A Comparative Study to Find an Optimal Algorithm for a stock Using Artificial Intelligence

Srikanth Satyala
Computer Science engineering
VIT-AP University
Guntur, India
srikanthsathyala@gmail.com

Devika T.S
Computer Science engineering
VIT-AP University
Guntur, India
devikasasi2019@gmail.com

Rayudu Vyshnavi Ram
Computer Science engineering
VIT-AP University
Guntur,
rayuduvyshnavi@gmail.com

Reeja S R
Computer Science engineering
VIT-AP University
Guntur, India
reeja.sr@vitap.ac.in

***Abstract*— Stock markets have different sectors in the market like the industrial sector, technology sector, finance sector, etc. different algorithms are used with analytical methods for efficient, profitable, and optimal trading using machine learning methods. Different mediums like stocks, cryptocurrency, NFT, etc can be traded as an asset. There are many types of trading algorithms like Mean Reversion, Factor-Based Investing ETF Rotation, Smart Beta, etc. and, applying Artificial intelligence models like Random Forest, Support Vector Machine, logistic regression, Naive Bayes, classification regression tree, and DNN models such as recurrent neural network, multilayer perceptron, deep belief network. Algorithmic trading is used for mainly short-term trading, but it can also be used for scalping. Both have some differences, so their respective algorithms are adjusted accordingly for better profits and smooth processing.**

Keywords*— *Artificial Intelligence, HFT, stocks, Short term trading, scalping, intraday trading

I. INTRODUCTION

Algorithmic trading is a pre-programmed method of running commands and trading instructions that are accounting for variables such as time, price, and volume. By applying artificial intelligence concepts to algorithms to get an OPTIMAL method that can return maximum profits. It can avoid small-scale man-made errors which can lead to a loss on a large scale and there will be no involvement of emotion which is a very risky and highly influencing factor in stocks. All the analytics are done using past data to make sure there are no losses or real money during the testing instead of real money.

There are many classification algorithms in artificial intelligence that can be classified based on trees, distance, probability, and neural networks. Every model has its pros and cons, that's why different types of classification models are compared to find which models are suitable for which market and find one that is suitable for all the sectors

For training, methods used are done in iterations. Every iteration model gets trained with artificial intelligence methods such as NB(Naive Bayes Classification), RF(Random Forest), MLP(Multilayer Perceptron), LR(Logistic Regression), CART(Classification and Regression Trees), KNN(K Nearest Neighbour), XGB(XG Boost), RNN(Recurrent Neural Network), SVM(Support Vector Machine) [10] and [11], DBN(Deep Belief Networks)[8] and [13] are used in different sectors. The annual return of rate (ARR), Winning ratio (WR), etc. as scaling factors to measure which gives the best profits and is optimal for each sector.

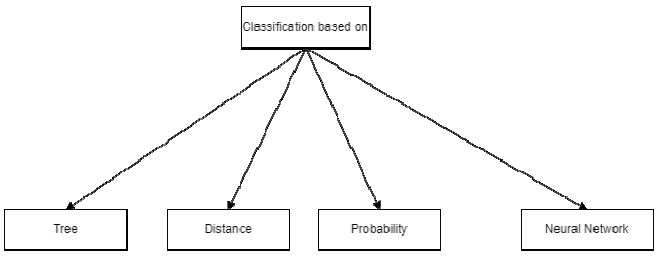

Fig. 1 multiple methods classification

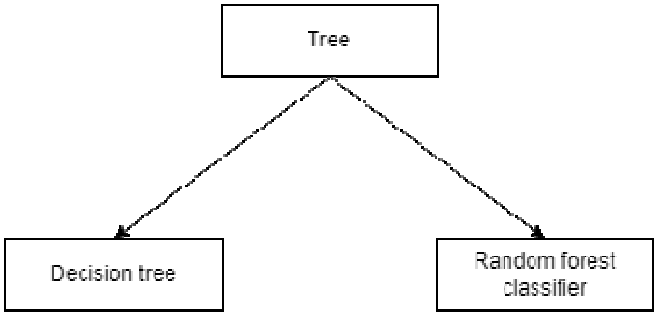

Fig. 2 Classified based on tree

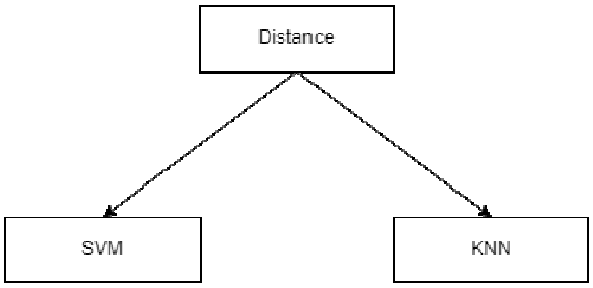

Fig. 3 Classified based on Distance

978-1-6654-9790-9/22/$31.00 ©2022 IEEE

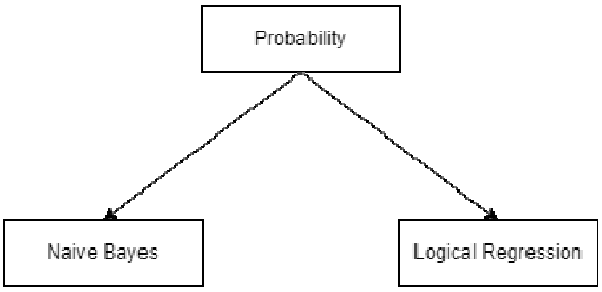

Fig. 4 Classified based on probability

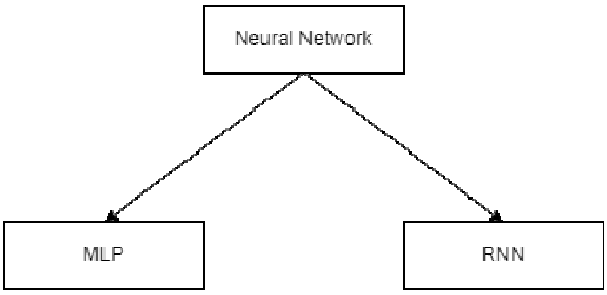

Fig. 5 Classified based on Nural Network

Based on the classification from Fig. 1, By further adding multiple methods that come under the classification where Fig. 2 is classified based on tree, Fig. 3 is classified based on Distance, Fig. 4 is classified based on probability, Fig. 5 is classified based on Nural Network.

II. Related Work

The author [1] experimented on humans about the relation between man and economical decisions made using humans with no prior knowledge of economics using external expert machine agents which are computers explaining the human's rules and methodologies. Cons are not covered as it is starting to develop.

The paper [2] reviewed near zero intelligence and improved algorithms which are not very flexible overall markets and concluded, that GDX is better than ZIP which is modifying and applying different algorithms to different markets based on strategies makes it more flexible and can be used for high-frequency trading. In [3] Upgraded the algorithms so that they covered the cons of previous algorithms which are not so flexible to all types of markets. Not only making them fast, cost-effective, and increasing trade volumes they also gained frequency Strategies. They also predicted future possibilities of threats and opportunities.

The author [4] computerized trading as the advancement of technology as time passed added new rules, strategies towards more profit and very less time using real data. It also covered risks of trading using computers and algorithms at high frequencies and more automation that can be done. Made a milestone leading trading to a new level that applies to real-life data and making actual profits that made many researchers realize the potential of artificial agent trading. Researched high-frequency trading methods like simulating analysis of gains in [5]. Selected actions of securing high-frequency trading. Also, limited issues of research in the past which are mostly time-based and some applying optimization to existing algorithms with new methods.

In paper [6] expanded review on the relation between man and computer including rules, models, strategies, negotiation, and dealing. It also covers some unpublished work that had a competition between algorithms vs humans towards profitability. Also discussed is the relation between human psychology on algorithms and how computers can trade as investors in the market and the agent decides that algorithms trade by themselves without humans taking part. [7] compared 14 different classifications on 7 different sectors to find an optimized method to gain more profit on stocks. RNN (Fig. 6) is a tree network that mainly works on Memorization and prediction best example can be voice assistants like google voice assistant and Siri which notes the questions that are not in the knowledge base to predict what solution can be given when a similar type of question is asked again [15] and [16]. It goes into the in-depth concept of patterns in sequential data and memory. From Fig. 1 one thing can be observed, that it runs in loops.

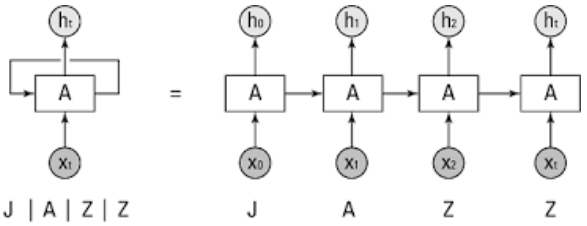

Fig. 6 (source: By Kevin Vu, Exact Corp)

Multilayer Perceptron (MLP Fig. 7) is not ideal for processing sequential data because it needs many parameters to process data. It has three layers: an input, an output, and a hidden layer. It is mostly applied for supervised learning.

III. Proposed Method

The evaluation, factors are used to train and measured to find the prediction's accuracy. Win Ratio (WR) in eqn.(1) is the number of winning trades/total number of trades used in the Kelly Criterion formula. Higher WR indicates good profit [9].

$$\text{Win Rate} = \frac{\#Won_{opp}}{\#Total_{opp}} \quad \text{…….. (1)}$$

The annual return rate (ARR) is the yearly rate calculated by money spent or earned at the end of the year divided by money invested at the start of the year in Eqn. (2). It also depends on the amount of time the stock period.

$$ARR = \left(\frac{End_{Val}}{Start_{Val}}\right)^{1/n} - 1 \quad \text{……. (2)}$$

SR is an evaluation measure in Eqn. (3). Suppose H is holding period and, m is number of periods. During *H* time, the annual return rate of the object is ARR_H, σ_H is the root-mean square deviation. Upon set R_f as zero,

$$ASR = \sqrt{m} * (ARR_h - R_f)/\sigma_h \quad \text{…….. (3)}$$

Drawdown is used to measure the highest "loss" vs the highest "profit". *MDD* represents highest decrease in value,

in the duration of τ, *Taking* D_t at any interval where r less than H in Eqn. (4)

$$D_t = max\left(0, maxp_{t_{ts(0,\tau}}\right);$$
$$MDD_h = maxD_{t_{ts(0,h)}}/maxp_{t_{ts(0,h)}} \quad \ldots\ldots\ldots (4)$$

In this equation p_t is the price at the time t; {ARR, ASR, MDD, and WR}. The above four factors are considered to compare optimality. These sectors are divided into 9 parts as the Fig. 7 below

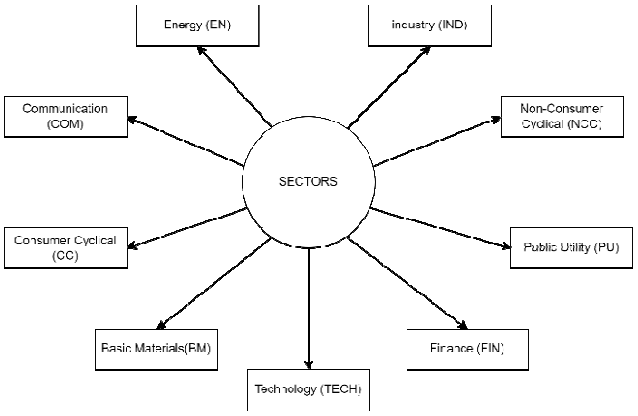

Fig. 7 The nine sectors

A. Data Set Environment

From the data collected we apply the following training, $D = \{(a1, b1), (a2, b2) \ldots., (an, bn)\}$, $ai = \{ai1, ai2 \ldots, ain\}$ is the input; n is count of total attribute; $bi = \{0,1\}$ which is classifier; $i = 1,2, \cdots, N$, where N is length of the sample. D is a data matrix of N*(n+1). The main goal of this training is to classify the labels. This paper uses the traditional Artificial intelligence methods like RF, SVM, BN, LR, XGB, CART, and DNN to forecast up and downs in price in the market. The training of these methods are done as represented in Table 1.

Table 1: Training methods with features and parameters

	Features	Label	Main parameters
LR	Matrix (250,44)	Matrix (250,1)	A specification for the model link function is logit
SVM	Matrix (250,44)	Matrix (250,1)	The kernel function used is Radial Basis kernel: Cost of constraints violation is L
CART	Matrix (250,44)	Matrix (250,1)	The maximum depth of any node of the final tree is 20 the splitting index can be Gini coefficient
RF	Matrix (250,44)	Matrix (250,1)	The Number of trees is 500. Number of variables randomly sampled as candidates at each split
BN	Matrix (250,44)	Matrix (250,1)	the prior probabilities of class membership are the class proportions for the training set.
XGB	Matrix (250,44)	Matrix (250,1)	The maximum depth of a tree is 10. the max number of iterations is 15 the learning rate is 0.3.

B. Learning Algorithm

The results are depicted using data of previous 100 days to predict the result of next 10 days. For 1,000 days of trading, it requires (1000–100) / 10 = 90 sessions for training the data to get 900 trading signal predictions. It is used to generate signals for trading as indicated in Algorithm 1.

Algorithm 1: signals for trading

```
1. N = length of Stock Code List #424 SPICS, and 185 CSICS. N = 424, 185.
2. L = Number of Samples #L = 2000
3. P = Length of Features #P = 44
4. k = length of Training Dataset #k = 250
5. n = Length of Testing Dataset/Length of Walk-Forward Window #n = 5
6. for i in 1:N
7.    Stock Data = SCLEI[i]
8.    M = (L-k)/n
9.    Trading Signal0 = NULL
10.   for j in 1:M
11.      New_Data = Stock Data[(k+n*(j-1)):(k+n+n*(j-1)), ]
12.      New_Train = New_Data[1:k,]
13.      New_Test = New_Data[(k+1): (k+n),1:P]
14.      Train_Model = Learning Algorithm(New_Train)
15.      Proba = Train_Model(New_Test)
16.      if Proba> = 0.5 then
17.         Trading Signal0 = 1
18.      else
19.         Trading Signal0 = 0
20.      End if
21.      Trading Signal = c(Trading Signal, Trading Signal0)
23.   End for
24.   return (Trading Signals)
25. End for
```

The data set is from SPICS, by analyzing the difference between parameters like ARR, WR, MDD, and ASR to find which algorithm is optimal for different industries. It can be inferred that Annualized Return of Rate (ARR) using Classification and Regression Trees is the highest CART is optimal in Basic Materials, Communication, Energy, and Industry. Deep Belief Network is optimal in Consumer Cycle and Public Utility. Multi-Layer Perceptron is the highest in Finance and Non-Consumer Cyclical. Sparse Auto Encoder is the highest in Technology. Frequency of model to the highest number of sectors. It is apparent that ASR of SAE is highest in EN, XGB is highest in COM and BM; RF is the highest in the CC, in the remining sectors. In the FIN, there isn't notable difference between the ASR's. The ASR of

CART < NB, SVM, XGB, and RF; In Industry sector, the ASR of RF > DBN; In the Non-Consumer Cyclical sector, the ASR of RF is > MLP, DBN. It is apparent that BAH is the highest of all sectors.

IV. RESULT ANALYSIS

The Optimized Trade Algorithm (TOTA) can be used in all sectors as an indicator. Setting some rules like, considering 2 algorithms a, b, "$a = b$" means performance of algorithm a is not so distinct from algorithm b, "$a>b$" shows that performance of a is better than b.For any sector, $i \in$ {NB, RF, LR, CART, KNN, XGB, RNN, MLP, SVM, DBN}, for evaluating factors $j \in \{WR, ARR, ASR, MDD\}$. Higher the value of j indicates model has better performance

First, by choosing the optimal methods for every sector that has better performance the benchmark. Secondly, the performance of the best method can be more precise than the BAH in every sector, while taking a strategy which reduce the risk of loss. By selecting the TOTA's that has better performance than the remining methods, as is evident from Table.2. From Table, Taking WR as indicator, MLP is the best and optimal method for all the sectors; Taking ASR as indicator in the CC, FIN, NCC, EN, and IND, any method can be used. Taking ARR as indicator, for FIN, MLP is the optimal algorithm, for remining sectors any method can be used Taking MDD as indicator, for FIN, CC, IND, NCC, and EN, and any method can be used in other sectors.

It is apparent from Table 3, that there is more than one optimal algorithm. There isn't much of a difference in performance of these algorithms. For example, for the TECH sector, takin WR as indicator, the algorithms that have better performance are DBN, SAE, and MLP. These multiple methods don't have much difference in WR. But it is apparent in Table 6, that all the optimal algorithms are not mentioned, and some are represented as "ATAU", which implies that there isn't much difference between the performance of all the algorithms, so some new rules need to be formed. For all the sectors, ASR returns only one optimal method that considers the risks and the returns, it can be considered as high importance for evaluation of the optimal method, better than the remining indicators.

MDD also describes potential risks from trading algorithms. WR doesn't show actual price of a stock, it only shows us performance of an algorithm on price of stock. Hence, $ASR>ARR>MDD>WR$ is order of indicator evaluation process, $a>b$ means a is a better indicator than b in terms of importance. The below rules are used to refine further. For example, Applying rules 2 to find TOTA's for FIN: WR ={ *SAE, DBN, MLP }*, ARR = *{MLP}* ASR = (ALL except)={ *GRU*, *LR*, *MLP*, *DBN*, *LSTM*, *NB*, *SVM*, *XGB*, *SAE*, *RNN*, *RF*}; MDD = { *GRU, RNN, RF, XGB, LSTM , SVM},* so $ASR \cap WR \cap ARR = MLP$ which means non null value that obeys all the rules(common factor) is MLP so the TOTA of FIN will be MLP.

Table 2: optimal algorithm

Industry	Indicator	SPICS
BM	WR	MLP, DBN, SAE
	ARR	Any trading algorithm can be used (ATAU).
	ASR	ATAU
	MDD	ATAU
CC	WR	MLP, DBN, SAE
	ARR	ATAU
	ASR	LR
	MDD	RF, XGB
COM	WR	MLP
	ARR	ATAU
	ASR	ATAU
	MDD	ATAU
EN	WR	ATAU
	ARR	ATAU
	ASR	SAE
	MDD	MLP, DBN, SAE
FIN	WR	MLP, DBN, SAE
	ARR	MLP
	ASR	Any trading algorithm can be used except CART.
	MDD	RNN, LSTM, GRU, SVM, XGB, RF
IND	WR	MLP, DBN, SAE
	ARR	ATAU
	ASR	RF
	MDD	GRU, CART, RF, LR, SVM, NB, XGB
NCC	WR	MLP, DBN, SAE
	ARR	ATAU
	ASR	RF
	MDD	RNN, GRU, CART, NB, RF, XGB
PU	WR	MLP, DBN, SAE
	ARR	ATAU
	ASR	ATAU
	MDD	ATAU
TECH	WR	MLP, DBN, SAE
	ARR	ATAU
	ASR	ATAU
	MDD	ATAU

Like this TOTA for all the other sectors can be obtained. As shown in Table.3

Table.3. TOTA sector details

Industry	SPICS
BM	MLP, DBN, SAE
CC	LR
COM	MLP
EN	SAE
FIN	MLP
IND	RF
NCC	RF
PU	MLP, DBN, SAE
TECH	MLP, DBN, SAE

It is apparent from Table.2, the total count of optimal algorithms that follow the Rule 2 is less. reason being, Rule 2 considers ASR vs the remining factor's significance. On the other hand, DNN a have a better performance than the remining methods in most sectors, and some others are best for different sectors. This shows us that on the dataset taken, TOTAs which can either select based off single indicator or better one on multiple indicators from all sectors. TOTAs are applied to perform the trading in every sector.

V. Conclusion

Compared to humans, algorithms do a better job at making profits on trading due to experience. Small changes in algorithms based on different strategies for different time segments make the algorithm directed towards respective profitable algorithms. That's why when observing both datasets, It can be inferred that different markets need different methods to have an optimal trade. By giving factors like ARR priority to find a benchmark and compare them to the remaining algorithms to find the most optimal model. Better Benchmarks for measuring optimality in the future can make finding a method very easy and can be used for High-frequency trading(HFT) using automation and risk management.

References

[1] Duffy, John. "Agent-based models and human subject experiments." Handbook of computational economics 2 (2006): 949-1011, Elsevier

[2] Cartlidge, J.; Szostek, C.; De Luca, M. and Cliff, D. (2012). TOO FAST TOO FURIOUS - Faster Financial-market Trading Agents Can Give Less Efficient Markets

[3] S. R. Reeja and N. P. Kavya, "Real time video denoising," 2012 IEEE International Conference on Engineering Education: Innovative Practices and Future Trends (AICERA), 2012, pp. 1-5, doi: 10.1109/AICERA.2012.6306745

[4] Kirilenko, Andrei A., and Andrew W. Lo. 2013. "Moore's Law versus Murphy's Law: Algorithmic Trading and Its Discontents."

[5] Goldstein, M.A., Kumar, P. and Graves, F.C. (2014), Computerized and High-Frequency Trading Beckhardt & Miller and Shorter

[6] Miller, R.S., & Shorter, G.W. (2016). High-Frequency Trading: Overview of Recent Developments.

[7] Bao, Te, and NEKRASOVA, Elizaveta and Neugebauer, Tibor and Riyanto, Yohanes E., Algorithmic Trading in Experimental Markets with Human Traders

[8] Dias, Norman. "Refleja.(2018). A quantitative report on the present strategies of Graphical authentication." International Journal of Computer Sciences and Engineering 6: 64-73

[9] Lv D, Huang Z, Li M, Xiang Y (2019) Selection of the optimal trading model for stock investment in different industries.

[10] Dias, N., Reeja, S.R. (2020). An Improvement of Compelling Graphical Confirmation Plan and Cryptography for Upgrading the Information Security and Preventing Shoulder Surfing Assault. In: Arai, K., Bhatia, R., Kapoor, S. (eds) Proceedings of the Future Technologies Conference (FTC) 2019. FTC 2019. Advances in Intelligent Systems and Computing, vol 1070. Springer, Cham.

[11] Soja Rani, S., Reeja, S.R. (2020). A Survey on Different Approaches for Malware Detection Using Machine Learning Techniques. In: Karrupusamy, P., Chen, J., Shi, Y. (eds) Sustainable Communication Networks and Application. ICSCN 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 39. Springer, Cham.

[12] Dias, N., Mouleeswaran, S.K., Reeja, S.R, A Systematic approach towards enhancing of Security and usability of graphical password through cognitive computing and data mining, Indian Journal of Computer Science and Engineeringthis link is disabled, 2021, 12(6), pp. 1789–1802

[13] S.R. Reeja, Rino Cherian, Kiran Waghmare, Jothimani , Chapter 7 - EEG signal-based human emotion detection using an artificial neural network, Editor(s): Hemanth D. Jude, Handbook of Decision Support Systems for Neurological Disorders, Academic Press, 2021, Pages 107-124

[14] Bao, Te, Elizaveta Nekrasova, Tibor Neugebauer, and Yohanes E. Riyanto. "Algorithmic trading in experimental markets with human traders: A literature survey." Handbook of Experimental Finance (2022): 302-322.

[15] Jose, J.M., Reeja, S.R. (2022). Anomaly Detection on System Generated Logs—A Survey Study. In: Shakya, S., Bestak, R., Palanisamy, R., Kamel, K.A. (eds) Mobile Computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol 68. Springer, Singapore. https://doi.org/10.1007/978-981-16-1866-6_59

[16] Reeja, S. R. "object detection,Tracking and Behavioural Analysis for static and moving background", International Journal of Early Childhood Special Education, Volume 14, Issue 05, Pages 788-795, 2022