Combination of SOM, SVR, and LMKNN for Stock Price Prediction

1st Maradona Jonas Simanullang
Universitas Potensi Utama
Medan, Indonesia
maradonajonas94@outlook.com

2nd Dr. Hartono, S.Kom,M.Kom
Universitas Potensi Utama
Medan, Indonesia
hartonoibbi@gmail.com

3nd Dr. Roslina, M.I.T
Universitas Potensi Utama
Medan, Indonesia
roslinanich@gmail.com

2023 International Conference of Computer Science and Information Technology (ICOSNIKOM) | 979-8-3503-6075-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOSNIKOM60230.2023.10364522

Abstract*—Support Vector Machine (SVM) is a strong AI calculation utilized for characterization and relapse errands. One of the shortcomings of SVM is that it doesn't offer direct help for multi-class grouping. The K-Nearest Neighbort (KNN) strategy is known as a basic, quick, and effectively implementable AI technique in different fields, in spite of the fact that its exactness can, in any case, be gotten to the next level. One central disadvantage of KNN is the position of new information into a class in light of most of the neighbors (Vote Larger part Framework), which can be dangerous when the good ways from each closest neighbor vary essentially from the distance of the testing information, frequently prompting misclassifications in KNN. The objective of this exploration is to propose a model that joins Self Organizing Map (SOM), Support Vector Reggression (SVR), and Local Mean-Based K-Nearest Neigbort (LMKNN) to work on the precision of stock cost forecast. From the experimental outcomes, the most elevated exactnesses were acquired for the accompanying stock information: 94.66% for BBCA with the Laplacian piece, 89.29% for ASII with the Gaussian bit, 98.64% for TLKM with the Gaussian part, 98.31% for BBRI with the Gaussian bit, and 96.31% for PGAS with the Gaussian bit.

Keywords— SVM, K-NN ,SVR, SOM, LMKNN

I. INTRODUCTION

AI is a part of Man-made reasoning that spotlights on the utilization of information and calculations to emulate the manner in which people learn without the requirement for explicit programming. AI can be utilized to procure new information and produce prescient models [1]. many machine learning methods can be used to generate predictive models, including Irregular Woodland (RF), AdaBoost, Backing Vector Machine (SVM), and K-Nearest Neighbor (KNN) [2]. Past examination has been led on stock cost expectation utilizing SV-KNNC and SOM, where this review was directed in the information arrangement stage, and the outcomes showed that KNN performs better for information characterization [3]. However, a drawback of KNN is the placement of new data into the class with the biggest number of neighbors (Vote Greater part Framework), which can be tricky when the good ways from each closest neighbor shift extraordinarily from the distance of the testing information [4]. SV-KNNC is a calculation that comprises SVM, K-Means, and KNN. The grouping brought about by this review was not agreeable, as the presentation of K-Means couldn't deliver productive information bunches, as proven by the low precision values obtained from a few tests [3]. SVM attempts to choose significant preparation information, and in the K-Means stage, there was a change by supplanting K-Means with SOM, which is a procedure that can perform information grouping to relegate loads to each preparing data*to work on the productivity of KNN [3].

II. LITERATURE STUDY

The proposed model in this review is a blend of SVR (Support Vector Regression) to deal with nonlinear information for getting the best vectors in characterizing datasets, SOM (Self Organizing Map) for bunching, and LMKNN (Local Mean-based K Nearest Neighbor) to address KNN [5].

A . *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a technique established in measurable learning hypothesis that has shown promising outcomes in giving improved results contrasted with different strategies [6]. SVM are AI instruments that dissect information and perceive examples or choice limits inside the dataset utilized basically for arrangement and relapse investigation [7]. SVM makes equal parcels by producing two equal lines to make a different space in a high-layered space utilizing the vast majority of its credits [7]. This plane is known as the hyperplane. It makes hyper-planes that have the biggest edge in the high-layered space, subsequently isolating the given information into classes and making edges. The edge addresses the most extreme distance between the nearest significant pieces of information of the two classes. The bigger the edge, the lower* the speculation mistake of the classifier. SVM gives the greatest adaptability of the multitude of classifiers [7].SVM functions admirably on high-layered datasets, and in any event, while utilizing a part, SVM maps the first information from its unique aspects to a moderately higher-layered space [7]. In contrast to Counterfeit Brain Organizations (ANN), where all chosen information is considered during the preparation cycle, SVM works in an unexpected way. Just a chosen subset of information adds to the development of the model utilized in the learned grouping [6]. SVM additionally contrasts with Closest Neighbor Draws Near, which stores all preparing information for expectation purposes. SVM just holds a little piece of the preparation information for use during the forecast. This is a strength of SVM as not all preparing information is viewed in each preparing cycle [5]. The information that is added to SVM is referred to as Help Vectors (SVs), consequently, the strategy is called Help Vector Machine [3]. Some of the kernels contained in SVM include: [8]:

1. Polynomial Degree h Kernel:
   The polynomial degree h portion stunt is reasonable for tackling grouping issues while the preparation dataset is as of now standardized. This portion stunt is communicated in Equation 1.

$$K(x_i\ x_j + 1) = \exp = (-\frac{||x_1 - x_2||^2}{2\sigma^2} \quad (1)$$

2. *Radial Basis Function*

The Spiral Premise Capability (RBF) portion stunt is the most usually involved piece for taking care of characterization issues on non-straightly divisible datasets. It offers incredible preparation and expectation precision. The RBF piece is communicated in Equation 2 :

$$K(x_1, x_2) = x_1^T x_2 \quad (2)$$

Where: Xi dan Xj = Pair of two training data

3. *The Sigmoid (Hyperbolic Tangent)*

The Sigmoid kernel is a kernel trick used in Support Vector Machines, which is an extension of artificial neural networks. It is expressed as follows:

$$K(x_1, x_2) = \tanh(\beta 0 x_1^T x_2 + \beta_1) \quad (3)$$

The part stunt gives a few benefits in light of the fact that during the SVM growing experience, to decide the help vectors, the client just has to realize the bit capability being utilized, without having to know the type of the non-straight capability. Among all the part deceives the spiral premise capability portion stunt yields the best outcomes in arrangement, particularly for information that can't be straightly isolated.

B. *Support Vector Regression (SVR)*

Support Vector Regression (SVR) expands the idea of a Help Vector Machine (SVM) to address relapse issues. SVM is a method used to separate a dataset into two classes using a hyperplane, which acts as a separating line. While SVM aims to divide the dataset into two distinct zones for classification purposes, SVR works oppositely by ensuring that all data points fall within a single zone while minimizing the value of epsilon (ε) [9]. The process of using SVR involves the following steps: The process of using SVR involves the following steps:

a. Preparing the training data.
b. Selecting the appropriate kernel, parameters, and regularization.
c. Creating a model to obtain coefficients.
d. Utilizing the obtained coefficients to build the estimator.

For a given landslide monitoring dataset y = f(x), non-straight SVR with a portion capability K(xi, x) is planned as displayed in Equation 4 [9]:

$$y = f(x) \sum_{i=1}^{n} \omega_i \ (x_i, x) + b \quad (4)$$

Where w and b are the weight vector and inclination, individually. The non-direct SVR with piece capability can be gotten through an advancement issue as displayed in Equation 5

$$min_{w,b} \frac{1}{2} \ ||\omega_i||^2 + C \sum_{t=1}^{T} |\varepsilon_t + \varepsilon_t| \quad (5)$$

where C, ω_t, and ω_t∗ t are the punishment boundary and two leeway factors, individually. By presenting the Lagrange multipliers a∗I, what's more, man-made intelligence, nonlinear SVR can be changed over completely to a double issue and communicated as follows:.

$$y = f(x) = \sum_{i=1}^{n} (a_i - a_i) \ K(x_i, x) + b \quad (6)$$

Different portions, including direct, polynomial, Gaussian, and sigmoid portions, have been proposed.

C. *Self Organizing Map (SOM)*

One Self Organizing Map (SOM), otherwise called the Kohonen Guide, is a brain network procedure intended to picture information by diminishing its dimensionality. It achieves this through the utilization of self-organizing neural networks, enabling human comprehension of complex datasets mapped into a lower-dimensional space [10]. In the SOM calculation, each group unit is related to a weight vector that fills in as a model of an info design connected to that particular bunch. Oneself getting sorted out process involves recognizing the bunch unit with loads nearest to the information vector design, frequently estimated utilizing the square of the base Euclidean distance, and assigning it as the victor [9]. The triumphant unit, alongside its adjoining units as per the geography of the group units, then, at that point, continues to refresh their separate loads. Subsequently, the Kohonen SOM shows explicit result responses to individual info designs, in this manner uncovering likenesses among individuals inside a similar group [10]. The course of SOM bunching envelops two essential advances: introduction and preparing, the two of which are additionally explained in the accompanying five phases [3]:

1. Initialization:

The underlying weight vectors of the SOM are introduced utilizing an irregular-based technique. Moreover, the size of the not entirely settled by utilizing the information boundaries of width and level as per the network aspects.

2. Best Matching Unit (BMU):

The BMU is distinguished through the hub with the nearest Euclidean distance to the information design. The condition for Euclidean distance should be visible in Equation 7.

$$d_{ij} \sqrt{\sum_{k=1}^{n} (X_{ij} - X_{ij})^2} \quad (7)$$

Where :

d_{ij}= The distance between occurrence I and j

$x_{i\,j\ dan}$ $X_{i\,j}$ = coordinates I and j

3. *Weight Updated*

During every emphasis, the loads for the hubs are refreshed inside the BMU's local sweep. The size of the general climate shrivels with every emphasis utilizing a Gaussian capability.

4. *Iteration*

The nature of the framed groups is approved involving a few notable measures for inner standards, for example, the Outline list, Davies-Bouldin file, and Dunn record. Interior standards approval is performed in light of the fact that the approval cycle depends on data from the actual information. In this way, the approval results are estimated in light of the vicinity of components inside bunches (attachment) and the distance between various groups (partition). Great bunches are described by least union and most extreme division

distances. The exploratory consequences of this examination are talked about in the ensuing segment.

D. Local Mean-Based K-Nearest Neighbor (LKMNN)

K-nearest neighbor (KNN) grouping is a notable and sample non-parametric technique in pattern classification. Be that as it may, its order exactness can be handily impacted by exceptions, especially in little example sizes. Nearby Mean Based K-Closest Neighbor (LMKNN) is an augmentation of KNN [12]. LMKNN has been demonstrated to further develop order execution and decrease the effect of exceptions, particularly in little information sizes [12].

The LMKNN interaction can be portrayed as follows [11]:

1. Determining the value of K.
2. Work out the distance between the test information and every one of the preparation information involving Euclidean distance as portrayed in Equation 9.

$$D(x,y) = ||x-y||_2 = \sqrt{\sum_{j=1}^{N}|x-y|^2} \quad (9)$$

3. Sort the information in good ways from littlest to biggest, choosing the top K distances for each class of information. Calculate the local mean vectors for each class using Equation 10.

$$w_c = argmin_{wj}\left(x, m_{wj}^k\right), j = 1,2,\dots,M \quad (10)$$

4. Decide the class of the test information by computing the closest distance to the neighborhood mean vector of each class Equation 11.

$$m_{wj}^k = \frac{i}{k}\sum_{i=1}^{k} yi^N, j^N \quad (11)$$

The precision, also known as confidence, is the proportion of the number of true positive cases correctly predicted as positive out of the total predicted positive cases. On the other hand, recall, also known as sensitivity, is the proportion of the number of true positive cases correctly predicted as positive out of the total actual positive cases [13]. The accuracy calculation can be seen in Equation 12 [13].

E. *Confusion Matrix*

The precision, also known as confidence, is the proportion of the number of true positive cases correctly predicted as positive out of the total positive cases. On the other hand, recall, also known as sensitivity, is the proportion of the number of true positive cases correctly predicted as positive out of the total actual positive cases [13]. The accuracy calculation can be seen in Equation 12 [13].

$$Accuracy\ (\text{ACC}) = \frac{TP}{P} - \frac{TN}{N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Explanation:

P = *Sample Positive*
N = *negative*
TP = *True positive*
TN = *True negative*(TN)
FP = *False positive* (FP)
FN = *False negative*

III. METHODOLOGY

Just help vectors (SVR) are utilized in the help vector machine to group concealed information. The SVR utilized for instance are those nearest to the hyperplane. Subsequently, the chosen SVR addresses preparing information which is sufficient to address all preparing information. In the choice cycle, the preparation information is all embedded into the SVM. Just a bunch of SVR is delivered as a consequence of SVM preparation. These SVRs are then kept in memory. The cycle likewise kills rehashing occasions in each class from the element space. Re-checking the SVR commitment to the grouping in the info space is required prior to utilizing them to bunch a query*instance to keep away from issues. To lay out the commitment of every SVR, a weight will be doled out. A higher*weight occasion is better so it will be more helpful in LMKNN bunching

$$w_i = \frac{n(class\,(xi))}{Total} \quad (12)$$

Prior to directing the gathering, an order is performed first to find the best vectors involving SVR for the information-gathering process utilizing SOM. The weighted SVR will be ordered utilizing the LMKNN model to notice the expectation results. Figure 1 is a blend of SVR + SOM + LMKNN. Figure 2 is SVR, Figure 3 is SOM and Figure 4 is LMKNN.

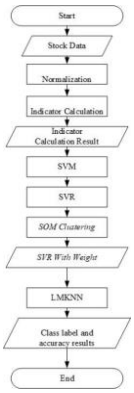

Figure 1. Model SOM SVR and LMKNN

Figure 1 The model comprises of SOM, SVR, and LMKNN for stock cost expectation.

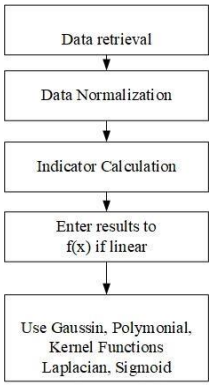

Figure 2. Model SVR (*Support Vector Regression*)

Figure 2 the model is SVR (Support Vector Reggression) prior to performing bunching.

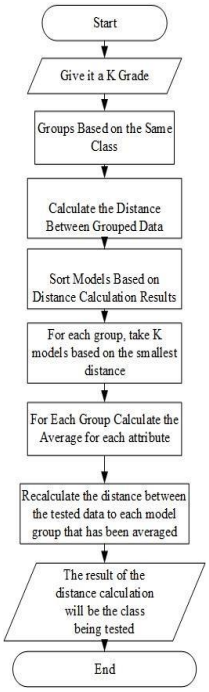

Figure 3. Model LMKNN (*Local Mean Based K-NearestNeighbor*)

Fig.3 is the flowchart for the information arrangement process utilizing assesses the precision of the expectations. The emphasis values are gotten by duplicating by five.

$$y = \frac{2 * x - (\max + \min)}{(\max + \min)} \qquad (12)$$

After standardization, the everyday stock price‟s specialized markers will be determined. Ten markers are utilized in this exploration which are: Remarkable Moving Normal Stochastic Oscillator*, Relative Strength File, Pace of progress and momentum*, and ADO/CLV (Gathering/Appropriation). Information preparing is the blend of crude informational indexes and specialized pointers. After the best vector is gotten, the best group is not entirely set in stone in the information bunching process utilizing Oneself Getting sorted preparationout Guide. After the best bunch is gotten, the information is grouped utilizing SOM. After the stock information is bunched, the heaviness of each stock information will be set. Stock information that has been. LMKNN. Support Vector is outfitted with a weight incentive for every information. Every piece of information has a class as well. The distance between these information will be processed. The mean is figured, and the class esteem is determined in the LMKNN cycle to improve the stock cost prediction‟s esteem. The Versatile direct combiner recipe is utilized in the standardization cycle. The day to day stock trade cost taken from the dataset fills in as the cycle of the Indonesia Stock Trade is utilized in this analysis. The chosen stocks are Telkom Indonesia, Astra Global, Bank Focal Indonesia, Organization Gas Negara, and Bank Rakyat Indonesia. These stocks were decided because of their steady pay and huge resources. The information is obtained from the Yahoo Money site. Then, The SVR vector from the SVM model will be determined.

IV. EXPERIMENT AND RESULT

The stock data from 31st January 2013 to 30th May 2023 (absolute of 2563 days) of 5 blue-chip stocks on

TABLE I. THE F-MEASURE‟ 'S COMPARISON OF LMKNNDAN KNN

Iterasi	**LMKNN**				
	TLKM	ASII	BBCA	PGAS	BBRI
5	98,64	89,39	96,66	96,96	98,31
10	98,64	89,39	96,66	96,,96	98,31
15	98,64	89,39	96,66	96,96	98,31
20	98,64	89,39	96,66	96.96	98,31
25	98,64	89,39	96,66	96,96	98,31
30	98,64	89,39	96,66	96,96	98,31
35	98,64	89,39	96,66	96,96	98,31
40	98,64	89,39	96,66	96,96	98,31
45	98,64	89,39	96,66	96,96	98,31

Iterasi	**KNN**				
	TLKM	ASII	BBCA	PGAS	BBRI
5	97,57	74,58	93,43	94,54	97,00
10	97,57	74,58	93,43	94,54	97,00
15	97,57	74,58	93,43	94,54	97,00
20	97,57	74,58	93,43	94,54	97,00
25	97.57	74,58	93,43	94,54	97,00
30	97,57	74,58	93,43	94,54	97,00
35	97,57	74,58	93,43	94,54	97,00
40	97,57	74,58	93,43	94,54	97,00
45	97,57	74,58	93,43	94,54	97,00

The LMKNN model shows better exactness for each organization while utilizing different emphasis boundaries. The worth of K utilized in both the KNN and LMKNN models is The Class Boundaries for SVR are as

per the following: TKLM utilizes the Gaussian portion, ASII utilizes the Laplacian bit, BBCA utilizes the Laplacian piece, PGAS utilizes the Gaussian bit, and BBRI utilizes the Gaussian bit. The bunch boundaries are produced by SOM for each organization utilizing 3 groups

V. CONCLUSION

In this review, the proposed model consolidates SVR (Support Vector Reggression) + SOM (Self Organizing ao) and LMKNN (Local Mean Based K-Neareset Neigbort). By changing the information grouping step, the presentation of SVR can be streamlined to acquire the best vector. By executing LMKNN in the information grouping step, the precision of the stock forecast can be gotten to the next level. Besides, the mix of LMKNN in the information bunching step upgrades the grouping system by consolidating neighborhood models. LMKNN uses the idea of k-closest neighbors to characterize information directs in light of their nearness toward adjoining focuses. This approach thinks about the nearby qualities of the information and can work on the exactness of the bunching results, accordingly improving the precision of stock forecast in this examination setting. By joining these strategies, the proposed model expects to defeat the constraints of the SV-KNNC + SOM approach and accomplish higher precision in stock forecast errands. Bank Rakyat Indonesia (BBRI), Astra Global (ASII JK), organization Gas Negara (PGAS JK), Telkom Indonesia (TLKM JK), and Bank Focal Indonesia (BBCA JK).. From the experimental outcomes, the most noteworthy correctnesses were gotten for the accompanying stock information: 94,66% for BBCA with the Laplacian part, 89,29% for ASII with the Gaussian piece, 98.64% for TLKM with the Gaussian portion, 98.31% for BBRI with the Gaussian bit, and 96,31% for PGAS with the Gaussian bit.

REFERENCES

[1] Albagmi, F. M., Alansari, A., Al Shawan, D. S., AlNujaidi, H. Y., & Olatunji, S. O. (2022). Prediction of generalized anxiety levels during the Covid-19 pandemic: A machine learning-based modeling approach. *Informatics in Medicine Unlocked*, *28*, 100854. https://doi.org/10.1016/j.imu.2022.100854

[2] Su, S., & Wang, J. (2023). Machine learning prediction of contents of oxygenated components in bio-oil using extreme gradient boosting method under different pyrolysis conditions. *Bioresource Technology*, *379*, 129040. https://doi.org/10.1016/j.biortech.2023.129040

[3] F. M. Sinaga, M. Jonas, Felix, and A. Halim, "Stock trend prediction using SV-KNNC and som," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019. doi:10.1109/icic47613.2019.8985731.

[4] Y. Wang, Z. Pan, and J. Dong, "A new two-layer nearest neighbor selection method for Knn Classifier," Knowledge-Based Systems, vol. 235, p. 107604, 2022. doi:10.1016/j.knosys.2021.107604

[5] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," Decision Analytics Journal, vol. 3, p. 100071, 2022. doi:10.1016/j.dajour.2022.100071

[6] P. Chhajer, M. Shah, and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction," Decision Analytics Journal, vol. 2, p. 100015, 2022. doi:10.1016/j.dajour.2021.100015

[7] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification,"

[8] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support Vector Machine Classification: Applications, challenges, and Trends. Neurocomputing, 408,

[9] J. Ma et al., "Metaheuristic-based support vector regression for landslide displacement prediction: A comparative study," Landslides, vol. 19, no. 10, pp. 2489–2511, 2022. doi:10.1007/s10346-022-01923-6

[10] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, "Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the World Using Self Organizing Maps," Chaos, Solitons & Fractals, vol. 138, p. 109917, 2020. doi:10.1016/j.chaos.2020.109917

[11] S. Liu, "Smote-LMKNN: A synthetic minority oversampling technique based on local means-based K-nearest neighbor," International Journal of Pattern Recognition and Artificial Intelligence, vol. 36, no. 05, 2022. doi:10.1142/s0218001422500197

[12] Mehta, S. et al. (2018) „A new nearest centroid neighbor classifierbased on k local means using harmonic mean distance",

[13] M.-T. Wu, "Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom," Scientific Reports, vol. 12, no. 1, 2022. doi:10.1038/s41598-022-07137-z.

[14] Syaliman, K. U., Nababan, E. B. and Sitompul, O. S. (2018) „Improving the accuracy of the k-nearest neighbor using local mean based and distance weight", Journal of Physics: Conference Series, 978(1). doi:10.1088/1742- 6596/978/1/012047. Local Mean Based K- Nearest Neighbor (LMKNN)

[15] W. R. Fadilah, D. Agfiannisa, and Y. Azhar, "Analisis Prediksi Harga Saham pt. Telekomunikasi Indonesia Menggunakan metode support vector machine," Fountain of Informatics Journal,vol. 5, no. 2, p. 45, 2020. doi:10.21111/fij.v5i2.4449

[16] W. R. Fadilah, D. Agfiannisa, and Y. Azhar, "Analisis Prediksi Harga Saham pt. Telekomunikasi Indonesia Menggunakan metode support vector machine," Fountain of Informatics Journal,vol. 5, no. 2, p. 45, 2020. doi:10.21111/fij.v5i2.4449

[17] Sivaram, M., Lydia, E. L., Pustokhina, I. V., Pustokhin, D. A., Elhoseny, M., Joshi, G. P., & Shankar, K. (2020). An optimal least square support vector machine-based earnings prediction of Blockchain Financial Products. IEEE Access,8,120321–120330. https://doi.org/10.1109/access.2020.3005808

[18] Abuzneid, M. A., & Mahmood, A. (2018). Enhanced human face recognition using LBPH descriptor, Multi-KNN, and back- propagation Neural Network. IEEE Access, 6, 20641–20651. https://doi.org/10.1109/access.2018.2825310