Evaluation of Pairs Trading Strategies using Machine Learning

Abstract Traditional Pairs Trading is a strategy that has heavily been dependent on and its efficiency limited by its two main processes, (a) identify correlated pairs by cointegration, a suitable but rudimentary tool, and (b) identify entries and exits based on the spread's deviation from the mean (SMR), a very simplistic policy. We propose alternative approaches to these two processes. For identifying the pairs, we use OPTICS and Hierarchical clustering. Both of these models study the returns of the individual securities and attempt to intelligently identify naturally occurring clusters in the data, transcending the linearity restrictions imposed by the cointegration pair-selection technique. For determining entry and exits, we introduce two forecasting techniques – ARIMA and LSTM - to study the expected changes in the spread and capitalize on any predicted divergences. For the purpose of this study, we select 120 equity ETFs as our universe for identifying eligible pairs, with our data ranging from 2018 to 2022. We observe that employing the clustering algorithms and forecasting techniques grants considerably better stability and more optimal profits. ARIMA strategy fetches a Sharpe Ratio of 9.07 and an annual ROI of 22.96%, and the LSTM strategy fetches a Sharpe Ratio of 2.77 and an annual ROI of 24.84%, in stark contrast to SMR's Sharpe Ratio of 1.87 and ROI of 20.96% under the OPTICS pair selection regime. The combination of OPTICS and LSTM proves most efficient with an ROI of 25%.

I. INTRODUCTION

Pairs trading is associated with identifying two securities that have historically correlated price movements. Traditionally this strategy attempts to capitalize on divergences in the prices of two related securities with the calculated expectation of the spread reverting to its mean. While the strategy is adequate in theory, it has long been plagued by two issues - its necessity to identify closely correlated pairs by cointegration, a suitable but rudimentary tool, and its possibly simplistic policy of identifying points of entries and exits based on the spread's deviation from the mean i.e. Standard Mean Reversion (SMR). Cointegration's primary drawback is its necessity to mandate linearity and consequently, its fails to recognize significant non-linear relationships that may exist between two related securities. This paper focuses on comparing the performance of novel non-linear pairs selection methodologies using data of 120 equity Exchange Traded Funds (ETFs). This work also explores the performance of a forecast spread-based trading model using both linear and non-linear predictions approach along with the standard mean reversion trading model.

The two main considerations behind pairs trading are the pair selection criteria and trading strategy. Depending on the number of assets available in the data sample, the computational complexity of the pair selection phase becomes an increasingly important factor. This is our focus in the pair selection part of the paper. After this, we present our trading strategies based on the set of selected pairs from each methodology. Here we present the rules for entering and exiting positions. In part IV we describe in detail the data processing steps and implementation of the models. In part V we present the results and performance metrics of different combinations of pairs selection method and trading algorithm. Finally, in part VI we draw conclusions based on the attributes of our model and data.

A. Related Research Overview

Pairs trading and more general statistical arbitrage strategies have been active research areas for several decades. The most promising recent developments include applications of machine learning and copulas combined with big data to model nonlinear relationships among price series. In addition to being a popular and profitable strategy, [7] finds that pairs trading performance is also robust in highly volatile market conditions.

To maximize the likelihood of finding profitable trading opportunities we analyze 5-minute data instead of daily prices because as [6] remarks, for high-frequency data we find greater levels of comovement which leads to more stock pairs being cointegrated. This allows us to impose stricter criteria for selecting pairs among all possible combinations and improve the performance of our trading strategy.

While copulas present a promising approach due to their ability to model the dependence structure among returns distribution, [3] argues that a direct application of the Misprice Index (MI) from [7] to high-frequency data is not suitable because it is extremely difficult to separately reflect the effect of MI and the proper selection of timing issues. Therefore, we believe that machine learning applications are more flexible and better suited for working with our data sample. We take inspiration from [9], [8], and [5] in using unsupervised learning to reduce the search space for pair selection, and forecasting methods for constructing some of our trading strategies.

II. PAIR SELECTION

Given our large universe of 120 equity ETFs our approach to selecting the most promising pairs to trade on consists of several steps. First, we produce a compact representation of each asset return via dimensionality reduction. We then implement two clustering algorithms of the new return series to make it easier to exploit stationary patterns among pairs. Finally, we iterate through all possible pair combinations within the reduced search space and select the candidates that meet several criteria based on the degree of cointegration and mean-reverting behavior.

A. Clustering Techniques

Given our large universe of 120 ETFs we decided to apply an unsupervised learning algorithm to cluster our data, as opposed to identifying pairs by considering all potential combinations in the original high-dimensional feature space. This approach enables us to only consider potential pairs within each cluster, greatly reducing the computational complexity of pair selection.

When considering alternative clustering techniques we opted for a non-linear density-based algorithm. The reasoning is having a large number of ETFs makes it more important to model nonlinear relationships. In addition, density-based algorithms like DBSCAN (Density-Based Spatial Clustering of Applications with Noise) don't require the specification of the number of clusters, unlike a linear-distance method such as k-means clustering (see Figure 1).

B. Dimensionality Reduction

Our clustering algorithm will not work very well if we apply it to the price series of the 120 ETFs because a high number of features means we are trying to identify clusters in a high-dimensional environment. Therefore, our approach consists in applying dimensionality re-

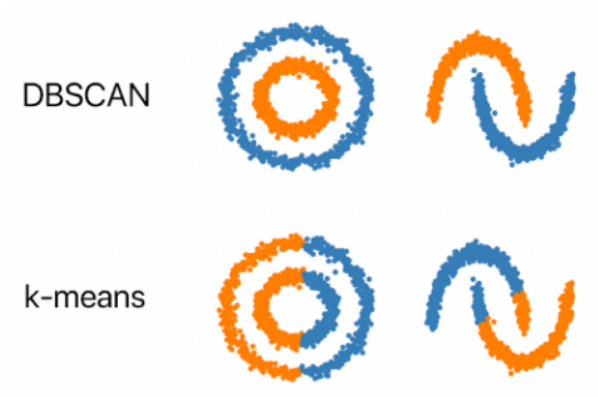

Figure 1: DBSCAN vs k-means. Source [1]

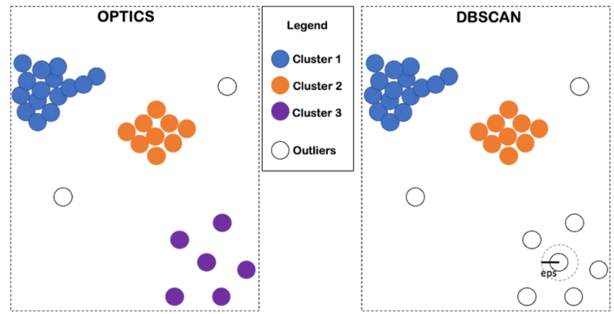

Figure 2: OPTICS vs DBSCAN. Source [10]

duction to get a representation of each ETF as a linear combination of 5 principal Components.

C. OPTICS Algorithm

Ordering Points to Identify the Clustering Structure (OPTICS) is an unsupervised learning algorithm that outlines the structure of a dataset and allows us to create density-based clusters. It is based on DBSCAN, another clustering algorithm that identifies clusters of any shape by tracking high-density regions in the data. However, it can only identify clusters of approximately constant density due to a fixed *eps* parameter, which indicates the maximum distance for points to be considered neighbors.

This shortcoming is addressed by OPTICS (see Figure 2). The reason is that we iteratively consider each point as the current center point to update the reachability distance of its neighbors. This is a desired property for our problem statement because a diverse set of ETFs can't be expected to have equally dense clusters.

For the implementation, we choose $MinPts = 3$, indicating that an area must have a minimum number of three points in order for a sample to be considered a core point and influence the reachability score list. The result is 7 clusters with 27 tickers for a total of 45 possible pairs. The number of potential pairs is quite reasonable and computationally feasible because we only consider pairs within each cluster and the density is approximately uniform.

D. Hierarchical Clustering

Although OPTICS is a powerful algorithm for identifying the underlying structure of the data, its flexibility makes it likely to overfit. Since we don't need to specify the cluster shape, number of clusters or neighbor distance parameter the algorithm can draw patterns from the noise of the data and lead to misleading associations. Therefore, we decided to also employ a common clustering technique.

Hierarchical clustering consists in iteratively merging the most similar data points based on a chosen distance measure. This approach is computationally efficient because we compare the centroid of each cluster, as opposed to each of its data points, with the remaining data points. Furthermore, this technique prioritizes the appropriate association among the most similar data points, which leads to high-quality pair selection.

III. TRADING STRATEGY

In this section, after selecting the most promising pairs, we apply a trading strategy based on three algorithms which are standard mean reversion model, ARIMA and LSTM.

The main idea of the strategy is to set long and short thresholds in advance and trade when the real spread (standard mean reversion model) or predicted spread change (ARIMA and LSTM) is outside the range of thresholds. Besides, the trades are closed once the spread change reverts back to the normal range.

To be specific, the initial balance of each pair selected above is equally weighted. A $1 investment is initially allocated for each pair. Once trading is triggered, the relative value invested in two legs follows the cointegration ratio β between the two stocks, which is shown in the following figure. The absolute investment amount is all the capital for each pair. That is to say, we separate balance account for each pair, and add profits to the initial capital for reinvestment. Also, it has to be pointed out that transaction costs are not considered in calculating profit and loss.

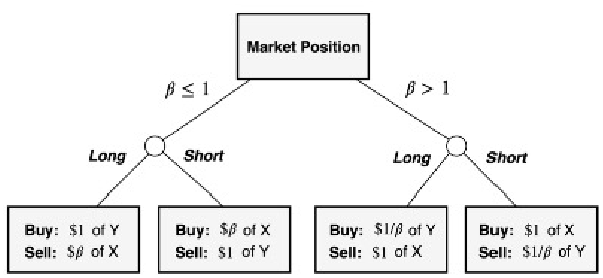

Figure 3: Trading Position. Source [8]

A. Standard Mean Reversion Trading Model

Standard mean reversion (SMR) assumes that the spread between the two assets follows a mean-reverting process. The idea is that when the spread deviates from its long-term mean, it is likely to revert to the mean at some point in the future. In the paper, we use normalized spread to generate trading signals. It can be shown in the formula (1),

$$P_{t+1} = \begin{cases} \text{long spread}, & s_t < \alpha_L \\ \text{short spread}, & s_t > \alpha_S \end{cases} \tag{1}$$

where s_t is the real normalized spread at time t, P_{t+1} is the position at time $t+1$, α_L and α_S are the thresholds.

B. ARIMA Forecast Trading Model

Standard mean reversion model calculates the historical mean and standard deviation of the spread and uses it to determine the entry and exit points for the trade. However, this method assumes that the market conditions remain constant, which is not always the case in real-world scenarios. Therefore, we introduce Autoregressive Integrated Moving Average model (ARIMA).

ARIMA combines three components: autoregression (AR), differencing (I), and moving average (MA). After differencing (I) the data to stationary, lags of the stationary series (AR) and lags of the forecast errors (MA) are used to construct the model, which can be written as

$$X_t = c + \Sigma_{i=1}^{p} \phi_{t-i} X_{t-i} + \epsilon_t + \Sigma_{i=1}^{q} \theta_{t-i} \epsilon_{t-i} \tag{2}$$

where p and q represent the order of the polynomials.

In the paper, ARIMA is applied to capture patterns and trends of spread, and trading signals are generated according to the forecast spread. It can be shown in the formula (3) and (4)

$$\Delta_{t+1} = (S^*_{t+1} - S_t)/S_t \tag{3}$$

$$P_{t+1} = \begin{cases} \text{long spread}, & \Delta_{t+1} > \alpha_L \\ \text{short spread}, & \Delta_{t+1} < \alpha_S \end{cases} \tag{4}$$

where S^*_{t+1} and S_t are predicted and real spread, respectively. Δ_{t+1} is the predicted spread change. Another thing to mention is that the thresholds in standard mean reversion model are determined in an absolute way, while the thresholds in forecast trading models are determined in quantiles.

C. RNN Forecast Trading Model

Though ARIMA is a popular statistical model in time series analysis, it is based on linear regression techniques. We would like to capture nonlinear patterns in time series data with machine learning techniques to see whether the trading strategy can perform better. The technique we choose is the Long Short-Term Memory model (LSTM).

LSTM is a type of recurrent neural network (RNN) that can be used for time-series analysis and prediction. Unlike traditional feedforward neural networks, which process input data in a single pass, LSTM neural networks use a memory cell that can store information over time and selectively forget or remember information based on its relevance to the current input. This allows LSTM networks to model long-term dependencies in time series data, making them well-suited for applications such as stock price prediction. The figure gives a basic structure of LSTM.

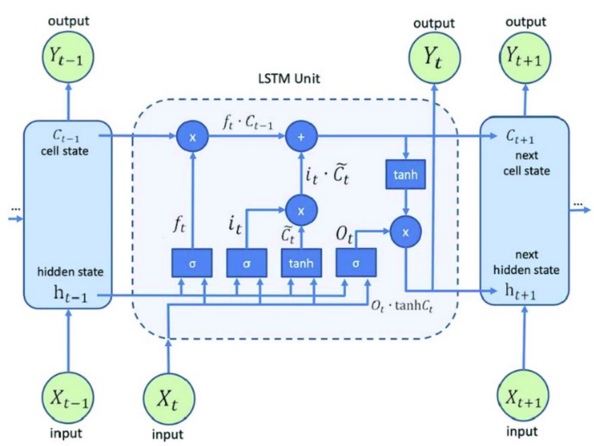

Figure 4: LSTM Structure. Source [11]

In the paper, the LSTM model is trained on historical spread data of each pair to predict the future spread. The trading formula is the same as the one in ARIMA. We will explain it briefly here. If the predicted spread change is larger than a certain threshold, we take a long position in the stock that is expected to outperform, and a short position in the other stock. Conversely, if the predicted spread is smaller than a certain threshold, we take a short position in the stock that is expected to underperform, and a long position in the other stock.

IV. DATA & IMPLEMENTATION

In this section, we are going to introduce the data and discuss the combination of two stages, i.e. pairs selection and trading.

A. Data

A total of 120 ETFs based on equity indices were included in the scope of this project. They were chosen to be the top 120 ETFs when judged by average volume so as to ensure that the data set consisted of the most liquid securities. The adjusted closing prices were considered in intervals of 5 minutes over a time span ranging from 2018-01-01 to 2022-12-30 for a combined sum of 99421 records, with the years 2018 to 2020 serving as the training set, 2021 as the validation set and 2022 as the test set.

Issues of data sanity are handled in two ways: first by linearly interpolating short stretches of missing data (limited to no more than 5 bars) and next by removing tickers wherein missing values persisted notwithstanding the linear interpolation. This effectively reduces the universe of ETFs considered to 71 tickers, yielding 136 pairs (assuming every pair is viable).

B. Implementation

Our trading strategy can be categorized by the clustering method used and the applied forecasting technique (unless standard mean reversion is in effect). For each of these categories, the trading strategy is first tested on the validation set to study the performance of the strategy overall and to glean the most profitable pairs which are then carried forward to be traded on the test set. The most profitable pairs are determined from the validation set by selecting the pairs with cumulative returns greater than the mean cumulative return of the total portfolio. This ensures that the best of the pairs are selected to maximize the expected returns in the test set. While there are initially 10 OPTICS pairs under consideration in the validation set, the number of profitable pairs is narrowed down to a range of 2 to 5 pairs (depending on the strategy used) to ensure optimal results. Similarly, while the Hierarchical pairs initially number 85, they are whittled down to between 11 and 36 pairs (again, depending on the strategy used). The quantiles for the OPTICS strategies were set at 20% (α_S) and 80% (α_L) while the Hierarchical strategies were set at 10% and 90% respectively.

V. RESULTS

Tables 1 and 2 show the performance metrics of the five different trading models implemented. Similarly, Figures 5 and 6 show the total account balance for the validation and test sets respectively. It is interesting to compare the OPTICS algorithm and Hierarchical clustering algorithm in terms of their performance. Optics algorithm has selected 10 pairs while hierarchical clustering has selected 85 pairs in the validation set. However, for the test set, only the most profitable pairs from the validation set are used.

Annual Return: We have considered the performance of S&P 500 for bench marking purposes. The annual return of S&P 500 for the year 2022 was -18.14%. Meanwhile, all our methods gave positive returns in the year 2022, proving that our trading algorithm beats the market significantly. Among the models, we find that Hierarchical clustering with ARIMA forecast trading model yielded the best returns (36.58%), beating the other models by more than 10% returns relatively. Again, Hierarchical clustering had the best annual returns on the validation set, but using the standard mean reversion trading model.

OPTICS vs Hierarchical clustering: Hierarchical clustering generally exhibited higher annual returns than the OPTICS pair selection, it is the OPTICS pairs selection with ARIMA forecast trading model that yields the best sharpe ratio (9.07), lowest maximum drawdown days (1 day), and the lowest portfolio maximum drawdown percentage (0.45%) in the test set. This performance is also consistent for validation data set. Hierarchical clustering yields the next best sharpe ratios in both validation and test cases.

Standard mean reversion vs forecast spread trading model: Based on Table 1, standard mean reversion outperforms forecast model in terms of annual ROI for both OPTICS and Hierarchical clustering in the validation set. However, based on Table 2, standard mean reversion gives the least annual ROI in the test set. Among the forecast models, machine learning based RNN model performs better than linear forecasting based ARIMA model in terms of annual ROI. ARIMA forecast trading model has exceptional sharpe ratio, lowest maximum drawdown days and the lowest maximum drawdown percentage using OPTICS algorithm in both validation and test set. It is noteworthy to point the forecast models have better sharpe ratio in the test set compared to the standard mean reversion model. Overall, forecast spread based trading models tend to have an edge over the standard mean reversion model based on the test set.

Number of trades: An interesting observation from Tables 1 and 2 is that the number of trades executed by standard mean reversion models is extremely low compared to the forecast trading models but with very high percentage of positive trades. On the other end of the spectrum, we can find the ARIMA forecast trading model generates over 40,000 trades, and if we include transaction costs, it might nullify the positive returns.

Table 1: Trading performance measures for the validation set (Jan-Dec 2021)

	OPTICS SMR	OPTICS ARIMA	OPTICS RNN	HC SMR	HC ARIMA
No. of pairs	10	10	10	85	85
Annual ROI (in %)	13.04	4.24	8.05	17.56	4.46
Sharpe Ratio	2.75	7.11	2.46	4.86	3.84
Max. Drawdown days	18	3	11	3	24
Portfolio Max DD (in %)	-2.97	-0.12	-1.43	-1.06	-0.38
No. of Trades	65	1074	1685	475	48933
No. of Pos. trades	45	815	1212	392	29821
No. of Neg. trades	20	259	473	83	19112
% of Pos. trades	69.23	75.88	71.93	82.53	60.94

Table 2: Trading performance measures for the test set (Jan-Dec 2022)

	OPTICS SMR	OPTICS ARIMA	OPTICS RNN	HC SMR	HC ARIMA
No. of pairs	5	2	2	36	11
Annual ROI (in %)	20.96	22.96	24.84	21.69	36.58
Sharpe Ratio	1.87	9.07	2.77	3.09	4.32
Max. Drawdown days	23	1	49	31	3
Portfolio Max DD (in %)	-5.83	-0.45	-5.44	-3.38	-5.13
No. of Trades	20	1135	1165	151	40462
No. of Pos. trades	18	892	821	126	24116
No. of Neg. trades	2	243	344	25	16346
% of Pos. trades	90.00	78.59	70.47	83.44	59.60

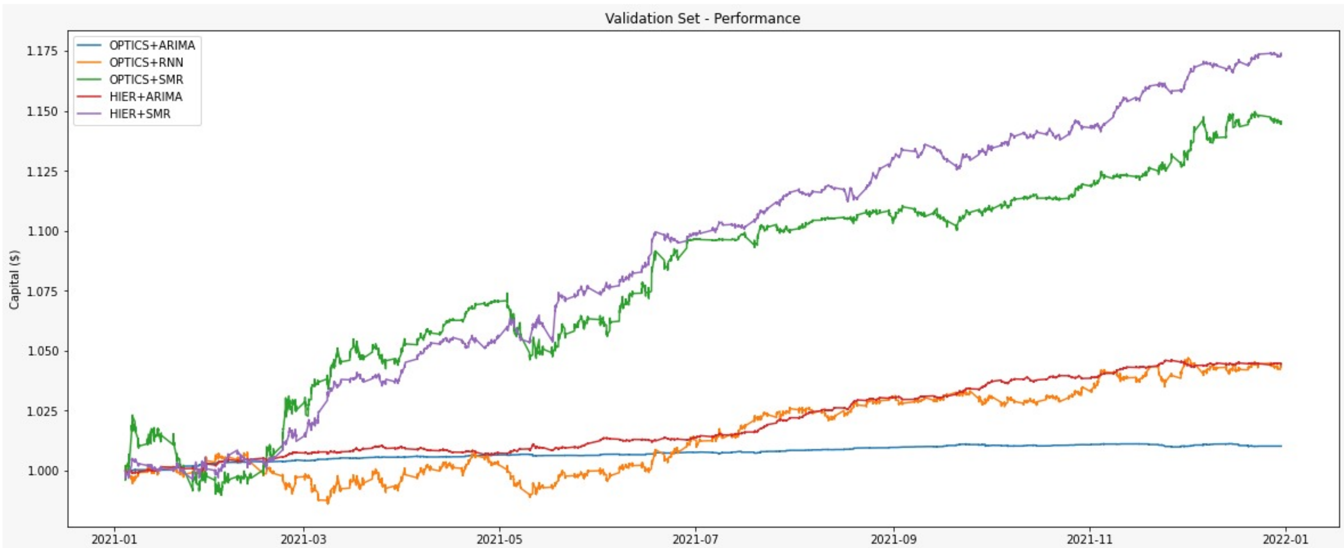

Figure 5: Total account balance for the validation set (Jan-Dec 2021)

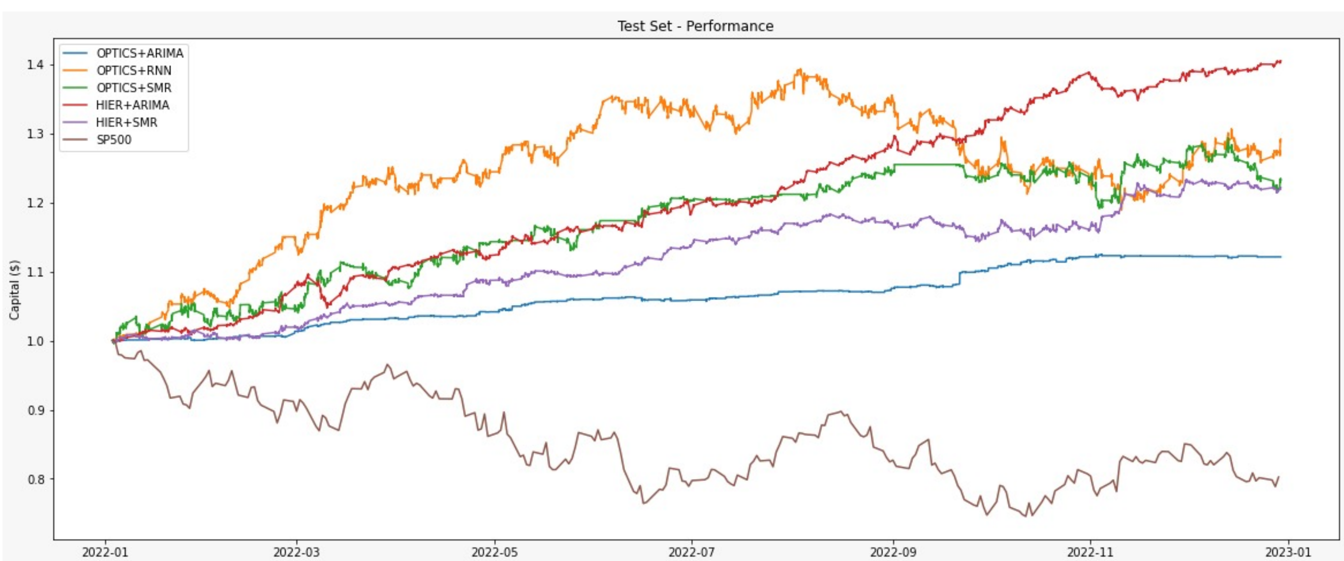

Figure 6: Total account balance for the test set (Jan-Dec 2022)

VI. CONCLUSION

Over the course of this project, we introduced machine learning in both stages of the Pairs Trading procedure - in the pair-selection step using OPTICS and Hierarchical clustering and in the trading step using ARIMA and LSTM. To begin with, OPTICS was more selective and chose only 10 pairs as opposed to Hierarchical clustering's more liberal selection of 85 pairs. Both trading models significantly improved the stability of the strategy by fetching considerably higher Sharpe Ratios compared to Standard Mean Reversion model. On average, OPTICS with LSTM yielded the best ROI and significant Sharpe Ratio compared to the other combinations. Hierarchical clustering paired with ARIMA yielded the best ROI of 36.58%, albeit at the cost of too many trades which would invite a high penalty if we were to consider transaction costs.

REFERENCES

[1] Andrewngai. *Understanding DBSCAN Algorithm and Implementation from Scratch*. 2020. URL: `https://rb.gy/1dgzyq`.

[2] Huafeng (Jason) Chen et al. "Empirical Investigation of an Equity Pairs Trading Strategy". In: *American Finance Association Meetings (AFA)* (2017). DOI: `https://doi.org/10.1287/mnsc.2017.2825`.

[3] Carlin C.F. Chu and Po Kin Chan. "Mining Profitable High Frequency Pairs Trading Forex Signal Using Copula and Deep Neural Network". In: (2018), pp. 312–316. DOI: `10.1109/SNPD.2018.8441125`.

[4] Evan Gatev, William Goetzmann, and K. Rouwenhorst. "Pairs Trading: Performance of a Relative-Value Arbitrage Rule". In: *Review of Financial Studies* 19.3 (2006), pp. 797–827. DOI: `https://doi.org/10.1093/rfs/hhj020`.

[5] Chulwoo Han, Zhaodong He, and Alenson Jun Wei Toh. "Pairs trading via unsupervised learning". In: *European Journal of Operational Research* 307.2 (2023), pp. 929–947. ISSN: 0377-2217. DOI: `https://doi.org/10.1016/j.ejor.2022.09.041`.

[6] Thomas A. Hanson and Joshua Hall. "Statistical Arbitrage Trading Strategies and High Frequency Trading". In: (2012), pp. 312–316. DOI: `10.2139/ssrn.2147012`.

[7] Hossein Rad, Rand Kwong Yew Low, and Robert Faff. "The profitability of pairs trading strategies: distance, cointegration and copula methods". In: *Quantitative Finance* 16.10 (2016), pp. 1541–1558. DOI: `10.1080/14697688.2016.1164337`.

[8] Simão Moraes Sarmento and Nuno Horta. "Enhancing a Pairs Trading strategy with the application of Machine Learning". In: *Expert Systems with Applications* 158 (2020), p. 113490. ISSN: 0957-4174. DOI: `https://doi.org/10.1016/j.eswa.2020.113490`.

[9] Chetan Tayal. "Application of Machine Learning in Market- Neutral Pairs Trading in the Indian Stock Exchange". In: 2021. URL: `https://www.jetir.org/papers/JETIR2107324.pdf`.

[10] Yufeng. *Understanding OPTICS and Implementation with Python*. 2022. URL: `https://rb.gy/wr4ox7`.

[11] Dingyi Zhou, Xiaoqing Zuo, and Zhifang Zhao. "Constructing a Large-Scale Urban Land Subsidence Prediction Method Based on Neural Network Algorithm from the Perspective of Multiple Factors". In: *Remote Sensing* 14 (Apr. 2022), p. 1803. DOI: `10.3390/rs14081803`.