*Quest Journals*
*Journal of Software Engineering and Simulation*
*Volume 7 ~ Issue 8 (2021) pp: 01-08*
*ISSN(Online) :2321-3795 ISSN (Print):2321-3809*
www.questjournals.org

**Research Paper**

# Pattern Extraction and Prediction in Time Series Data Mining

[1.] Emmanuel N. Nwajiobi
*Dept of Computer Science , NwaforOrizu College of Education Nsugbe, Anambra State, Nigeria*

[2.] Sylvanus O. Anigbogu,
*Dept of Computer Science, Nnamdi Azikiwe UniversityAwka, Anambra State, Nigeria*

[3.] Gloria N. Anigbogu
*Dept of Computer Science, Nnamdi Azikiwe UniversityAwka, Anambra State, Nigeria*

***ABSTRACT***
*Data mining (DM) can be seen as a process that analyses a large amount of data to find new and hidden information, particularly patterns and relationships buried in data. One of the ultimate goals of data mining is prediction. Predictive data mining is the most common type of data mining that has the most direct business application. In this paper, we tookseveral time series datasets and then applied an algorithm developed and implemented for this work to extract interesting and previously unknown patterns from the datasets. The extracted patterns were used for classification and prediction of future time series patterns. For ease of reference, a pattern refers to a portion of a time series that can be identified as a unit.To facilitate pattern detection, extraction and prediction, we pre-defined patterns as up, down and flat, and having equal length (three data points). Each pattern represents a segment (subsequence) of the time series. The algorithm was tested with fifteen different historical time series datasets obtained online from (a) Nigerian Stock Exchange (NSE), (b) Dow Jones Industrial Average, (c) Nasdaq, and (d) S&P 500 via yahoo finance. Each dataset consisted of 5158 data points, covering the period 2000-2020. The algorithm captured all the pre-defined patterns in the datasets and was able to correctly predict future patterns of 11 out of the 15 different datasets (73 % accuracy). Our algorithm is a veritable tool for time series data mining operations. The methodologies used to develop the system were prototyping and object oriented analysis and design methodology (OOADM); while the tools used consisted of MYSQL (for database implementation), PHP (for backend production), HTML, JAVASCRIPT and CSS for front end development.*

***KEY TERMS:*** *Datamining, time series, prediction, pattern, pattern extraction, time series representation*

*Received 18 July, 2021; Revised: 01 August, 2021; Accepted 03 August, 2021*  *Published with open access at www.questjournals.org*

## I. INTRODUCTION

Due to the availability of large storage systems, fast computer systems and efficient information systems, majority of activities in companies and organisations generate large amounts of data which are typically saved in databases. Notwithstanding this advancement in technology, the question of what to do with the huge amounts of data stored in databases are not always easily obvious or answered in most situations by owners of such large databases. As such, computational algorithms are needed to analyse the datasets to find patterns and relationships buried in data. Massive data sets are rarely profitable; their real worth lies in the possibility to extract useful information for making decisions or for understanding the phenomena that generated such data.

To this extent, information retrieval is no longer enough anymore for decision-making. Thus, the availability of these huge collections of data now created new needs that will help us make better and informed decisions, including making predictions about the future. These new needs include automatic summarization of data, extraction of information buried in stored data, discovery of patterns in raw data, and prediction of future patterns. With the availability of these enormous amounts of data stored in files, databases, and other repositories, it is therefore very important and necessary to develop powerful means of analysing and interpreting the data, as well as extracting interesting patterns that would help in decision-making and prediction.

To make these large data sets more useful, we need techniques to analyse them with a view to finding out something surprising and interesting from the gathered data. In this regard, we are faced with the problem of how to find patterns from the datasets and show that the patterns are useful, informative and predictable. Data mining techniques can be used to discover patterns from large time series datasets and also predict future patterns.

Time series is a collection of observations made sequentially in time (Abdullah, 2016). It is an ordered sequence of values (real numbers) of a variable or variables measured, observed or calculated at regular time intervals over a period of time. According to Pohl and Bouchachia (2012), the following activities can be performed on a time series data: detecting motifs, recognizing and extracting patterns, finding correlation between time series or finding similar time series. Similarly, analysis of a time series can be said to comprise three processing steps, namely: (a) Abstraction (or representation), (b) Mining and Discovery of trends and patterns, and (c) Prediction (Pohl and Bouchachia, 2012). The main focus of this work is pattern extraction and prediction. Figure 1 shows the three stages of time series analysis.

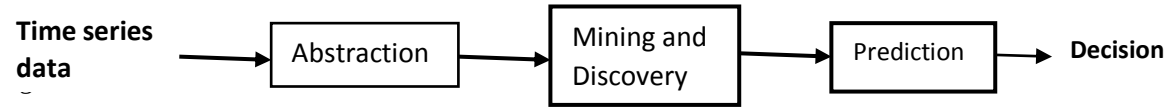


Figure. 1: Processing stages in time series analysis (source: Pohl and Bouchachia, 2012)

Any information of the sequential nature can be processed by pattern recognition algorithms to make the sequences comprehensible and enable its practical use. The term pattern recognition connotes automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities actions such as classifying the data into different categories can be taken (Bishop, 2006). These regularities in data are referred to as patterns in this paper.

Raj et al (2015), posited that pattern recognition is a multi-disciplinary subject covering the following fields: statistics, engineering, artificial intelligence, computer science, psychology and physiology, etc. They noted that computer-based automated pattern recognition systems are required when: (a) the human senses fail to recognize patterns, (b) there is need to automate and speed up the recognition process. Considering the volume of data generated by businesses and companies these days, it is obvious that pattern recognition is inevitable in exploring the data for information buried in the data. People measure things like blood pressure, annual rainfall, value of stock, etc.And therefore such time series also occur in medical, scientific and business domains. Time series reveals the temporal behaviour of the underlying mechanism that produced the data.

However, as the amount of data generated by business houses increases, there is therefore the need to explore new ideas and algorithms to analyse it in order to gather information necessary for decision making and predictions. The type of time series data considered in this paper were mostly those that can generate forecasts, like stock closing price. Based on the foregoing, this research work, proposed a new and novel pattern recognition algorithm/model that can: (a) efficiently detect and extract patterns of interest buried in time series datasets, (b) predict future times series patterns of a given times series application domain. Time series datasets collected via yahoo finance website from different sources were used to test and validate the model.

## II. LITERATURE REVIEW

The goals of any time series analysis or data mining tasks on time series are usually to identify the nature of the phenomenon represented by the sequence of observations and possibly find and predict patterns of the time series variable. Hence, the need to extract patterns from the observed time series and more or less formally describe and predict it. Some related works on time series prediction include:

Kimoto, Asakawa, Yoda and Takeoka, (1990) used several learning algorithms and prediction methods to predict the Tokyo stock exchange prices index (TOPIX). The proposed system used neural network that learned the relationships between the various factors. The output of the system was the best time to buy and sell stocks. They executed simulation of buy and sell stocks to evaluate the system. In their study, vector curve, turnover ratio, foreign exchange rate and interest rate were used as input variables. Trading profit using the system proved better than using ordinary buy and hold strategy.

Trippi and Desieno (1992) in their work performed daily prediction of up and down direction of S&P 500 Index Futures using artificial neural network (ANN). Input variables in the study were technical variables for a two-week period to the trading day: open, high, low, close price, and the price fifteen minutes after the market opening of the current trading day. The output variable was a long or short recommendation. They performed composite rule generation procedure to generate rules for combining outputs of networks. They reported prediction accuracy of 45.3% to 52.8%.

Robert et al (2009), established a financial time series forecasting model by clustering stocks in Taiwan Stock Exchange Corporation (TSEC). The forecasting model integrated a data clustering technique, a fuzzy decision tree (FDT), and a genetic algorithm (GA) to construct a decision-making system based on historical data and technical indexes. The set of historical data was divided into k sub-clusters by adopting K-means algorithm. GA was then applied to evolve the number of fuzzy terms for each input index in Fuzzy Decision Tree so that the forecasting accuracy of the model can be further improved. Different forecasting models were generated for each sub-cluster. In other words, the number of fuzzy terms in each sub- cluster was different. According to their study, the proposed Genetic Algorithm Fuzzy Decision Tree (GAFDT) model had the best performance when compared with other approaches on various stocks in TSEC.

Lee, Lin, Kao and Chen (2010) proposed an effective approach to stock market prediction. The method they proposed converted each financial report to feature vector and used hierarchical agglomerative clustering to divide the feature vector into clusters and then applied K-means for each sub-cluster so that most feature vectors in each sub-cluster belonged to the same class. Then, for each sub-cluster, a centroid was chosen as the representative feature vector and finally this feature vector was employed to predict the stock price movements.

Babu, Geethanjali and Satyanarayana (2012) proposed the use of an effective clustering method, HRK (Hierarchical agglomerative and Recursive K-means clustering) to predict the short-term stock price movements. They used the proposed framework to classify stock time series based on similarity in their price trends. Result of their model HRK outperform support vector machine (SVM) in terms of accuracy and average profit, even as their work used financial report as features.

Senthamarai; Sailapathi; Mohamed and Arumugam (2012) proposed techniques which were able to predict whether future closing stock price will increase or decrease. They combined five methods of analyzing stocks to predict if the day's closing price of a stock would increase or decrease. The methods are Typical Price (TP), Bollinger Bands, Relative Strength Index (RSI), CMI and Moving Average (MA). The results of their technique showed that the algorithm was able to predict if the following day's closing price would increase or decrease. The algorithm performed well on half of the stocks and not so well on the other half of the stocks since it was able to generate both increase and decrease predictions. Thus, the algorithm could perhaps be used as a buying or selling signal, or be used to give confidence to a trader's prediction of stock prices.

Kuo-Ping, W; Yung-Piao, W and Hahn-Ming, L. (2014) presented a model to predict the stock trend based on a combination of sequential chart pattern, K-means and AprioriAll algorithm. The stock price sequence was cut short into charts by sliding window. The resulting charts were clustered by K-means algorithm to form chart patterns. Thus, the chart sequences were now successfully converted into chart pattern sequences, such that the frequent patterns in the sequences can be extracted by AprioriAll algorithm. The existence of frequent patterns implies that some specific market behaviors often appear, therefore, the corresponding trend can be predicted. Experimental results showed that the proposed system can produce better index return with fewer trades. As a result, the proposed method can make profits on the real market, even in a long-term usage.

Shunrong, Haomiao and Tongda (2015) proposed the use of data collected from different global financial markets as the input features to a machine learning algorithm such as support vector machine (SVM) to predict the stock market index movement. Various machine learning based models were proposed for predicting daily trend of US stocks, and numerical results obtained suggested high accuracy. In addition, a practical trading model was built upon their trained predictor and the model generated higher profit compared to selected benchmarks. Hence, they were convinced that index value of stock markets and commodity prices can provide useful information in the prediction process.

These studies on time series domain point to the fact that patterns in time series can repeat themselves. Therefore, detection of patterns similar to those that have occurred in the past can readily provide useful information about the future of time series movement. These algorithms already proposed in the literature modelled time series behaviour. However, the existing studies and algorithms focused more on finding efficient techniques for stock market time series price prediction. Furthermore, issues bordering on mining time series historical datasets to extract patterns and predict future patterns were lacking in the existing researches. And since, patterns are fundamental characteristics of any time series, it cannot be overlooked in any time series data mining process.

Against this backdrop, this work seeks to complement existing researches in time series data mining by designing and implementing a robust automated data-driven pattern recognition and prediction model for time series data mining. The model has an efficient algorithm for time series representation, pattern extraction and prediction.

## III. SYSTEM METHODOLOGY AND DESIGN

This work approached the issue of pattern recognition, representation and prediction of time series from the data mining perspective, rather than from the statistical point of view. This was informed by the fact that statistical tools can not suffice for large time series datasets analysis with respect to pattern extraction and

prediction. As a result, the pattern recognition approach applied was an unsupervised learning since there was no prior labelling or classes of patterns unto which new patterns can be mapped to. However, to facilitate the task of pattern recognition, patterns were defined as either Up, Down or Flat, with fixed lengths of either 3, 5 or 10 data points (i.e. days).

The algorithm started with a historical time series dataset which it received as input. Prior to that, the dataset should have been preprocessed by removing blank cells of data and transforming (normalization process) the dataset into the range [0,1], such that the highest value is 1 and the least value in the series is 0. After this normalisation process, the pattern recognition algorithm can be applied to the resulting dataset to identify patterns of interest and thus represent them with symbols. All patterns identified were symbolized and stored in a database for future uses and manipulation. In order to make prediction of future time series patterns, the immediate past data points (3 days' data points) were collected and entered into the system. Then the module for prediction was invoked to complete the process of predicting the next pattern. The methodologies applied in the development of the system were prototyping and object oriented analysis and design methodology (OOADM); while the tools used consisted of MYSQL (for database implementation), PHP (for backend production), HTML, JAVASCRIPT and CSS were used for front end development.

### 3.1 The Pattern Extraction and Prediction Algorithm

In this section, we propose an algorithm for pattern detection, extraction and representation (using symbols). For ease of identification of patterns, extraction and representation, we pre-defined patterns as either Up (U), Down (D) or Flat (F). Each pattern represents a segment, and can be drawn as shown in figure 2.

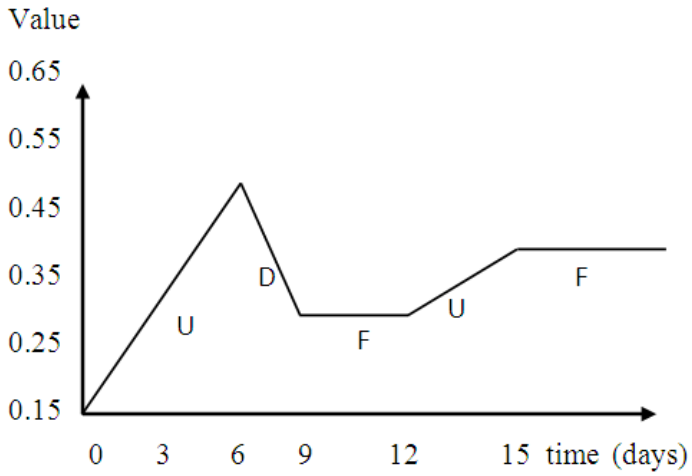


**Figure 2:** Visualisation of patterns of a time series, showing the up, down and flat patterns. where value = average value of data points in a segment: Up, Down or Flat pattern

From figure 2, the symbolic representation of the time series is UDFUF. Therefore, a time series of length 25 (data points) has been reduced to a string of UDFUF (which is five characters).

### 3.2 The algorithm for pattern detection, extraction and symbolic representation:

Input: S, Segment_size
Output: Pattern string symbols (for Up, Down and Flat patterns)
Repeat
Initialize tup = tdn = 0;
For (i = 0; i ≤ Segment_size - 1, i ++) {
df = (i + 1) – i;
if df is positive, tup++ //augment increasing pattern variable
if df is negative, tdn++ //augment decreasing pattern variable
}
If tup = Segment_size - 1 or tdn = Segment_size - 1 then, pattern is Up or Down respectively
otherwise pattern is Flat.
Calculate segment average; //Call SegmentAvg( ) function
Store segment MinDate, MinValueMaxDate, MaxValue, Segment_Symbol (U,D,F), SegmentAvg
Until end of S is reached.

*The Prediction algorithm*

This algorithm is responsible for performing prediction operations leading to the prediction of future time series pattern.

1. Start
2. Enter new set of data
3. Call normalization sub-algorithm to normalize the data
4. Detect pattern of the new set of data (pat)
5. Calculate its average (avg)
6. Open the patterns turning point database table
7. Find match for average (avg) and pattern (pat) in the database
8. If match is found, get the next pattern ('np') in the database
9. If match is not found, then increment or decrement average (avg) until a match is found, get 'np'
10. Display np as predicted future pattern
11. return

## IV. RESULTS AND DISCUSSION

The model cum algorithm was tested with real-valued discrete univariate time series data, mostly stock market data, obtained online from Nigerian Stock Exchange (NSE) and Yahoo websites. The algorithm achieved 100% success in detecting and extracting the three pre-defined patterns. Out of the 15 different datasets used for experimental prediction, the system was able to predict 11 correctly and missed 4. Thus, it achieved 73% success, which is an impressive and acceptable outcome.Below are some of the outputs from the system.

**Table 1:** Sample raw and normalized time series data. It has 5158 records (data points).

| Record No | Date | Value | Normalised Value | Year |
|---|---|---|---|---|
| 1 | 2000-01-03 | 11357.5097656250 | 0.8104355931 | 2000 |
| 2 | 2000-01-04 | 10997.9296875000 | 0.6239265800 | 2000 |
| 3 | 2000-01-05 | 11122.6503906250 | 0.6886174083 | 2000 |
| 4 | 2000-01-06 | 11253.2597656250 | 0.7563626170 | 2000 |
| 5 | 2000-01-07 | 11522.5595703125 | 0.8960445523 | 2000 |
| 6 | 2000-01-10 | 11572.2001953125 | 0.9217924476 | 2000 |
| 7 | 2000-01-11 | 11511.0800781250 | 0.8900903463 | 2000 |
| 8 | 2000-01-12 | 11551.0996093750 | 0.9108479023 | 2000 |
| 9 | 2000-01-13 | 11582.4296875000 | 0.9270983338 | 2000 |
| 10 | 2000-01-14 | 11722.9804687500 | 1.0000000000 | 2000 |
| 11 | 2000-01-18 | 11560.7197265625 | 0.9158377051 | 2000 |
| 12 | 2000-01-19 | 11489.3603515625 | 0.8788245916 | 2000 |

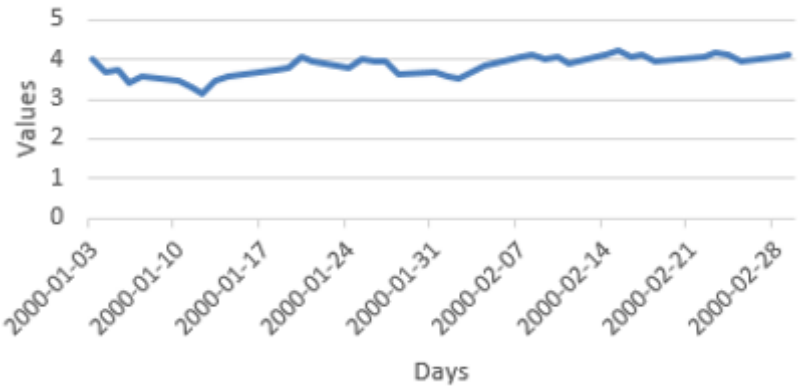


**Figure 3**: This shows the raw data plot without patterns extracted for two (2) months

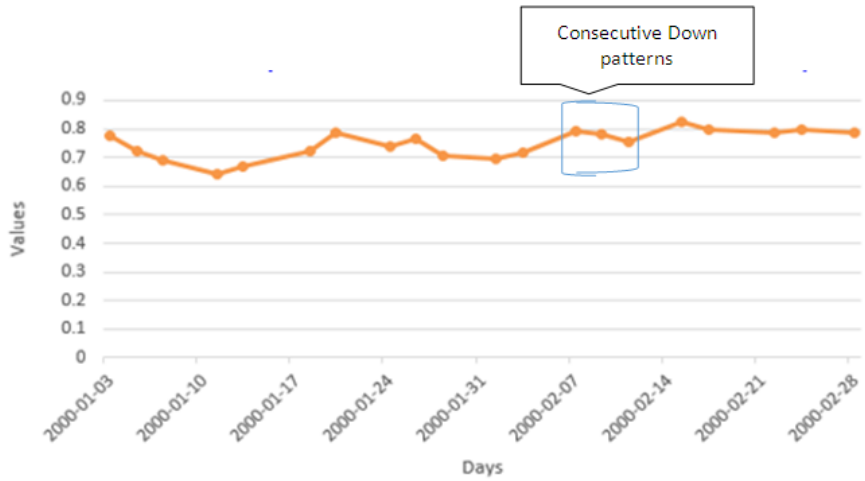


**Figure 4:** This shows the normalized data plot of the extracted patterns for the same 2 months.

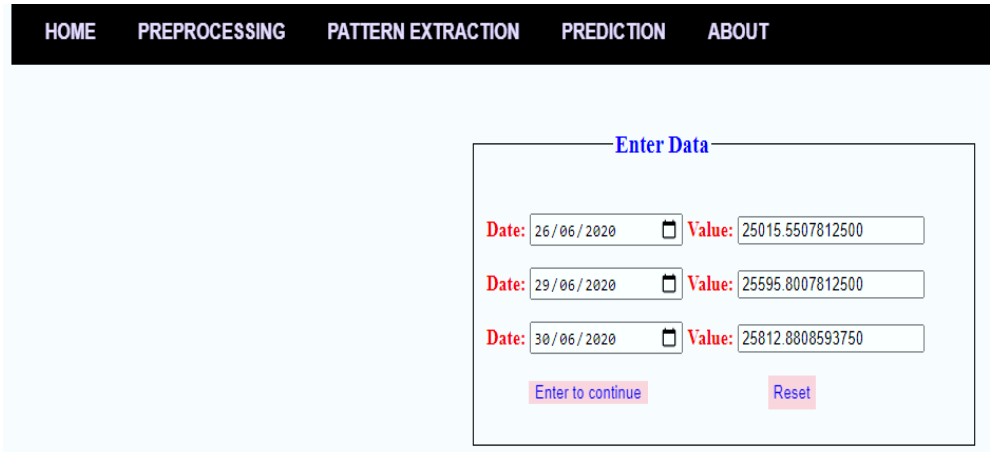


Figure 5: This is the pattern prediction window.

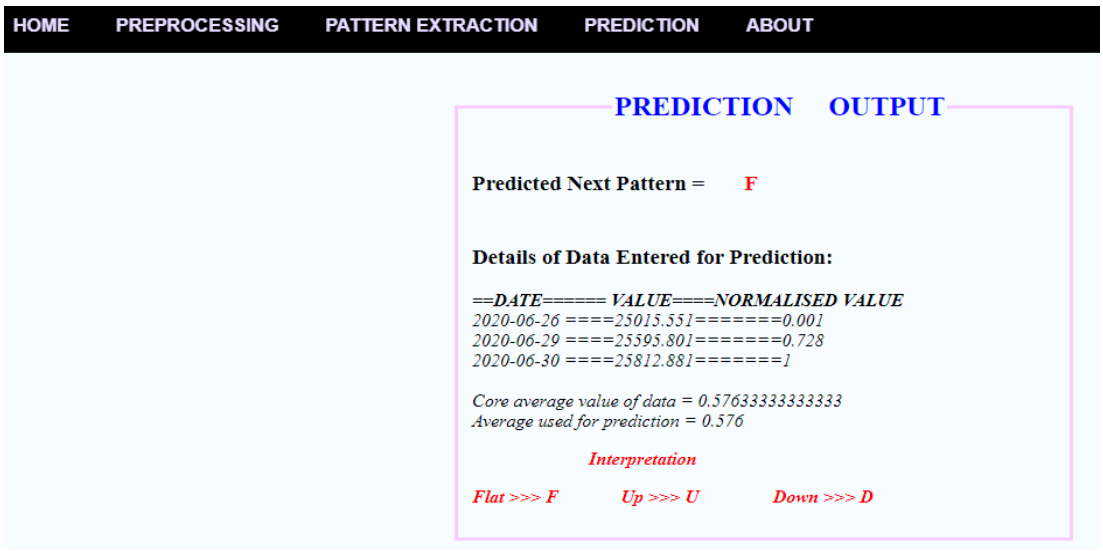


Figure 6: The prediction output from the system.

Table 1 presented a portion of the historical dataset, which has 5158 records (data points). The table showed both the raw and normalized values and their timestamps (date and year). As already noted, figure 3 presented the raw data plotted without pattern extracted for two (2) months; while figure 4 presented thenormalized data plot of the extracted patterns for the same 2 months. In comparing figure 3 and 4, the patterns in figure 4 looked straight lines in Up, Down and Flat positions. Furthermore, figure 4 conspicuously

showed patterns in the series, unlike in figure 3 which showed everything as a curve with no indication of where there is an occurrence of a pattern.

Figure 5 shows the pattern prediction window where the user has to enter current values needed to predict the pattern of the next three (3) days. Finally, Figure 6 shows prediction output from the system. The data for the predicted next three days were used to validate the output of the system. As recorded in our experiments, the system correctly predicted 11 out of the 15 datasets used in the prediction experiment.

## V. SUMMARY AND CONCLUSION

Time series analysis cuts across the following activities: time series representation, pattern recognition, similarity search and prediction. This work explored development of time series pattern recognition and prediction algorithms from the data mining perspective, and successfully developed an algorithm to mine time series datasets for patterns and also predict future patterns using symbols (U, D, F) for Up, Down and Flat patterns respectively. Our algorithms were very efficient, easy to understand and implement towards finding patterns in a time series and predicting future patterns. Researchers in time series analysis from different perspectives like statistics, economics and data mining will benefit immensely from the contributions of this work. Further research can be carried out to explore the applicability of the algorithms to other time series domains aside stock market data. In addition, consideration of multivariate time series datasets can also be explored.

## REFERENCES


[1]. Abdullah, M; Suman, N and Jie, L (2016). *Similarity Search on TS Data: Past, Present and Future*. CIKM2016 Tutorial, obtained from http://www.cs.unm.edu/~mueen/Tutorial/CIKM2016Tutorial.pdf on 17/4/2017.

[2]. Álvarez, F.M. (2010*) Pattern sequence analysis to forecast time series*. Unpublished thesis work. Universidad Pablo De Olavide De Sevilla

[3]. Babu, M.S; Geethanjali, N and Satyanarayana, B (2012). Clustering Approach to Stock Market Prediction. International Journal of Advanced Networking and Applications, Volume: 03, Issue: 04, Pages:1281-1291

[4]. Badhiye, S.S; Hatwar, K, S. and Chatur, P.N (2015). *Trend based Approach for Time Series Representation.* International Journal of Computer Applications, Volume 113, No. 16, (0975 – 8887).

[5]. Bishop, C.M (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer Science+Business Media, LLC.

[6]. Burcu, K; Serhan, O and Bora K (2011). Application of Symbolic Piecewise Aggregate Approximation (PAA) Analysis To ECG Signals. Artificial Intelligence & Design Lab, Mechanical Engineering Department, Computer Engineering Department, Izmir Institute of Technology, 35430 Izmir, Turkey. A pdf file obtained 15/4/2017

[7]. Ding, H; Hui D.; Goce T.; Scheuermann P.; Xiaoyue W., and Keogh E. (2008) " *Querying and Mining of Time Series Data: experimental comparison of representations and distance measures"*. Proceedings of the VLDB Endowment VLDB Endowment, Volume 1 Issue 2, pp 1542-1551.

[8]. Han, J. and Kamber, M. (2001). *Data mining concepts and techniques*. San Francisco: Morgan Kaufman

[9]. Jiangling, Y; Yain-Whar, Si and Zhiguo, G (2011). *Financial Time Series Segmentation Based On Turning Points*. Proceedings of 2011 International Conference on System Science and Engineering, Macau, China - June 2011.

[10]. Jingpei, D; Weiren, S; Fangyan, D and Kaoru, H (2013). Piecewise Trend Approximation: A Ratio-Based Time Series Representation. Journal of Abstract and Applied Analysis, Volume 2013. http://dx.doi.org/10.1155/2013/603629

[11]. Keogh, E. (2010) *Data Mining Time Series Data*, in Lovrić (Ed.), International Encyclopaedia of Statistical Science. New York, USA: Springer

[12]. Keogh, E. (2007). *Mining Shape and Time Series Databases with Symbolic Representations.* SIGKDD 2007 Tutorial.

[13]. Keogh, E., Chakrabarti, K., Pazzani, M. &Mehrotra, S. (2001). *Locally adaptive dimensionality reduction for indexing large time series databases*. In proceedings of ACM SIGMOD Conference on Management of Data. Santa Barbara, CA, May 21-24. pp 151-162.

[14]. Keogh, E.J; Chu, S; Hart, D and Pazzani, M.J (2001). *An Online Algorithm for Segmenting Time Series*. In ICDM Proceedings of the IEEE International Conference on Data Mining. Washington DC, USA: IEEE Computer Society, pp. 289 - 296.

[15]. Keogh E., S. Lonardi, Ratanamabatana C.A (2001). *Towards parameter-free data mining*. In proceedings of Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining

[16]. Kimoto, T.; Asakawa, K.; Yoda, M. and Takeoka, M., (1990) Stock market prediction system with modular neural networks. Proceedings of the IEEE International Joint Conference on Neural Networks, 1990, pp. 11-16.

[17]. Kuo-Ping, W; Yung-Piao, W and Hahn-Ming, L. (2014). Stock Trend Prediction by Using K-Means and AprioriAll Algorithm for Sequential Chart Pattern Mining. Journal Of Information Science And Engineering No 30, pp.653-667

[18]. Lee A. J. T; Lin M.-C.; Kao R.-T.; and Chen K.-T. (2010). An Effective Clustering Approach to Stock Market Prediction. PACIS 2010 Proceedings. Paper 54.

[19]. Lin, J; Keogh, E; Lonardi, S and Patel, P. (2002) *Finding Motifs in Time Series.* ACM SIGKDD, July 23-26,Edmonton, Alberta, Canada.

[20]. Nguyen, Q.V.H and Duong, T.A (2007). *Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Series.* International Symposium on Information Technology Convergence, IEEE Computer Society.

[21]. Pohl, D and Bouchachia, A (2012). *Financial Time Series Processing: A Roadmap of Online and Offline Methods*. A pdf file downloaded on April 5, 2017.

[22]. Prasanna, S. and Ezhilmaran, D. (2013). *An analysis on Stock Market Prediction using Data Mining Techniques*; International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 4 No. 02; p. 49-51.

[23]. Raj, M.P; Swaminarayan, P.R; Saini, J.R and Parmar, D.k (2015). *Applications of Pattern Recognition Algorithms in Agriculture: A Review.* Int. J. of Advanced Networking and Applications Volume: 6 Issue: 5 Pp: 2495-2502

[24]. Robert K.L; Chin-Yuan F.; Wei-Hsiu H and Pei-Chann C (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. An International Journal of Expert Systems with Applications, Vol.36, No.2, pp. 3761-3773

[25]. Senthamarai, K.K.; Sailapathi, P.S; Mohamed, M.S andArumugam, P. (2012) Financial Stock Market Forecast using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12
[26]. Singh, S (2000). *Pattern Modelling in Time-Series Forecasting*. Cybernetics and Systems - An International Journal, vol. 31, issue 1
[27]. Trippi, R.R. and Desieno D (1992) Trading equity index futures with a neural network. The Journal of Portfolio Management, Vol 5 (3)
[28]. Wu, H; Salzberg, B and Zhang, D (2004). *Online event-driven subsequence matching over financial data streams*, in Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, Paris pp. 23-34.
[29]. Zhang, Z; Jiang, J; Liu, X; Lau, W.C; Wang, H; Wang, S; Song, X and Xu, D (2010). *Pattern recognition in stock data based on a new segmentation algorithm*, in Lecture Notes in Computer Science, pp. 520-525, Springer Berlin / Heidelberg.