•

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

ANALYSIS OF STOCK MARKET PREDICTION USING VARIOUS MACHINE LEARNING TECHNIQUES

Vilas Alagdeve[1*], P T Karule[2], Mamani Bandopadhyay[3], Seema P Nehete[4], Samuel Biswas[5], Sanjay Shahaji Kadam[6], Subhadip Goswami[7]

[1*]Assistant Professor, Yeshwantrao Chavan College of Engineering, India
[2]Professor, Yeshwantrao Chavan College of Engineering, India
[3]Assistant Professor, Brainware University, India
[4]Assistant Professor, Datta Meghe College of Engineering, India
[5]Assistant Professor, Haldia Institute of Management, India
[6]Assistant Professor, Bharati Vidyapeeth College of Engineering, India
[7]Assistant Professor, Bharat Institute of Engineering & Technology, India
Email:[1*]vilas.alag@gmail.com; [2]ptkarule@gmail.com; [3]mamani.banerjee1985@gmail.com; [4]seema.nehete@dmce.ac.in; [5]sam_biswas@gmail.com; [6]sanjaykadam23@gmail.com; [7]mrrnt007@gmail.com

Abstract

The **objective** is to compare machine learning models (SVM, Random Forest, ANN, Naive Bayes) for stock price prediction using historical data, assessing accuracy under different input methods for reliable forecasting. This study compares two **methods** of input data: one using ten technical characteristics (open, high, low, close prices) and the other using trend-deterministic data. Four prediction models—SVM, Random Forest, ANN, and Naive Bayes—are tested on ten years of historical data (2003-2012) from Reliance Industries, Infosys Ltd., CNX Nifty, and S&P BSE Sensex. The models' accuracy is assessed, and the efficacy of regression analysis is explored for improving forecasts. The study **finds** that forecasting stock price movements in 23 Indian stock markets remains challenging due to high risk and volatility. It compares two data input methods—one using technical characteristics (open, high, low, close prices) and another using trend-deterministic data—across four prediction models: Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), and Naive Bayes. Historical data from 2003-2012, covering Reliance Industries, Infosys Ltd., CNX Nifty, and S&P BSE Sensex, is used to assess the models' accuracy. The results indicate that traditional methods like technical and fundamental analysis are not always reliable, and regression analysis may be enhanced by incorporating more factors. ANN is highlighted as a promising approach for stock market prediction, though it requires further refinement. The study emphasizes the need for more advanced and diverse techniques from computer science and economics to improve stock price forecasting, while also recommending areas for future research in this field. The study's **novelty** lies in comparing advanced prediction models (SVM, Random Forest, ANN, Naive Bayes) using both technical and trend-deterministic data, highlighting the potential of ANN for improving stock forecasts.

Keywords: Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), Naive-Bayes

1. Introduction

The uncertainties involved make it difficult to predict stocks and stock price indices. Before purchasing a stock, investors conduct two different kinds of analysis. The fundamental analysis comes first. To determine whether or not to invest in this, investors consider factors such as the inherent worth of stocks, the state of the economy and industry, the political environment, etc. Technical analysis, on the other

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

hand, involves analyzing market activity-generated facts, like previous prices and volume, to evaluate equities. Instead of attempting to determine a security's fundamental value, technical analysts study stock charts to spot patterns and trends that could indicate how a stock will perform going forward. According to Malkiel and Fama's (1970) efficient market theory, stock prices may be predicted using transaction data because they are informational efficient [1]. This makes sense because a lot of erratic variables, such as the nation's political climate and the company's reputation, will begin to affect stock prices. According to the experimental results, when the technical parameters in the first method are represented as continuous values, the Random Forest model performs better than the other three models. Furthermore, when the technical parameters are given as trend-deterministic data, all models perform better. Because the stock market is by its very nature nonlinear, making predictions about it is difficult and demanding. Numerous methods have been developed over the years to forecast stock trends. Originally, stock trends were predicted using traditional regression techniques. Non-stationary time series data, such as stock data, can be analyzed using non-linear machine learning algorithms. For stock and stock price index movement prediction, two machine learning techniques that are most frequently employed are Artificial Neural Networks (ANN) and Support Vector Machines (SVM). By building a network of neurons ANN mimics how our brains function to learn. In order to predict the behavior of financial markets, Hassan, Nath, and Kirley (2007) devised and implemented a fusion model that included the Hidden Markov Model (HMM), Artificial Neural Networks (ANN), and Genetic Algorithms (GA) [2]. The daily stock prices are converted into separate sets of values using ANN, which are then fed into an HMM.A prediction system that was helpful in predicting the Taiwan stock market's mid-term price trend was created by Wang and Leu in 1996. Vapnik (1999) created the well-known SVM method that looks for a hyper plane in a higher dimension to divide classes. Support vector machines, or SVMs, are a particularly particular class of learning algorithms distinguished by the use of kernel functions, capacity control of the decision function, and solution scarcity [3]. Huang, Nakamori, and Wang (2005) used SVM to estimate the weekly movement direction of the NIKKEI 225 index in order to study the predictability of financial movement direction. SVM was contrasted with Elman Back propagation Neural Networks, Linear Discriminant Analysis, and Quadratic Discriminant Analysis [4]. The experiment's findings demonstrated that SVM performed better than the other classification techniques. Kim (2003) employed SVM to forecast the direction of the Korea Composite Stock Price Index (KOSPI) daily stock price movement [5]. The first set of qualities consisted of twelve technical indications. In this study, SVM was compared with case-based reasoning (CBR) and back-propagation neural networks (BPN). The experimental data clearly showed that SVM performed better than BPN and CBR. The stock market is volatile and non-linear by nature. Stock market predictions have been made using a variety of conventional and statistical techniques. The most often used techniques involve combining various learning algorithms with artificial neural networks (ANNs). An artificial neural network's (ANN) capacity to learn from and generalize from non-linear data trends makes it a good fit for problem domains like stock market prediction. Better prediction accuracy is achieved by the ANN than by the conventional approach because of its ability to adapt to the data pattern and relationship between the input and output.

2. Classical Stock Market Estimation

To forecast the closing price or the moving price of the stock market, numerous conventional techniques have been used. Two key theories—the random walk theory [8] and the efficient market hypotheses (EMH) [7]—are employed in the traditional approach to stock market prediction. In 1964, Fama presented the efficient market hypothesis [7]. Based on historical stock data, EMH hypotheses state that future stock prices are unpredictable. The unbalanced stock is immediately detected when fresh information enters the system and is promptly removed by the appropriate adjustment in price [6]. There are three variations of the EMH: weak, semi-strong, and powerful [7]. Historical data are employed in weak EMH stock price prediction. All available information—historical, public, and private—including insider knowledge—is utilized in the strong EMH to forecast stock prices. However, according to the random walk hypothesis, stock prices are independent of previous stock performance [9]. Because the past data does not match the present stock price trend, these patterns should not be taken advantage of. Technical analysis and fundamental analysis are two traditional methods for stock market forecasting [4]. A numerical time series method called technical analysis is used to forecast stock markets utilizing charts to make predictions based on past data the principal instrument [10]. This method seeks to extract data. Time series mining is the process of identifying a trend by analyzing historical data [6]. The study of the variables influencing supply and demand is known as fundamental analysis [11]. According to the basic analysis, the primary method for predicting the stock price is the collection and interpretation of information. This analysis's trading opportunity takes advantage of the lag time between an event's occurrence and the market's reaction to it. The economic data of businesses (such as annual and quarterly reports), auditor's reports, balance sheets, and income statements are crucial pieces of information utilized in fundamental analysis. Since news also captures the present supply and demand dynamics in the market, news is important for fundamental research as well. Due to the rise in computational power, which allows computers to analyze larger data sets more correctly in less time, these traditional methodologies are currently becoming less effective. Nonetheless, these methodologies continue to serve as the foundation for novel approaches to artificial intelligence, including machine learning and computational intelligence.

Predicting the swings of the stock market index is essential for the creation of profitable market trading strategies. Traders can choose whether to buy or sell an item by choosing an effective forecasting model. Investors may benefit from accurate stock market index movement prediction. Forecasting changes in the Stock Market Index is a very difficult and complex task. In order to predict stock prices for a company with a higher level of accuracy and dependability, this study applied machine learning techniques. The primary contribution of the experts was the incorporation of the ELR-ML Model as a stock price calculation mechanism.

3. Literature Survey

Because of its learning, mapping, generalizing, and self-organizing qualities, NN has been demonstrated to be able to forecast volatility and non-linear stock market values. This segment will showcase a selection of the projects completed by academics that use and deploy neural networks in financial stocks forecasting the market. The most often utilized NN architecture is feed forward NN in predicting the stock market. This is the most basic NN, where the Feed forward NN information only flows in one direction. The data travels from the input layer to one or more hidden layers, and then from those layers to the output layer. [12] used a three-layer feed forward neural network, comprising an input layer, a hidden layer, and an output layer, to forecast the price of IBM's common stock on a daily basis. The study employed a dataset of 5000 days to carry out its experiment. One thousand days of the 5000 days of data were used for training, while the remaining days were used for testing. Although the NN's performance is not up to par, they did offer insightful advice on how to use NN for stock market prediction. A NN model was used in [13] to forecast the closing value of the Indian S&P CNX Nifty 50 Index. The study examined ten-year data sets of the S&P CNX Nifty 50 Index closing values, spanning from January 1, 2000, to December 31, 2009. Of the ten years of data, four years were used for validation. Stock market prediction using Artificial Neural Networks (ANN), Support Vector Machines (SVM), and regression techniques involves applying sophisticated machine learning methods to forecast future stock prices or trends. Artificial Neural Networks are designed to mimic the human brain's ability to learn and recognize patterns, making them well-suited for capturing the complex, non-linear

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

relationships inherent in financial markets. By training on historical stock data, ANN models can identify patterns that might indicate future price movements. Support Vector Machines, particularly in their regression form (Support Vector Regression, SVR), are used to find the optimal boundaries that separate data points, aiming to create a function that closely follows the data with minimal deviation. This technique is effective in handling high-dimensional data and can be particularly useful in making predictions where the relationship between input features and stock prices is not straightforward. Finally, regression techniques, including linear and non-linear regression models, are employed to establish a relationship between independent variables (such as economic indicators, technical indicators, and company fundamentals) and the stock price. These models attempt to predict future prices based on the relationships learned from past data. Combining these techniques can create a more robust and accurate prediction model, leveraging the strengths of each method to better understand and forecast stock market movements.

4. Prediction Model

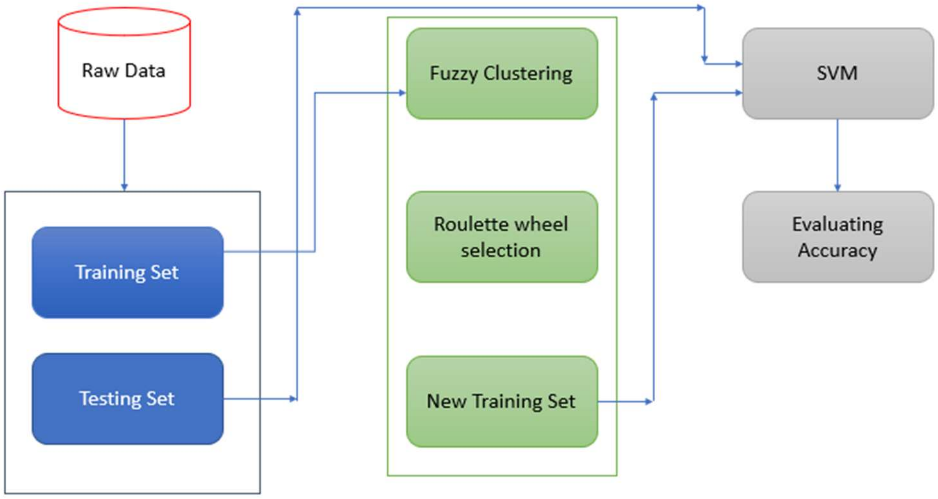

Figure 1. A demonstration of the novel fuzzy selection support vector machine (SVMFS) approach.

4.1 ANN Model:

Artificial Neural Networks (ANNs) are a dense network of interconnected neurons that are activated in response to inputs. They are inspired by the way biological neural networks function. In this study, we use a three layer feed-forward neural network. Ten technical indicators, each represented by ten neurons in the input layer, serve as the network's inputs. A single neuron in the output layer uses the log sigmoid transfer function. A continuous value output between 0 and 1 is the outcome of this. To anticipate whether the movement will be up or down, a threshold of 0.5 is employed. Prediction is regarded as the upward movement if the output value is more than or equal to 0.5, and the downward movement otherwise. Tan sigmoid was used by every neuron in the hidden layer as the transfer function. The weights are updated using a gradient descent with momentum method, whereby weights are adjusted at each epoch to enable the achievement of a global minimum. To establish parameters for every stock and index, we have carried out extensive parameter setting tests. The momentum constant (MC), number of epochs (EP), value of learning rate (LR), and number of hidden layer neurons (N) make up the parameters of the ANN model. The parameter setting experiments test 10 levels of n, nine levels of

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

mc, and ten levels of EP in order to efficiently determine them. The value of LR is initially set at 0.1. For a total of 3600 ANN treatments, two indices and two stocks are taken into consideration. For comparison experiments on a comparison data set, the top three ANN models are determined by combining the parameters in the best way possible to achieve the best average of training and holdout performances. The learning rate (LR) for these top-performing models is adjusted within the interval [0.1, 0.9]. Researchers offer guidelines to help and direct others in conducting ANN forecasting experiments. The output and the outcomes are validated during the first step of data preparation.

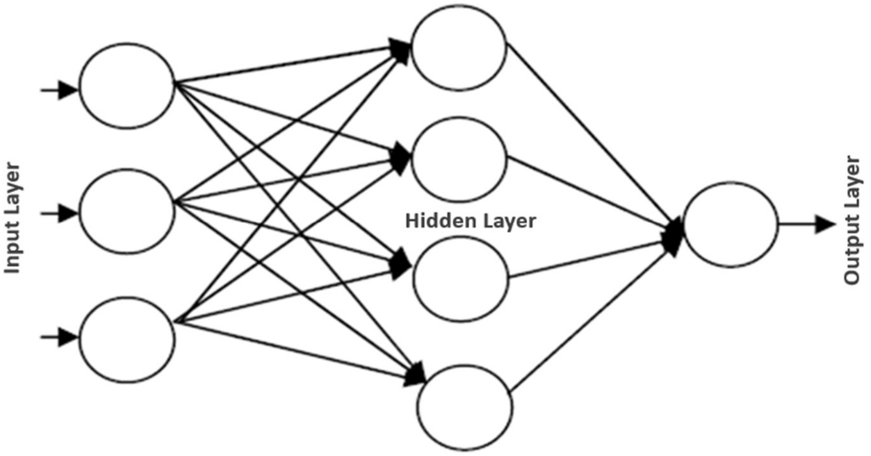

Figure 2. ANN Architecture and Multi-Layer Perceptron (MLP)

Numerous ANN architectures, including feed forward, recurrent, and spiking NN, have been constructed. Furthermore, there are other kinds of neural networks (NNs), including temporal NNs, radial basis function NNs, self-organizing NNs, single-layer NNs, and Multi-Layer Perceptrons (MLPs) [14]. Natural language processing (NNs) has been evaluated and used in a wide range of research and applications, such as disease detection [15][16], speech recognition [17][18], data mining [19][20], image processing [21][22], forecasting, robot control [23][24], and many more.

Marcus O' Connor and William Remus put forward the first guide in [25]. The following are the steps to follow:

1. Before estimating the NN model, clean up the data.
2. Before estimating the model, scale and de-seasonalize the data. Normalization of the raw data used as the ANN's input is accomplished through scaling. To analyse non-seasonal trends in the data, de-seasonalization involves removing the seasonal component from the time series.
3. Select the ideal beginning point by using the relevant techniques.
4. Employ specific techniques to steer clear of local optima.
5. Keep growing the network until the fit doesn't significantly improve.
6. When analyze and estimating neural networks, utilize holdout samples and pruning procedures.
7. Be sure to choose software with built-in features to prevent the drawbacks of NN.
8. Create believable neural networks and shrink them to make the model more palatable.
9. Employ additional methods to confirm the validity of the NN model.

4.2 SVM Model:

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

It was Vapnik (1999) who originally invented support vector machines (SVMs) [3]. Support vector 400 machines fall into two basic categories: support vector regression (SVR) and support vector classification (SVC). SVM uses a high dimensional feature space as part of its learning algorithm. Points in SVM are classified by allocating them to one of two disjoint half spaces, either in the pattern space or in a higher-dimensional feature space, according to Khemchandani and Chandra (2009) [26].

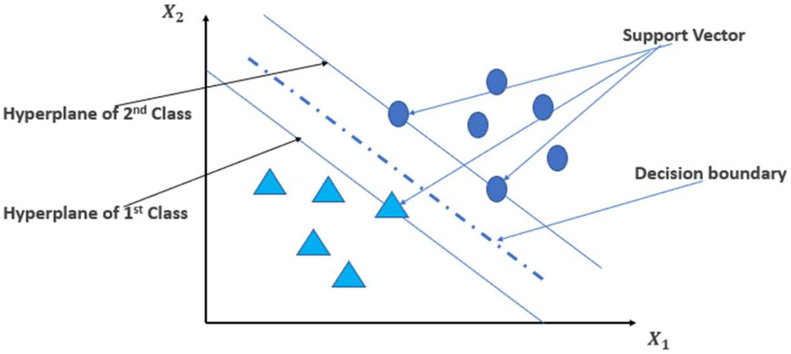

Figure 3. Maximum margin of the separating hyperplane is determined by support vectors.

The 407 maximum margin hyper plane identification is the primary goal of the support vector machine. According to Xu, Zhou, and Wang (2009) [27], the goal is to maximize the margin of difference between positive and negative examples. The ultimate decision boundary is determined as the greatest margin hyper plane. Assume that $a_i \in L^d, i = 1, 2, \dots N$ generates a set of input vectors with class labels $b_i \in \{+1, -1\}L^d, i = 1, 2, \dots N$. SVM maps the input vectors $a_i \in L^d$ into a high dimensional feature space $\theta(a_i) \in K$. A kernel function $G(a_i b_j)$ performs the mapping $\theta(.)$. The resulting decision boundary can be defined as follows:

$$g(a) = \operatorname{sgn}\left(\sum_{i=1}^{N} b_i \beta_i \,.\, G(a, a_i) + c\right)$$

The normalization parameter c regulates the trade-off between margin and misclassification error. We employ the radial basis kernel function and the polynomial one. These parameter choices result in 20 and 40 SVM treatments using polynomial and radial basis kernel functions, respectively, for a single stock. Two indices and two equities are taken into consideration for a total of 240 SVM treatments. The best average of training and holdout performances for each of the polynomial kernel SVM and radial basis kernel SVM parameter combinations are chosen as the top two SVM models for comparison studies.

4.3 Regression:

Regression analysis is a statistical method used in statistics to estimate the relationships between variables. When examining the relationship between a dependent variable and one or more independent variables, it encompasses a variety of modelling and analysis methodologies. More precisely, regression analysis clarifies how changes in any one of the independent variables, while holding the other independent variables constant, affect the dependent variable's usual value. Regression analysis also involves characterizing the variation of the dependent variable around the regression function, which can be explained by a probability distribution. There is a lot of overlap between the usage of regression

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

analysis and machine learning in the prediction and forecasting domains. Regression analysis can also be used to investigate the types of relationships between independent factors and dependent variables, as well as to identify which variables are associated to which one.

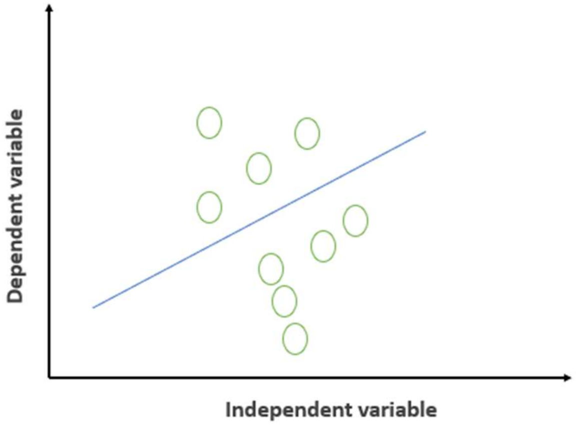

Figure 4. Stock market linear regression model.

The first kind of regression analysis to be thoroughly researched and applied in a wide range of real-world scenarios was linear regression [28] This is due to the fact that models with linear dependence on unknown parameters are simpler to fit than those with non-linear parameter dependence, and it is also simpler to ascertain the statistical characteristics of the resulting estimators. Although the least squares approach is frequently used to fit linear regression models, there are other approaches as well. For example, one can minimize the "lack of fit" in a different norm (as with least absolute deviations regression) or minimize a penalized version of the least squares loss function (as in ridge regression, which uses the L2-norm penalty, and lasso, which uses the L1-norm penalty). On the other hand, models that are not linear can be fitted using the least squares method. As a result, despite their close relationship, the phrases "least squares" and "linear model" are not interchangeable.

Thanks to computers' increased speed and accuracy in processing vast volumes of data, machine learning (ML) enables investors to anticipate stock market price indices more accurately than they could using conventional methods. Numerous machine learning techniques are capable of producing highly accurate stock market predictions. To increase accuracy, some suggest creating unique machine learning models. While SVM and ANN are also utilized, NNs make up the majority of ML models. It is our goal that these results will further the understanding of researchers regarding potential methods for stock market prediction.

5. Conclusion

The application of artificial neural networks (ANN) to stock market prediction is reviewed in this paper. While NN has produced results that are considered acceptable, many researchers are currently working to increase the accuracy of stock market prediction by utilizing hybrid methods and accounting for a greater number of external factors. It is well-known that the dynamic world of the stock market is non-linear, volatile, and susceptible to numerous external influences. This paper can serve as a starting point for individuals who are interested in working with NN for stock market prediction. Because the stock market is so volatile, making predictions is a crucial part of the highly difficult and complex process of working in the stock market. According to this study, NN is the most generally utilized approach, and its difference from other ways is rather considerable. It is surprising to learn that, with four publications apiece, 2015 and 2016 have the most research papers discussing NNs. In comparison, there are just three papers mentioning NNs in 2013, the year with the most published papers in this study. One of the researchers' predictions of the Shanghai Stock Exchange's closing price in the least amount of time, along with the daily highest and lowest prices, demonstrate NN's efficacy in stock market prediction. Additional study may benefit from the use of the NN model. The second-highest strategy to be employed is SVM.

6. References

1. Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work⁄. The Journal of Finance, 25, 383–417.
2. Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. Expert Systems with Applications, 33, 171–180.
3. Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10, 988–999.
4. Huang,W., Nakamori, Y., &Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32, 2513–2522.
5. Kim, K.-j. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55, 307–319.
6. S. Sneha, "Applications of ANNs in Stock Market Prediction: A Survey," International Journal of Computer Science & Engineering Technology, vol. 2(3), 2011.
7. E. F. Fama, "The Distribution of the Daily Differences of the Logarithms of Stock Prices," Unpublished Ph.D Dissertation, University of Chicago, 1964.
8. F. Pegah, "Stock trend prediction using news articles a text mining approach," Master thesis, Luleå University of Technology, 2007.
9. E. F. Fama, "Random Walks In Stock Market Prices," Financial Analysts Journal, vol. 21(5), pp. 55–59, 1965.
10. M. J. Pring, "Technical Analysis Explained," New York (NY), McGraw-Hill, 1991.
11. M. C. Thomsett, "Mastering Fundamental Analysis," Chicago: Dearborn Publishing, 1998.
12. H. White, "Economic prediction using neural networks: the case of IBM daily stock returns," IEEE International Conference on Neural Networks, vol. 2, pp. 451-458, 1988.
13. M. Majumder, and A. Hussian, "Forecasting of Indian Stock Market Index Using Artificial Neural Network".
14. C. Peterson, and T. Rögnvaldsson, "An introduction to artificial neural networks," Proc. 1991 CERN Summer School of Computing, CERN Yellow Report 92-02, pp. 113-170, 1992.
15. K. A-S. Qeethara, "Artificial Neural Networks in Medical Diagnosis, " International Journal of Computer Science Issues, vol. 8(2), 2011.
16. Y. K. Irfan, P. H. Zope, and S. R. Suralkar, "Importance of Artificial Neural Network in Medical Diagnosis disease like acute nephritis disease and heart disease," International Journal of Engineering Science and Innovative Technology (IJESIT), vol 2(2), 2013.
17. W. Chris, "Introduction to Speech Recognition Using Neural Networks," European Symposium on Artificial Neural Networks, 1998.
18. N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and Speaker Recognition System Using Artificial Neural Networks and Hidden Markov Model," International Conference on Communication Systems and Network Technologies, pp. 311-315, 2012.
19. Y. Singh, and A.S. Chauhan, "Neural networks in data mining," J. Theo. and App. Inf. Tech, vol. 5, pp. 37-42, 2009.
20. S. Nirkhi, "Potential use of Artificial Neural Network in Data Mining," International Conference on The 2nd Computer and Automation Engineering (ICCAE), 2010, vol. 2, pp. 339-343, 2010.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

21.]M. Egmont-Petersen, D. de Ridder, and H. Handels, “Image processing with neural networks, a review,” Pattern Recognition, vol. 35, pp. 2279–2301, 2002.
22. J. A. Ramírez-Quintana, M. I. Chacon-Murguia, and J. F. Chacon-Hinojos, “Artificial Neural Image Processing Applications: A Survey,” Engineering Letters, vol. 20(1), 2012.
23. K.-O. Chin, J. Teo, and S. Azali, “Artificial Neural Controller Synthesis In Autonomous Mobile Cognition,” International Journal of Computer Science (IJCS). vol 36(4), 2009.
24. K.-O. Chin, and J. Teo, “Evolution and Analysis of Self-Synthesized Minimalist Neural Controllers for Collective Robotics using Pareto Multi-objective Optimization,” World Congress on Computational Intelligence, pp. 2172-2178, 2010.
25. W. Remus, and M. O'connor, “Neural Networks For Time Series Forecasting,” in Principles of Forecasting: A Handbook for Researchers and Practitioners, J. S. Armstrong, editor, Norwell, MA: Kluwer Academic Publishers, 2001.
26. Khemchandani, R., Chandra, S., et al. (2009). Knowledge based proximal support vector machines. European Journal of Operational Research, 195, 914–923.
27. Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. Expert Systems with Applications, 36, 2625–2632.
28. Larose, D. T. (2005), “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2,John Wiley & Sons, Inc.