xn)}{n}$	x denotes the number of keywords and n is the total number of keywords present in each day.

the selection of kernel function is an important for effectiveness of SVR. Where, there's no such mature theory proposed in selection of SVR kernel function [37].

B. RECURRENT NEURAL NETWORKS

We also apply Recurrent neural networks (RNN) to get more accurate results. Recurrent neural networks are powerful form of neural network which is designed to handle the sequence of dependencies, and it is often used for time series prediction [41]. RNN accomplish similar task for every element in sequence and its current output is depending on previous calculations. The architecture of a single hidden-layer of RNN provides the connection present between each unit that forms a direct cycle. In our work RNN, used input values of the t^th day $xt = (xt, 1..., xt, m)$ where m-vector indicates the features described in prior subsections. The algorithm iterates over the following equations [1]:

$$ht = tanh(Uxt + Wht1 + b) \tag{8}$$

$$ot = tanh(Vht + c) \tag{9}$$

Here, ht denotes the hidden state calculated on the bases of previous hidden states $ht - 1$ and input xt for current time step. ot in equation 9 is the predicted output and considered a stock price indicator for next trading. RNN trained three parameters U, V and W where U indicate (input-to-hidden), V shows (hidden-to-hidden), and W denotes the (hidden-to-output). RNN trained itself on the bases of long arbitrarily information in sequence. It is due to vanishing gradient issue, RNN is hard to learn long-term dependencies. To tackle this issue [1] proposed Gated Recurrent Units (GRU), where rt and zt are known as reset gates, which are utilize for the combination of new input xt with earlier memory $ht1$ for computing st. st determines a "candidate" hidden state.

Update gate (zt) helps t in the calculation that how much space previous memory should require. Following equations are used for calculation of GRU [1]:

$$zt = \alpha(xtUz + ht1Wz + bz) \tag{10}$$

$$rt = \alpha(xtUr + ht1Wr + br) \tag{11}$$

$$st = tanh(xtUs + (ht - 1 \odot rt)Ws + bs) \tag{12}$$

$$ht = (1 - zt) \odot (st + zt) \odot (ht - 1) \tag{13}$$

where $\alpha(x)$ is hard-sigmoid function and $\odot$ represents component wise multiplication. In our work, we applied 2-hidden-layer GRU component and capture the higher-level of feature interactions in between different time phases. Units in second hidden-layer are intended similar to the first hidden layer. To train the RNN we, input the feature vectors of a specified time period from $t0$ to tn, as training data and observed values as a target value i.e. $xt, t = 1 \ldots n$ and $yt, t = 1 \ldots n$ correspondingly. Here, we calculate the dependent variable which is $Yi = \frac{Ci}{C0}1, i = 1..n$, wherever, $C0, \ldots, Cn$ are taken as closing price. Historic data of previous s days is used to predict the price of the n trading day. The starting parameters of GRU unit set by using predefined seed as a guarantee repetitive of RNN models. GRU uses back

propagation approach to train the parameters, by minimizing difference between the ot (output) and observed values yt. For performance evaluation of our proposed model, the total time gap is divided into two steps. Data from $t0tm1$ used for training (GRU parameters) and predict tm tn as dependent data. In second step, GRU parameters are updated after new prediction are calculated i.e. $ot + 1$ where yt are the input into the GRU module for training. It simulates real-world situation for the new stock price because the new price can be obtained on a daily basis and can be utilized as an input for training.

Another very powerful technique based on RNN is LSTM. It has the ability to deal with sequential data and is highly suitable for training and testing the stock market value prediction. These are capable of learning long-term dependencies among data. The underlying working principle of LSTM is same as GRU, except that these have some additional gates. The addition of memory cells can help combat vanishing gradients. It consists of four units: an input gate, an output gate, a forget gate and a self-recurrent neuron. These gates control the interactions between neighboring memory cells and the memory cell itself. The input gate controls the influence of the input on the memory cell, while the output gate controls the amount of memory to retain. Lastly, the forget gate is used to control how much history to remember or forget.

C. AUTO ARIMA BASICS

Autoregressive Integrated Moving Averages (ARIMA) is a statistical model used to predict and analyze time -series data [43]. ARIMA has been widely used to describe the model structure. ARIMA basically establishes the relation between some delayed observations and current observation by applying the moving average. ARIMA has three standard representations; p (lag order): represents the number of lag observations which are included in the model, d (degree of differencing): represents the number of times the differences are calculated for raw observations, and q: represents the size of moving average window. A model is created by configuring the above-specified terms for forecasting a variable. A value of 0 can be used for the parameter with the element that the model will not use.

D. RANDOM FOREST

Random Forest (RF) is an ensemble of classifiers, which makes predictions by combining the results from many individual decision trees [44]. It is like bagging method but offers an improved method of bootstrapping. It generates a set of classification and regression tree (CART)-like classifiers. To output the final result, averaging is performed over all predictions from these sampled trees in case of regression or by majority vote in case of classification. In our case, we used the boosting method for its simplicity. Technically, the feature sampling is used to generate a subset of data. The number of features used for splitting is an adjustable user-defined parameter. It is worth noting that limiting the number of features for split can reduce the computational complexity of the algorithm. In addition, it can help process high dimensional data efficiently and define relatively deeper trees. The final results are then obtained by averaging over the individual results obtained from each subtree.

VI. EXPERIMENTAL RESULTS

This section describes the experimental setup and the quantitative results obtained from the proposed model. As explained above, the experiments are done by using two the datasets (i.e. time series and (public mood) from twitter) collected from Dec 31st (2014 to Dec 2015). We employed Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as metrics for evaluating the proposed method. Table 6 summarizes the parameter values selected. The prediction performance of all classifiers is investigated for close price prediction on the data obtained for the year 2014 – 2015 for Apple Inc.

A. EXPERIMENTAL SETUP

Our main objective is to perform a day-ahead stock close price prediction. We used two weeks (14 days) historical samples as input to train the model and then predict the stock close price of the next day. The rolling recursive strategy was employed for processing both training and testing data. The time-series data was transformed into M x N matrix using phase space reconstruction method. Where M represents the number of days which was set to 14 and N is the number of samples. Before running the final experiments, we first divided the data into training (80%) and testing (20%) data sets. We used cross-validation to identify the optimal parameters for the classifiers. We applied a grid search algorithm to identify the optimal parameters for each classifier.

SVR with radial basis function (RBF) was selected due to its good performance in this experiment. This required two important parameters, cost (C) and gamma (γ). The optimal value selected for C=275 and for γ=0.1. One of the main challenge in applying support vector regression is to define the parameters value of proposed model. Resolving that issue, optimized parameter values are implemented in the model. Every time dataset while prediction experiments is divided into two subsets: training set and test set. Table 3 shows kernel selected parameters value by using grid search which helps us in producing better results for this analyzation. We use regression as regressions performance is depended upon kernel that's why we select radial (RBF) kernel, because it is popularly used with SVM. The (RBF) kernel on the two samples x and x', represents as feature vectors in any input space. We use kernel optimal parameter values shown in Table 3 and compare it with [10] results.

- g kernel parameter
- d degree parameter
- c penalty parameter

For RF, we fine-tuned two parameters, the number of decision trees (nt) and the maximum number of features considered at each split (nf). In our experiments, we empirically

TABLE 3. Kernel parameters setting

Regression model	Kernel	C	G	D
SVR	Rbf	250	0.01	3
SVR+HP	Rbf	275	0.1	3
SVR+FMHP	Rbf	285	0.1	3
SVR+FMHP+Novel Features	Rbf	285	0.1	3

set the values for both parameters, i.e. nt =40 while nf = 4. The literature review indicated that the RF is less sensitive to nf therefore, we set it to a constant value. we performed convergence tests for the RF on our training set to find the optimal values for its parameters. It is worth mentioning that initially, the accuracy of the RF increased as we increase the number of trees. However, after reaching the number of the tree to 40, we did not see any further improvement in terms of out of bag error (OOB). Therefore, we selected this value as the optimal value for training. Table 4 summarizes the OOB for 10 iterations for a time window to 30 days on the training data. The choice of splitting criteria was defined using Gini impurity.

TABLE 4. Optimal parameter values for RF from OOB for 10 iterations with time window of 30 days

No. Trees	OOB
5	0.325889
10	0.295832
20	0.105478
30	0.095524
40	0.056235
50	0.065822
80	0.190352
100	0.125560
200	0.345323
500	0.452200

For ARIMA, the (p, d, q) were empirically calculated as (1,0,2). These values were manually selected as not all possible combinations could be tested due to time constraints. The values which produced optimal values were selected for training the model. Table 10 summarizes the various combination of the parameters and their corresponding standard error of regression (SER). It shows that the best relative results were obtained for the parameter setting of (p, d, q) = (1, 0, 2). The smallest Bayesian information criterion (BIC) obtained was 3.5042 and relatively smallest standard error of regression (SER) of 0.443804.

TABLE 5. Optimal parameter values for ARIMA

ARIMA	BIC	SER
(0,0,1)	6.3242	0.719390
(0,1,0)	4.2265	0.880561
(0,1,2)	4.9397	0.556230
(1,0,0)	8.6980	0.971576
(1,0,2)	3.5042	0.443804
(1,0,1)	8.5138	0.532737
(1,1,0)	8.0925	0.612315
(2,0,0)	9.1971	0.733480
(2,1,2)	6.2086	0.596781
(2,2,1)	7.0138	0.560580

B. HODRICKPRESCOTT FILTER AND FULLY MODIFIED HODRICKPRESCOTT FILTER

The presence of noise in the data may influence the performance of the machine learning algorithm. As the stock market, volume varies every day and don't express any signs of predicting stocks in the stock market, therefore it is difficult to understand the whole trend. For better understanding and analyzing the trend, it essential to filter the noise from the dataset. In this research, we evaluated two different types of noise filters and compare their efficiency on the analysis of financial time series.

It is quite extensive to extract the cyclical component from a financial time series, end point bias issue (EPB) of HP filter as explained in Section 2. FMHP filter uses nonlinear or linear increased in penalization (which minimized cumulative loss) on terminals by fixing end point observations to penalizing end point observations' weights. FMHP outperforms all other conventional filters e.g. HP, BK and CF filter. End point performance (EPB) of Fully Modified HP filter is evaluated in this research and found best as compare to HP Filter. When we apply FMHP filtering on our dataset we found that the extraction of trending and detrending is better than HP Filters, where FMHP automatically adjust smoothing parameter according to the macroeconomics time series. The use of FMHP Filter provides a better estimation of the cyclical behaviour to analyze financial data series. Economic managers and stockholders will get better knowledge about the state and change to occur in economic dynamics and thus they will be better able to take the essential step and stabilization measures at right movement. Before performing the regression, we need to use a different filter technique for each experiment to filter the noise and normalize the data values on each attribute separately. These filter techniques include: Hodrick Prescott filter, Christiano Fitzgerald filter, Baxter King filter but for improvement and getting better results found HP filter perform well in all three that's why we only implement the Hodrick Prescott filter and Fully Modified Hodrick Prescott filter (FMHP) [40]. Comparison of both (Hodrick Prescott filter and Fully Modified Hodrick Prescott filter) trend and cyclical component shown in Figure 5 and 6. The main goal of this filter to decompose financial time series into several more series with respect to common frequencies. And here we decompose our dataset into the trend and a cyclical component where we only used trend component as our requirement. Regressive model using various libraries and packages including: e1071, lattice, kernlab. Table 7 shows the comparison of our filtering technique Fully Modified Hodrick–Prescott filter (FMHP Filter) with Hodrick–Prescott filter (HP Filter) results using AAPL stock historical series and found that FMHP Filter performs better than HP Filter.

C. EVALUATION

To evaluate the proposed method, two statistical evaluation techniques were used, which are Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Per-

TABLE 6. Evaluation Metrics

Evaluation Metric	Formulas	Description
Accuracy	$A = \frac{n_{corr}}{N} \times 100\%$ [10]	n_{corr} is the number of training days and N denotes the number of total trading days.
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(Y_t\text{-}O_t)^2}$	O_t denotes prediction of Y_t
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{N}\sum_{t=1}^{N}\lvert\frac{Y_t - O_t}{Y_t}\rvert$	O_t is the prediction of Y_t and is measure as the qualitative performance.

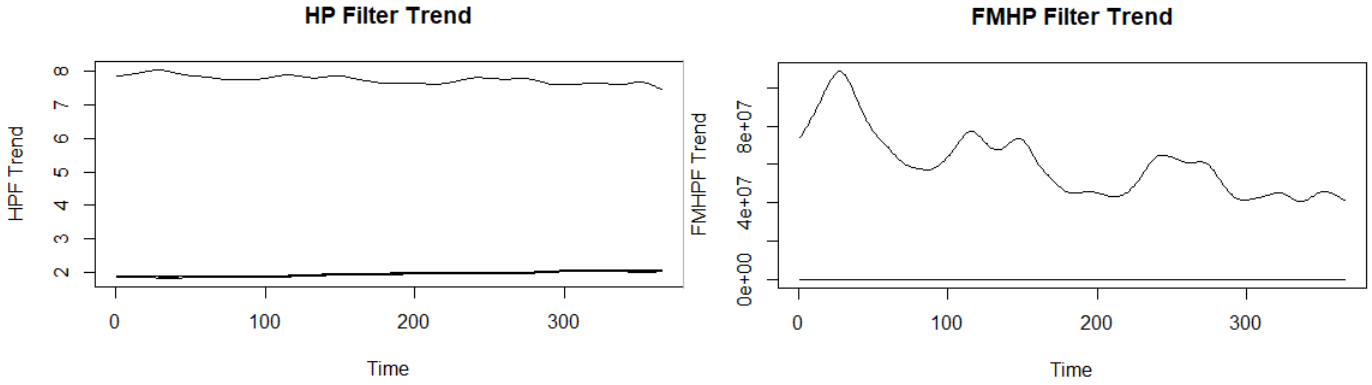

FIGURE 5. Comparison of HP and FMHP Filter Trend

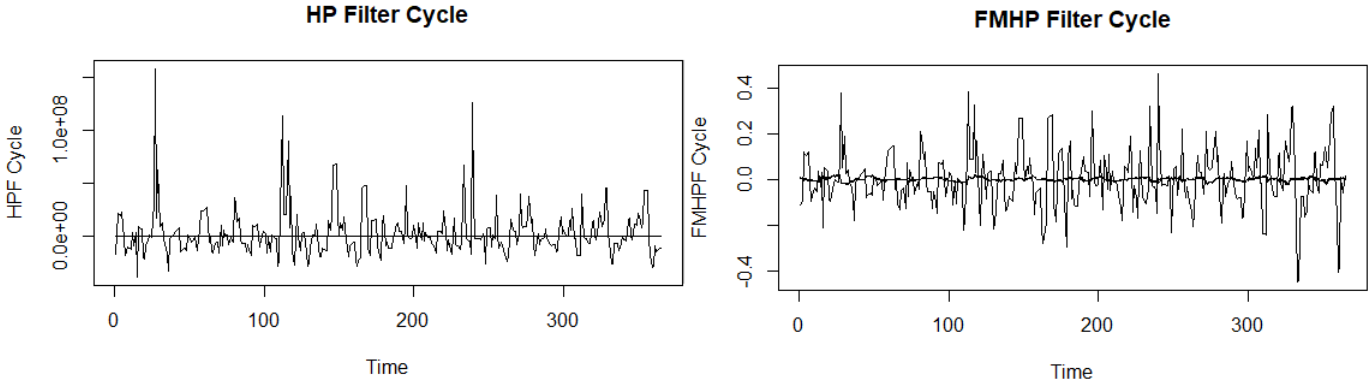

FIGURE 6. HP and FMHP Filter Cycle Comparison

TABLE 7. Comparison with HP Filter [10] with FMHP Filter

Years	SVR	SVR + HP	SVR+FMHP
2014-2015	0.25	0.23	0.2

formance evaluation matrices shown in Table 6 are used to compute error rate between the original and predicting stock price for the evaluation of our proposed model. Accuracy is adopted to evaluate qualitative performance for prediction measurements. Accuracy calculates the consistency present in predicting direction of stock price between predictions and observations. Several experiments were performed to investigate the effect of the filtering approach on the prediction of the stock value. Before applying the regression algorithms, first, we used a filtering technique for each experiment to filtering noise and normalize our data values separately. In addition, experiments were also performed without applying the filters to find the effect of filtering on the accuracy of the machine learning algorithms. These filtering methods include Hodrick Prescott filter and Fully Modified Hodrick Prescott filter. The aim of using these filters to decompose our data into various other series with common frequencies such as trend and a cyclical component.

1) SVR Results

The SVR was trained on the training data set for a time window of 30 days and then tested the accuracy of prediction on our test data set. The results obtained for forecasting the close price of Apple stock using SVR with Hodrick–Prescott and Fully Modified Hodrick–Prescott filter with several combinations of technical and sentiment features are summarized in Table 8. The results indicate that the SVR with HMHP out performed HP filter and SVR results. This is a promising result which suggests that the FMHP filter is effective in reducing the noise before applying the SVR. Using HP improve approximately 0.09 in terms of RMSE over SVR alone while

FMHP improved 0.16 over SVR.

TABLE 8. Evaluation Matrices with our Model Results using SVR

Dataset with Novel Features	Year	Accuracy	MAPE	RMSE
AAPL+SVR	2014-2015	66%	0.25	0.14
AAPL+SVR+HP	2014-2015	67.01%	0.23	0.11
AAPL+SVR+FMHP	2014-2015	68%	0.2	0.08
AAPL+Tec+SVR+HP	2014- 2015	68.22%	0.19	0.08
AAPL+Sent+SVR	2014-2015	68.99%	0.17	0.06
AAPL+Tec+FMHP +Sent+SVR	2014-2015	69.81%	0.14	0.08

2) RF Results

The RF classifier was also trained in a similar fashion on the test data set as SVR. The results obtained using RF are summarized in Table 9. In the case of RF, the highest accuracy was obtained for the combination of RF with HP filter. Similarly, the accuracy of RF with FMHP filter was also slightly higher than RF alone in terms of both RMSE and MAPE. The results indicate that the SVR with HMHP outperformed HP filter and SVR result alone. This is a promising result which suggests that the FMHP filter was effective in reducing the noise before applying the SVR. Using HP improve approximately 0.09 in terms of RMSE over SVR alone while FMHP improved 0.16 over SVR.

TABLE 9. Results obtained using RF

Method	No. Trees	No. Features	MAPE	RMSE
RF	40	4	4.64	5.24
RF+HP	40	4	3.95	5.78
RF+FMHP	40	4	2.15	4.22

3) ARIMA Results

Table 10 summarizes the results obtained for ARIMA model for prediction of Apple Inc. close price. These results were obtained for the optimal parameter values (p, d q) as (1, 0, 2). We did not perform any differencing on the original data to change the stationarity. The best results were obtained for ARIMA + HP with MPAE and RMSE on the test were 3.13 and 4.38 respectively. It can be seen from Table xx that the results for ARIMA+FMHP and ARIMA + HP are very close. It is also interesting to note that the results for RF+HP and ARIMA + HP are very close in terms of MAPE and RMSE on the same data used for testing.

TABLE 10. Results obtained for ARIMA

Method	MAPE	RMSE
ARIMA	6.13	8.45
ARIMA +HP	3.13	4.38
ARIMA +FMHP	3.59	5.17

4) Recurrent Neural Network (RNN) Results

We used two popular variations of the RNN namely LSTM and GRU. The overall architecture and parameters for both variations were the same. We used four layers with 50 units in each layer with hyperbolic tangent function as its activation. Learning rate was set to 0.01 and Adam optimizer was used. Since the data was small, therefore, no dropout was considered. We used the hyperbolic tangent function because its derivative approaches late to 0 which helps in learning longer sequences.

The results obtained for LSTM and GRU are close to each other which are summarized in Table 11. Both LSTM and GRU produced the lowest MAPE and RMSE when combined with FMHP.

TABLE 11. Results obtained using RNN variations (LSTM and GRU)

Method	MAPE	RMSE
LSTM	1.45	2.89
LSTM +HP	1.40	2.70
LSTM +FMHP	1.33	2.56
GRU	1.49	2.95
GRU +HP	1.38	2.89
GRU +FMHP	1.36	2.73

Initially, we tested the worthiness of our proposed features, especially sentiment features. We used a 2-layer RNN model combination with GRU for the prediction of the stock market index to examine different feature subsets. Results are shown in Table 12. Tec and $Sent$ indicate the technical and sentiment features respectively. We adopt different types of seeds for the initialization of our model. The average is measured based on 100 rounds using 0 to 99 seeds of experiments [1]. Adding sentiment feature, accuracy of our model increases significantly from 69.22% to 70.88% and errors decrease moderately as shown in Table 13.

TABLE 12. Comparison of our Model with [10]

Dataset	Error	Years	SVR	SVR+HP	SVR+FMHP
Multiple [10]	MAPE	2013–2014	0.25	0.08	-
AAPL	MAPE	2014-2015	0.25	0.23	0.2
AAPL	RMSE	2014 -2015	0.14	0.11	0.08
AAPL	Accuracy	2014-2015	66%	67.01%	68%

TABLE 13. Evaluation Matrices with our Model Results using RRN

Dataset with Novel Features	Year	Accuracy	MAPE	RMSE
AAPL+RRN	2014-2015	67%	0.23	0.12
AAPL+RRN+HP	2014-2015	68.01%	0.2	0.09
AAPL+RRN+FMHP	2014-2015	69%	0.17	0.05
AAPL+Tec+RRN+HP	2014- 2015	69.22%	0.14	0.04
AAPL+Sent+RRN	2014-2015	70.81%	0.11	0.04
AAPL+Tec+Sent+RRN+FMHP	2014-2015	70.88%	0.1	0.04

5) Comparison of proposed Model with other studies

Our proposed model using SVR and RNN outperforms single RNN and RNN-Boost models when combines (FMHP filter

TABLE 14. Comparison with RRN-Boost [1] with our Model

Model	MAPE (%)	RMSE (%)	Acc (%)
Single RNN [1]	26.21	2.17	65.28
RNN-Boost[1]	24.31	2.05	66.54
Tech+Sent+RNN-Boost	23.97	2.23	70.17%
(AAPL+Tech+FMHP) + Sent +SVR	0.23	0.08	69.81%
(AAPL+Tech+FMHP) + Sent +RNN	0.1	0.04	70.83%

and (Technical and Sentiment Features)) and gets accuracy of 69.81% as shown in Table 14.

In Table 12, comparison of our proposed model with (A novel hybrid model based on HPF and SVR) [10] is shown, where our model outperforms (A novel hybrid model based on HPF and SVR models). For comparison we use AAPL Historical time series data and then by implementing HP filter and FMHP filter we can analyze our model results in which our model out performs SVR+HP Model.

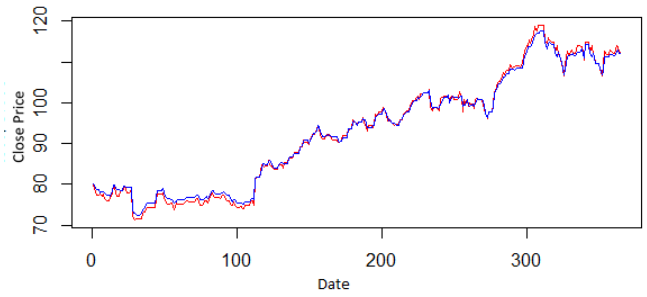

FIGURE 7. APPL + SVR

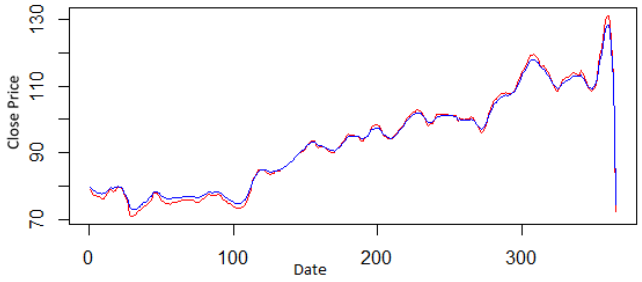

FIGURE 8. AAPL + HPF + SVR

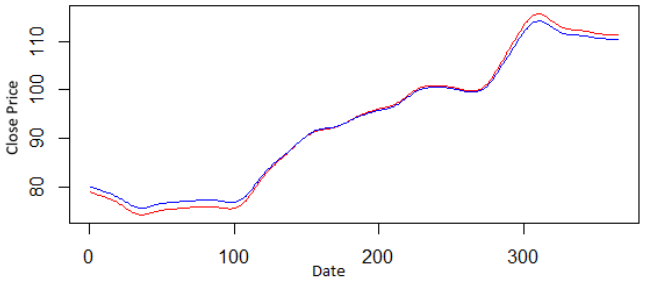

FIGURE 9. AAPL + FMHPF + SVR

6) Proposed Framework Results

We apply SVR model on APPL (Historical time series) data to check accuracy of Hodrick–Prescott filter, Fully Modified Hodrick–Prescott filter with the combination of novel technical and sentiment features. In Figures 7, 8, 9, 10, blue line shows the original Close Price where red line shows, accuracy in predicting AAPL stocks Close Price.

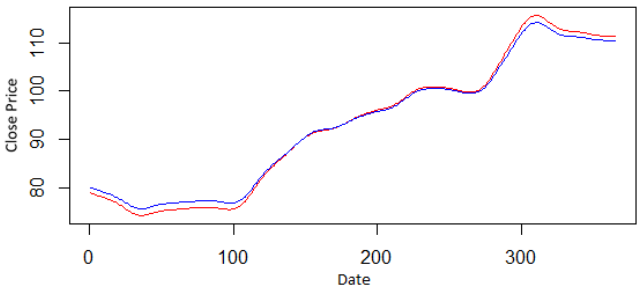

FIGURE 10. AAPL + FMHPF + Tec+ Sent Features + SVR

VII. DISCUSSION

HP filter technique minimizes fluctuations in time series against parameters which approaches linear trend. Where FMHP filter is an extended form of HP filter which not only produced trend and cyclical component but also lower the endpoint base (EPB) and performs comparatively better than HP filter in an analysis of movements. Exploring filtering analysis, we found that trend component produced by the HP filter preserved financial series curve, whereas, trend component produced by BK filter or CF filter tends to slightly change the time series curve and FMHP filter maintain that curve and hence improving endpoint bias. To compute the performance of predicting stock price and our proposed model we conducted several experiments using different ML and DL approaches. Both technical and content features were combined to improve prediction accuracy. The performance of machine learning methods was investigated on original data as well as applying some noise reduction techniques including HP and FMHP. After filtering analysis, we found that trend component produced by FMHP filter maintain that curve and also improve endpoint bias and performs better than HP filter. Because of which, we found the small difference in the performance. On the other hand, prediction outcomes vary from one dataset to another due to the increase in the number of features and time-series data structure. Successfully, the combination of both, the time series and sentiment dataset work well with both ML and DL models. We found that both HP and FMHP are effective for noise reduction and work efficiently for improving the prediction accuracy of the model. In addition, the technical and content features complement each other and hence, help reduce MAPE and RMSE error rates.

VIII. CONCLUSION

The main objective of this paper is to investigate the best combination of machine learning and noise reduction techniques for the prediction of the closing price. In addition, both technical and content features were combined to improve prediction accuracy. Two types of features were obtained, technical features were derived from the historical stock data while content features were obtained from official accounts of Twitter Inc. We used five different machine learn-

ing approaches, three tradition and two deep learning-based approaches. Two approaches for time series data denoising were evaluated: FMHP and HP. We performed several experiments in combination with machine learning approaches for prediction of Apple Inc. stock value prediction using 14 days of historical data.

Conferring to our experiments, we achieved that the proposed model using our hybrid technique and combination of ML and DL models, the FMHP noise filter along with new technical and content features is a powerful predictive tool for analyzing stock market price prediction, content and financial time series. The stocks market price is not only depending on time series data but macroeconomic factors and other external factors such as news greatly impact the stock market price. These types of limitations lead to some problems which needed to be solved for future research. In future, we would like to extend our existing work to improve the prediction accuracy for longer historical data and also to reduce the processing time for deep learning methods. Moreover, the hyper-parameter tuning is also complex, therefore, an automatic hyper-parameter selection approach will be adopted to obtain optimal parameters.

REFERENCES

[1] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Leveraging of social media news to predict stock index movement using RNN-boost," Data & Knowledge Engineering", Nov. 2018 vol 118, pp. 14–24.
[2] W. Long, L. Song, and Y. Tian, "A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity," Expert Systems with Applications,vol.118,pp.411- 424, Mar.2019.
[3] Galicia, Antonio and Talavera-Llames, R and Troncoso, A and Koprinska, Irena and Martínez-Álvarez, Francisco, "Multi-step forecasting for big data time series based on ensemble learning", Knowledge-Based Systems, 830–841, vol 163, Elsevier, 2019
[4] Tidhar, Ron and Eisenhardt, Kathleen M, "Get rich or die trying… finding revenue model fit using machine learning and multiple cases", Strategic Management Journal, 1245-1273, vol 41, Wiley Online Library, 2020
[5] Gandhmal, Dattatray P., and K. Kumar. "Systematic analysis and review of stock market prediction techniques." Computer Science Review 34 (2019): 100190.
[6] Kim KJ. Financial time series forecasting using support vector machines Neurocomputing. 2003;55(1–2):307–19.
[7] Cao, Li-Juan, and Francis Eng Hock Tay. "Support vector machine with adaptive parameters in financial time series forecasting." IEEE Transactions on neural networks 14.6 (2003): 1506-1518.
[8] Ghasemiyeh, Rahim, Reza Moghdani, and Shib Sankar Sana. "A hybrid artificial neural network with metaheuristic algorithms for predicting stock price." Cybernetics and Systems 48.4 (2017): 365-392.
[9] Dash, Rajashree, and Pradipta Kishore Dash. "A hybrid stock trading framework integrating technical analysis with machine learning techniques." The Journal of Finance and Data Science 2.1 (2016): 42-57.
[10] M. Ouahilal, M. E. Mohajir, M. Chahhou, and B. E. E. Mohajir, "A novel hybrid model based on Hodrick–Prescott filter and a support vector regression algorithm for optimizing stock market price prediction," Journal of Big Data, vol. 4, no. 1, Dec.2017.
[11] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." Computers & Operations Research, vol. 32, no. 10 (2005): 2513-2522.
[12] Long, Wen, Zhichen Lu, and Lingxiao Cui. "Deep learning-based feature engineering for stock price movement prediction." Knowledge-Based Systems 164 (2019): 163-173.
[13] J.-L. Seng and H.-F. Yang,"The association between the stock price volatility and financial news a sentiment analysis approach," Kybernetes, vol. 46, no. 8, pp. 1341–1365, Sep. 2017.
[14] Chniti, Ghassen, Houda Bakir, and Hédi Zaher. "E-commerce time series forecasting using LSTM neural network and support vector regression." Proceedings of the International Conference on Big Data and Internet of Thing. 2017.
[15] Badics MC. Stock market time series forecasting with data mining methods Finance Econ Rev.2014;1 (4):205–25.
[16] S. Jeon, B. Hong, and V. Chang, "Pattern graph tracking-based stock price prediction using big data," Future Generation Computer Systems, vol. 80, pp. 171–187, Mar.2018.
[17] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," Expert Systems with Applications, vol.79, pp. 153–163, Aug. 2017.
[18] Higo M, Nakada SK. How can we extract a fundamental of trend from an economic time series? IMES Discussion Paper Series (98-E-5). Bank of Japan: Institute for monetary and economic studies; 1998.
[19] Henrique, Bruno Miranda, Vinicius Amorim Sobreiro, and Herbert Kimura. "Stock price prediction using support vector regression on daily and up to the minute prices." The Journal of finance and data science 4.3 (2018): 183-201.
[20] Nahil, Anass and Lyhyaoui, Abdelouahid, "Short-term stock price forecasting using kernel principal component analysis and support vector machines: the case of Casablanca stock exchange", Procedia Computer Science, 161-169, vol. 127, Elsevier, 2018
[21] de Almeida, Bernardo Jubert, Rui Ferreira Neves, and Nuno Horta. "Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage." Applied Soft Computing 64 (2018): 596-613.
[22] P.D. Azar, A.W. Lo, The wisdom of twitter crowds: predicting stock market reactions to form meeting via twitter feeds, J. Portf. Manage. 42 (5) (2016) 123–134.
[23] Ding, Xiao, et al. "Deep learning for event-driven stock prediction." Twenty-fourth international joint conference on artificial intelligence. 2015.
[24] Chen, Weiling, et al. "Leveraging social media news to predict stock index movement using RNN-boost." Data & Knowledge Engineering 118 (2018): 14-24.
[25] Cervantes, Jair, et al. "Data selection based on decision tree for SVM classification on large data sets." Applied Soft Computing 37 (2015): 787-798.
[26] M. Yu and C. Guo, "Using news to predict Chinese medicinal material price a index movements," Industrial Management & Data Systems, vol. 118, no. 5, pp. 998–1017,Jun. 2018.
[27] Kara, Yakup, Melek Acar Boyacioglu, and Ömer Kaan Baykan. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert systems with Applications 38.5 (2011): 5311-5319.
[28] J. Garcke, T. Gerstner, M. Griebel, Intraday foreign exchange rate forecasting using sparse grids, in: parse Grids and Applications, Springer, 2012, pp. 81–105.
[29] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. a Listening to chaotic whispers A deep learning Framework (model) for a news-oriented stock trend prediction. In Proceedings of the Eleventh ACM of International conference on Web Search and Data Mining. ACM, Los Angeles, California, USA, pages 2018, 261–269.
[30] X. Zhang, J. Shi, D. Wang, and B. Fang, "Exploiting investors and a social network for stock prediction in China's market," Journal of Computational Science, vol. 28, pp. 294–303, Sep. 2018.
[31] P. Hájek, "Combining of bag-of-words and sentiment features of annual reports to predict abnormal stock returns," Neural Computing and Applications, vol. 29, no. 7,pp. 343–358, Apr. 2018.
[32] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, X. Deng, Exploiting topic-based on twitter sentiment for stock prediction, ACL (2) 2013 (2013) 24–29
[33] Oliveira Nuno , Cortez Paulo , Areal Nelson ,The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices, Expert Systems With Applications 73 (2017) 125–144
[34] X. Zhou, Z.Pan ,G.Hu ,S.Tang, and CZhao,"Stock Market Prediction on the High-Frequency Data Using Generative Adversarial Nets," Hindawi Mathematical Problems in Engineering vol. Article ID 4907423, 11 pages, feb.2018
[35] Y.Song, J.W.Lee, J.Lee," A study on novel filtering and relationship between input-features and target-vectors use in a deep learning model for stock price prediction,"Applied Intelligence Springer Science Business Media, LLC, part of Springer Nature 2018, vol. 23, pp. 153–163, Aug. 2017.

[36] P.-F. Pai, L.-C. Hong, and K.-P. Lin, "Using Internet Search Trends and Historical Trading Data for to Predicting Stock Markets by the Least Squares Support Vector Regression Model," Computational Intelligence and Neuroscience, vol. 2018, pp. 1–15, Jul. 2018.
[37] Hamilton, James D, "Why you should never use the Hodrick-Prescott filter". The Review of Economics and Statistics, 831-843, vol. 100, MIT Press, 2018
[38] McDermott, C.J, "An Automatic Method for Choosing the Smoothing Parameter in the HP Filter", Unpublished, IMF, Washington DC,(1997).
[39] Bloechl.A, "Reducing the Excess Variability of the Hodrick-Prescott Filter by (HPF)Flexible Penalization", Munich Discussion, Paper No. 2014-1, University of Munich.
[40] Muhammad Nadim Hanif, Javed Iqbal, M. Ali Choudhary, 2017. "Fully Modified HP Filter," SBP Working Paper Series 88, State Bank of Pakistan, Research Department.
[41] Xu, Yumo, and Shay B. Cohen. "Stock movement prediction from tweets and historical prices." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2018.
[42] Di Persio, Luca, and Oleksandr Honchar. "Artificial neural networks architectures for stock price prediction: Comparisons and applications." International journal of circuits, systems and signal processing 10.2016 (2016): 403-413.
[43] Richards, Daniel W., and Gizelle D. Willows. "Monday mornings: Individual investor trading on days of the week and times within a day." Journal of Behavioral and Experimental Finance 22 (2019): 105-115.
[44] Nti, Kofi O., Adebayo Adekoya, and Benjamin Weyori. "Random forest based feature selection of macroeconomic variables for stock market prediction." American Journal of Applied Sciences 16.7 (2019): 200-212.
[45] S. Fatima and G. Hussain, "Statistical models of KSE100 index using hybrid financial systems," Neurocomputing, vol. 71, no. 13–15, pp. 2742–2746, 2008.
[46] M. B. Patel and S. R. Yalamalle, "Stock Price Prediction Using Artificial," vol. 3, no. 6, pp. 13755–13762, 2014.
[47] K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock Price Prediction Using K-Nearest Neighbor Algorithm," Int. J. Business, Humanit. Technol., vol. 3, no. 3, pp. 32–44, 2013.
[48] S. AL Wadi, M. Almasarweh, and A. A. Alsaraireh, "Predicting Closed Price Time Series Data Using ARIMA Model," Mod. Appl. Sci., vol. 12, no. 11, p. 181, 2018.
[49] M. Almasarweh and S. AL Wadi, "ARIMA Model in Predicting Banking Stock Market Data," Mod. Appl. Sci., vol. 12, no. 11, p. 309, 2018.
[50] A. M. El-Masry, M. F. Ghaly, M. A. Khalafallah, and Y. A. El-Fayed, "Deep Learning for Event-Driven Stock Prediction Xiao," J. Sci. Ind. Res. (India)., vol. 61, no. 9, pp. 719–725, 2002.
[51] Y. Baek and H. Y. Kim, "ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module," Expert Syst. Appl., vol. 113, pp. 457–480, 2018.
[52] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015, pp. 2823–2824, 2015.
[53] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," PLoS One, vol. 12, no. 7, 2017.
[54] Ansarul Haque , Tamjid Rahman, "SENTIMENT ANALYSIS BY USING FUZZY LOGIC, "International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4,No. 1, February 2014.
[55] X. Zhou, Z.Pan ,G.Hu ,S.Tang, and CZhao,"Stock Market Prediction on the High-Frequency Data Using Generative Adversarial Nets," Hindawi Mathematical Problems in Engineering vol. Article ID 4907423, 11 pages, feb.2018

KHALID IQBAL received his PhD degree in Applied Computer Technology from University of Science and Technology Beijing in 2014, and the B.Sc., and MS(CS) degrees from University of the Punjab, Lahore and SZABIST Karachi respectively. He was awarded a fully funded scholarship by the China Scholarship Council for the entire duration of his PhD studies. He also won the excellent researcher (or excellent international research student award) from University of Science and Technology Beijing in the year 2012-13. He is currently an Assistant Professor in the Department of Computer Science, COMSATS University Islamabad, Attock Campus. He is also the director of Pattern Recognition, Images and Data Engineering (PRIDE) research group where graduate students are working under his supervision in different fields such as Data Mining, Social Networks, Image Processing and simulation-based Models. He has worked on Bayesian network application for Privacy Preserving of XML association rules and Text Localization in Scene Images. His research work has been published in several international conference proceedings and reputed journals. His research interests include Pattern recognition, Machine Leaning, Data Mining and Social Networks. He is the recipient of the CSC scholarship and QCRI/Boeing Travel grant.

SIDRAH IJAZ received the B.S. degree in Software Engineering from COMSATS Islamabad, Pakistan, in 2017, the M.S. degree in Computer Science in 2020.Currently working as a data analyst in Rani foods, doing freelancing as a part time job. Also worked as a Software Developer and a Project Manager for Pakistan's largest organization (Allama Iqbal Open University).

SYED ATTIQUE SHAH received the Ph.D. degree from the Institute of Informatics, Istanbul Technical University, Istanbul Turkey. During his Ph.D., he studied as a Visiting Scholar with the National Chiao Tung University, Taiwan, University of Tokyo, Japan and Tallinn University of Technology, Estonia, where he completed the major content of his thesis. He currently works as assistant professor and chairperson at the Department of Computer Science, BUITEMS, Quetta Pakistan. His research interests include big data analytics, cloud computing, information management and Internet of Things.

AKHTAR JAMIL was born in Hassanabad, Hunza, Gilgit-Baltistan, Pakistan. He received the B.S. degree in computer science from the University of Karachi, Karachi, Pakistan, in 2008, the M.S. degree in computer software engineering from the National University of Science and Technology, Islamabad, Pakistan, in 2011, and the PhD degree in Remote Sensing and GIS from Yildiz Technical University, Istanbul, Turkey, in 2018. Dr. Akhtar is currently working as Assistant Professor at the Department of Computer Engineering, Sabahattin Zaim University, Istanbul, Turkey. His current research interests include machine learning, deep learning, pattern recognition, data analytics, and remote sensing

DIRK DRAHEIM received the Ph.D. degree from Freie Universität Berlin and habilitation degree from Universität Mannheim, Germany. Currently, he is full professor of information systems and the head of the information systems group of Tallinn University of Technology, Estonia. The information systems group conducts research in large and ultra-large-scale IT systems. He is also an initiator and a leader of numerous digital transformation initiatives.