NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Self-Organizing Maps for Trading Stocks

A Case Study on S&P 500

Carolina Sofia Antunes Caldeira

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Self-Organizing Maps for Trading Stocks

A Case Study on S&P 500

by

Carolina Sofia Antunes Caldeira

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics.

Supervised by

Fernando Bação, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 2025/06/20]

Carolina Caldeira

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to my supervisor, Professor Fernando Bação, for his excellent guidance and support throughout the development of my master thesis. I am truly grateful for the opportunity to work under his supervision. I would also like to express my gratitude to Farina Pontejos, for her revision and important inputs that enriched my thesis. Her attention to detail and important suggestions helped me to refine ideas and improve the research project. Finally, I would like to thank my family, colleagues at Deloitte and friends for their strong support and constant encouragement throughout this journey.

ABSTRACT

This thesis explores the use of Self-Organizing Maps (SOMs) for designing an adaptive trading strategy aimed at outperforming the traditional buy-and-hold approach. A Hierarchical SOM (HSOM) architecture is proposed, where technical indicators are grouped into four categories: trend, momentum, volume, and volatility, used to train separate SOMs. The outputs of these SOMs are then integrated into a final layer SOM, forming a hierarchical structure. Based on model's outputs, three distinct trading strategies were developed and backtested on a two-years out-of-sample S&P 500 data. Results show that all HSOM-based trading strategies outperformed the traditional buy-and-hold approach, in both return and risk-adjusted performance, with the Consistent Strategy achieving the highest final capital and return on investment. Furthermore, the HSOM demonstrated lower topographic error and superior results when compared to a benchmark SOM, reinforcing the effectiveness of a layered architecture for financial time series analysis. The component planes analysis further revealed meaningful associations between features and market regimes. Overall, this work contributes to the growing field of machine learning in finance by proposing an interpretable and data-driven framework for adaptive algorithmic trading strategy design.

KEYWORDS

Self-Organizing Maps; Trading Strategy; Technical Analysis; Financial Time Series; Algorithmic Trading

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AD	Accumulation/Distribution Index
ADOSC	Accumulation/ Distribution Oscillator
ADX	Average Directional Movement Index
AMH	Adaptive Market Hypothesis
ANN	Artificial Neural Networks
AOBV	Average On-Balance Volume
AP	Affiliate Propagation clustering
AR	Annualized Return
ATR	Average True Range
AUC	Area Under Curve Score
AUDJPY	Australian Dollar/ Japanese Yen
AUDUSD	Australian Dollar/ US Dollar
AVGL	Average Loss per Losing Trade
AVGP	Average Profit per Profitable Trade
AVGR	Average Return per Trade
AY	Annual Yield
B&H	Buy-and-Hold traditional strategy
BB	Bollinger Bands
BB_LW	Bollinger Bands Lower line
BB_MD	Bollinger Bands Middle line
BB_UP	Bollinger Bands Upper line
BMU	Best Matching Unit
BOP	Balance of Power
BP	Back Propagation neural network
C-E	Combination of Clustering and Ensemble learning

C-INDEX	Concordance Index
CAE	Convolutional Autoencoder
CALMR	Calmar Ratio
CCI	Commodity Channel Index
CH	Chaikin Oscillator
CR	Cumulative Return
CRISP-DM	Cross Industry Standard Process for Data Mining
CSI300	China Securities Index 300
CTI	Correlation Trend Indicator
DA	Direction Accuracy
DCPERIOD	Dominant Cycle Period
DCPH	Dominant Cycle Phase
DEA	Difference Exponential Average of MACD
DEC	Decreasing
DFSOM	Deep Fuzzy Self-Organizing Map algorithm
DJIA	Dow Jones Industrial Average index
DM	Diebold-Mariano test
DMA	Different of Moving Average
DX	Directional Movement Index
EBSW	Even Better Sine Wave
ECH	iShares MSCI Chile ETF
EM	Expectation-Maximization clustering
EMA	Exponential Moving Average
EMA_20	20-day Exponential Moving Average
EMA_200	200-day Exponential Moving Average
EMA_5	5-day Exponential Moving Average

EMA_60	60-day Exponential Moving Average
EMH	Efficient Market Hypothesis
EURGBP	Euro/ British Pound
EURJPY	Euro/ Japanese Yen
EWZ	iShares MSCI Brazil ETF
FLRs	Fuzzy Logic Relationships
FTS	Fuzzy Time Series
GC	Gray Correlation
GHSOM	Growing Hierarchical Self-Organizing Map
GMM	Gaussian Mixture Model clustering
GRU	Gated Recurring Unit
HOG	Histogram of Oriented Gradients
HSOM	Hierarchical Self-Organizing Map
IBS	Integrated Brier Score
INC	Increasing
INPH	In-phase component of phaser components
IR	Information Ratio
IVV	iShares Core S&P 500 ETF
KAMA	Kaufman Adaptive Moving Average
KDJ	Stochastic Indicator (derived from STOCH)
KM	K-means clustering algorithm
LSTM	Long Short-Term Memory Network
LSW	Lead Sine wave
LT_KDJ	Long-Term Stochastic Indicator
LT_MA	Long-Term Moving Average
MA14	14-day Moving Average

MA21	21-day Moving Average
MA7	7-day Moving Average
MACD	Moving Average Convergence Divergence
MACD_ SIGNAL	Moving Average Convergence Divergence Signal Line
MACD_HIST	Moving Average Convergence Divergence Histogram
MACD-ALMA	Hybrid MACD with Arnaud Legoux Moving Average
MAE	Mean Absolute Error
MAFE	Mean Absolute Forecast Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MDD	Maximum Drawdown
MFI	Money Flow Index
MLP	Multilayer Perceptron
MOM	Momentum indicator
MR	Maximum Retracement
MSE	Mean Squared Error
N225	Nikkei 225 Index
NART	Normalized Average True Range
NN	Neural Network
OBV	On-Balance Volume
P/L	Profit/Loss Ratio
PCA	Principal Component Analysis
PR	Profit Rate
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PVR	Price-Volume Rank

QE	Quantitative Evaluation
QUADR	Quadrature Component of phaser components
RF	Random Forest
RMSE	Root Mean Squared Error
RMSFE	Root Mean Squared Forecast Error
ROC	Rate of Change
ROCP	Rate of Change Percentage
ROR	Rate of Return
RRMSE	Relative Root Mean Squared Error
RSI	Relative Strength Index
RSI_14	14-day Relative Strength Index
RSI_21	21-day Relative Strength Index
RSI_7	7-day Relative Strength Index
S&P 500	Standard & Poor's 500 index
SMA	Simple Moving Average
SMA_14	14-day Simple Moving Average
SMA_20	20-day Simple Moving Average
SMA_200	200-day Simple Moving Average
SMA_21	21-day Simple Moving Average
SMA_40	40-day Simple Moving Average
SMA_5	5-day Simple Moving Average
SMA_50	50-day Simple Moving Average
SMA_60	60-day Simple Moving Average
SMA_7	7-day Simple Moving Average
SMA40	40-day Simple Moving Average
SMAPE	Symmetric Mean Absolute Percentage Error

SOM	Self-Organizing Map
SPDB	Shanghai Pudong Development Bank
SR	Sharpe Ratio
SSOM	Supervised Self-Organizing Map
ST_KDJ	Short-Term Stochastic Indicator
ST_MA	Short-Term Moving Average
STOCH	Stochastic Oscillator Indicator
STOCH_D	Stochastic Oscillator Smoothed
STOCH_K	Stochastic Oscillator
STOCHRSI_D	Stochastic Relative Strength Index Smoothed
STOCHRSI_K	Stochastic Relative Strength Index
SVM	Support Vector Machine algorithm
SVR	Support Vector Regression algorithm
SW	Sine wave
TAIEX	Taiwan Capitalization Weighted Stock index
TAKMV	Technical Analysis, K-Means clustering, Mean-Variance portfolio optimization
TEMA	Triple Exponential Moving Average
TEMA_20	20-day Triple Exponential Moving Average
TEMA_200	200-day Triple Exponential Moving Average
TEMA_5	5-day Triple Exponential Moving Average
TEMA_60	60-day Triple Exponential Moving Average
TN	Total Number of Trades
TN-	Number of Non-Profitable Trades
TN+	Number of Profitable Trades
TR	True Range
TRIX	Triple Exponential Average

U-Matrix	Unified Distance Matrix
V%	Volume Change (%)
VI-	Downtrend Vortex Indicator
VI+	Uptrend Vortex Indicator
VOL	Volume
VQ	Vector Quantization
WILLR	Williams %R
ZS	Z-Score

1. INTRODUCTION

In the era of economic globalization and rapid technology advancement, the generation and accumulation of financial data have grown at an extraordinary rate. Modern markets generate huge amounts of data, reflecting the complex influence of economic, political, and psychological factors on investors' behaviour. This exponential growth has far exceeded the capacity of manual analysis, making it imperative to adopt automated and intelligent methods in order to extract valuable insights.

Within this landscape, stock price movements are typically considered as a time series problem (Wang C., 2022). Financial time series data are characterized by non-linear patterns, long-term trends, cyclical fluctuations, and sudden irregularities, making them uniquely challenging to interpret (Niaki & Hoseinzade, 2013). Therefore, data mining and machine learning techniques have emerged as critical tools, enabling market participants to make timely and knowledge-driven decisions while mitigating risk (Dash & Dash, 2016)

One of the most significant transformations in modern finance is the rise of the algorithmic trading, also known as quantitative trading, which leverages advances in computing power, data availability, and machine learning to develop sophisticated trading strategies. These strategies aim to improve decision-making, reduce human biases, and enhance forecasting accuracy (Salehpour & Samadzamini, 2023). This shift has fundamentally reshaped how investment decisions are made, influencing both market behaviour and the financial industry (Wilhelmina et al., 2024).

In this context, Self-Organizing Maps (SOMs), an unsupervised learning algorithm proposed by Kohonen (1982), offer a powerful approach for clustering and visualizing high dimensional data. Their capacity to map complex input data into low-dimensional, structured representations, makes them particularly valuable for identifying patterns in market data and detecting shifts in market regimes (Deboeck & Kohonen, 2000; Kohonen, 2013; Kossakowski & Bilski, 2017). Despite their potential, the application of SOMs in trading systems remains relatively underexplored compared to supervised learning techniques.

This thesis explores the use of Self-Organizing Maps to develop a clustering-based trading strategy, using historical price and volume data enriched with technical indicators. By leveraging SOMs' capabilities, the goal is to provide investors with accurate information on the decision to buy, hold or sell a stock. The study focuses on the S&P 500 index, a leading benchmark that represents the performance of the 500 largest companies in the United States of America across diverse sectors. Its diverse composition and high liquidity make it a great focus for financial research and algorithmic trading applications.

The main research question this thesis seeks to address is: How can Self-Organizing Maps be utilized to design an adaptive trading strategy that outperforms traditional strategies such as Buy-and-Hold?

To answer this question, the research is structured around the following sub-questions and associated objectives:

- Does a layered architecture in Self-Organizing Maps improve the quality of the trading signals generated?

The objective is to compare a hierarchical SOM with a benchmark SOM, using the same technical indicators and similar model tunning, to evaluate whether a hierarchical structure enhances SOM model performance in financial applications.

- How effectively can Self-Organizing Maps cluster financial market behaviours based on categorized price and volume technical indicators?

The purpose is to assess the capability of SOMs to organize and visualize meaningful market behaviours using categorized technical indicators related to trend, momentum, volume, and volatility.

- Does the trading strategy based on Self-Organizing Maps outperform the traditional Buy-and-Hold strategy?

The goal is to evaluate the effectiveness of the SOM-based trading strategy by comparing its profitability and risk-ajusted returns with the buy-and-hold benchmark strategy.

2. RESEARCH BACKGROUND

2.1 FINANCIAL THEORIES

This subchapter consists of the discussion of some relevant financial theories that are used as a benchmark to analyse the results (see Figure 1). Firstly, the Random Walk Hypothesis (RWH) is presented as the foundational framework of the financial theories. It establishes the idea upon which modern financial theories are built. Afterwards, the fundamentals and critics of the Efficient Market Hypothesis (EMH) are discussed, which leads to the most recent approach, the Adaptive Market Hypothesis (AMH). This theory combines the principles of the Efficient Market Hypothesis with behavioural finance.

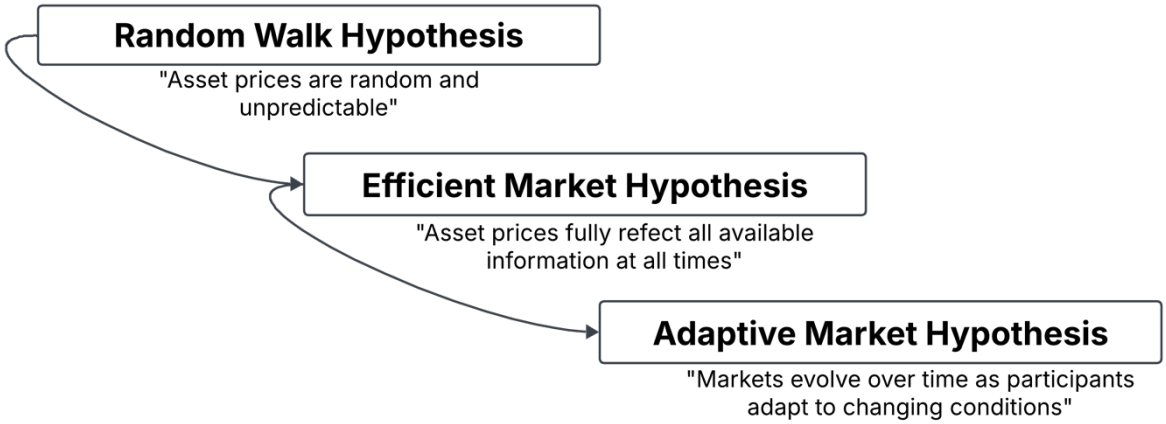

Figure 1 - Financial Theories

Paul Samuelson, an American economist, formalized the Random Walk Hypothesis (RWH) in 1965. He argues that since all available information is reflected in current prices, future price movements are driven solely by the arrival of new, unpredictable information. As a result, historical prices cannot be used to predict future ones (Samuelson, 1965). His work demonstrated that price randomness is not due to inefficiency or irrationality, but rather the consequence of an efficient market, where participants act on all available information.

Later, an economist named Eugene Fama provided empirical evidence supporting the RWH. His model makes two assumptions: successive price changes are independent, and past price changes provide no information about future movements. The New York Stock Exchange historical stock price data was used, and the results concluded that past price data cannot reliably predict future price movements, since price changes reflect new information, which arrives randomly and unpredictably (Fama, 1965). In summary, this theory challenges the validity of technical analysis.

Therefore, the Efficient Market Hypothesis (EMH) was formalized in 1970 by Eugene Fama. It is a foundational concept in financial economics that defends market prices, at any point in time, fully reflect all available information (Fama, 1970). This hypothesis implies that it is

impossible to consistently outperform the market through superior skill or timing, as asset prices already incorporate all publicly available and historical data. In practice, investment outcomes are driven by chance rather than expertise.

Fama categorized market efficiency into three distinct categories, based on the type of information reflected in prices:

- Weak-form efficiency: Market prices reflect only historical price information;
- Semi-strong efficiency: Prices adjust immediately and accurately to all publicly available information, such as earnings announcements or economic reports;
- Strong-form efficiency: Prices incorporate all information, both public and private, including insider knowledge, which makes it impossible for any individual to achieve abnormal returns based on exclusive information.

This framework highlights that assets are consistently priced at their fair value, enabling fair transactions and excluding consistent market outperformance based on informational advantages (Fama, 1970).

In the next 20 years, Fama did more research on market efficiency and asset-pricing models. While reaffirming the hypothesis, he introduced refinements to account for practical limitations. He acknowledged that the assumption of perfect efficiency, where information and trading costs are zero, is an idealized benchmark (Fama et al., 1991).

Other two economists, Stanford J. Grossman and Joseph E. Stiglitz, also argue that the market cannot be perfectly efficient, or there is no profit to gathering information which is so quickly reflected in market prices, what would lead to insufficient reason to trade, and markets would eventually collapse (Grossman & Stiglitz, 1980). Instead, a more realistic version of the EMH recognizes that prices reflect information only to the extent that the marginal benefits of obtaining and acting on the information exceed the associated costs. This pragmatic approach connects the EMH to real-world economic constraints, accommodating scenarios where market inefficiencies exist but are limited by the costs of exploiting them (Fama et al., 1991).

Fama also highlighted critical challenges in testing the EMH, particularly the joint-hypothesis problem, which states that any observed inefficiencies in the market might not necessarily reflect true inefficiencies, instead, they could indicate flaws or limitations in the chosen asset-pricing model. This issue complicates the interpretation of anomalies, making it unclear whether they reflect true market inefficiencies or deficiencies in the underlying pricing models (Fama et al., 1991).

Burton Malkiel critiques the rigidity of the EMH in explaining real-world market behaviours like short-term momentum, long-run return reversals, or predictable seasonal patterns (Malkiel, 2003). He defends that short-term momentums were never significant to guarantee returns, and they will never be useful for investors after they are known. Regarding predictable seasonal patterns, he says that they are very small effects, which are not worth it

in comparison to the transaction costs in trying to exploit them. As a result, he upholds that this market behaviours will occur over time and probably persist for short periods, however, he thinks that the belief that the stock market is remarkably efficient in its utilization of information will continue (Malkiel, 2003).

In addition, EMH has been widely criticized mostly by psychologists and behavioural economists, who argue that the theory is based on false assumptions regarding human behaviour. Some critics of the EMH state that investors are often irrational, and some behavioural biases discussed in the literature are overconfidence (Gervais, 2001), herding (Chiang & Zheng, 2010) and loss aversion (Rieger, 2022).

On the other hand, the MIT professor Andrew Lo, in 2007, introduces a more flexible framework to understand market behaviour, the Adaptive Market Hypothesis (AMH), a theory that incorporates behavioural finance and evolutionary psychology with the principles of EMH. It suggests that markets evolve over time as participants adapt to changing conditions.

Lo defends that behavioural biases are evolutionary heuristics that can lead to inefficiencies under certain conditions. He also suggests that the relationship between risk and return is not constant but evolves as market participants adapt to new challenges and opportunities. This perspective contrasts with the static risk-return trade-off implied by traditional models, and accommodates both efficient and inefficient market behaviours as part of a broader, adaptive process (Lo, 2007).

However, AMH faces some critics since it is a qualitative method. Although it outperforms the EMH in explaining the real-world situations of financial markets, it is necessary to establish a quantitative model for it (Li, Li & Xiao, 2021).

Despite all the research, theoretical and empirical, there is still no consensus among economists on the validity of the Efficient Market Hypothesis (Ayunku, 2020). However, it is an essential and important theory to understand in the context of this study, trading stocks.

2.2 Self-Organizing Map

This subchapter provides a comprehensive introduction to the Self-Organizing Map (SOM), the core model underlying this research. A thorough understanding of how this model operates is essential for correctly interpreting the methodology and results presented throughout this thesis. By presenting the fundamental principles, structure, and training procedure of the SOM, this section prepares the reader with the necessary conceptual foundation to follow the next chapters.

The SOM model builds upon the concept of classical vector quantization (VQ), a data compression technique introduced by Lloyd (1957). VQ partitions the input space, typically composed of multi-dimensional feature vectors, into a finite set of regions. Each region is represented by a codebook vector, which serves as the reference point for the vectors falling within that region. The primary goal in VQ is to minimize the mean quantization error, defined

as the distance between the input vector and its nearest codebook vector (Kohonen, 2013). For computational simplicity, the Euclidean distance is frequently used in this context.

Extending the principles of vector quantization, the Self-Organizing Map (SOM), also known as the Kohonen Map, was introduced by Teuvo Kohonen in the 1980s (Kohonen, 1982). SOM is an artificial neural network designed for unsupervised learning, that not only performs quantization but also preserves the topological properties of the input data, meaning that similar input vectors are mapped to adjacent regions in the output space. This topological ordering capability distinguishes SOM from classical VQ and makes it a powerful tool for data visualization, clustering, and dimensionality reduction.

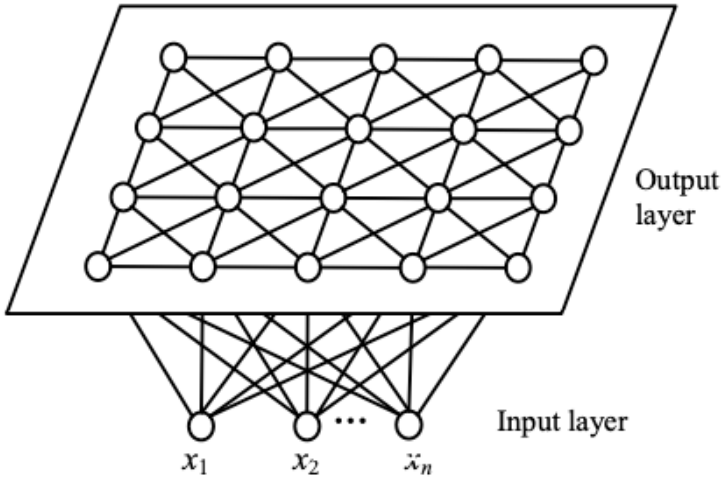

Figure 2 - Structure of SOM (Han & Wang, 2008)

Structurally, SOM consists of two layers: an n-dimensional input layer, and a typically two-dimensional output layer, often referred to as the map or grid of neurons (Mingoti & Lima, 2006). A visualization of the SOM structure can be found in Figure 2. Each neuron is represented by a weight vector of the same dimension as the input vectors. When input data is presented to the network, it is mapped onto this grid in a way that preserves the data's intrinsic structure. Commonly, the output neurons are arranged in regular topologies, with hexagonal grids being the most recommended due to their visual clarity and improved neighbourhood representation. In addition, the shape and size of the map can be optimized based on the distribution of the input data, with Kohonen (2013) suggesting that the dimensions should align with the two largest principal components of the dataset.

The SOM algorithm mathematically works as follows:

```
Let  X  be the set of n training patterns x1, x2,..xn
     W  be a pxq grid of units wij where i and j are their
        coordinates on that grid
     α  be the learning rate, assuming values in ]0,1[, initialized
        to a given initial learning rate
     r  be the radius of the neighborhood function h(wij,wmn,r),
        initialized to a given initial radius
1  Repeat
2    For k=1 to n
3      For all wij∈W, calculate dij = || xk - wij ||
4      Select the unit that minimizes dij as the winner wwinner
5      Upadate each unit wij∈W: wij = wij + α h(wwinner,wij,r) || xk - wij ||
6    Decrease the value of α and r
7  Until α reaches 0
```

Figure 3 - SOM algorithm (Bação et al., 2005)

The learning process of the SOM is competitive and iterative, as illustrated in Figure 3. It begins with the random initialization of the neuron's weight vectors. For each input vector, the algorithm identifies the Best Matching Unit (BMU), commonly referred to as the "winning neuron", which is the neuron whose weight vector is closest to the input, typically based on the Euclidean distance. Following this, the weight vector of the BMU, as the ones within its neighbourhood, defined by the neighbourhood function, are adjusted to the direction of the input data (Kohonen, 1990). The rate at which the weights are updated is the learning rate, which reduces through every iteration, guaranteeing convergence of the algorithm to an optimum value of the error function (Mingoti & Lima, 2006). Similarly, the radius of the neighbourhood function also decreases as the training progresses. By the end of training, weight updates are applied only to the winning neuron, reflecting a transition from global to local learning (Deboeck & Kohonen, 2000).

Through this process, the SOM evolves from an initially unstructured state to a coherent map where similar inputs are clustered together in neighbouring regions, making it an intuitive and interpretable tool for pattern recognition and data exploration. It is particularly suitable for the financial applications explored in this thesis, where the identification of structures in multi-dimensional data is crucial for the development of robust trading strategies.

3. LITERATURE REVIEW

Developing an adaptive trading strategy system based on Self-Organizing Maps presents a significant challenge due to the complexity of financial data, which makes uncovering hidden patterns and generating profitable trading strategies a difficult task. Previous studies developed several methods to analyse and cluster stock market data based on historical prices and technical indicators. These studies serve different purposes, including predicting price movement, portfolio optimization and stock price forecasting. To report the most appropriate and representative past studies for this thesis' scope, the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) was applied.

3.1 PRISMA FRAMEWORK

PRISMA is a structured approach design for conducting and reporting systematic literature reviews and meta-analyses. It ensures transparency, rigor and reproducibility when documenting existing literature, making it an excellent guideline for academic research (Page et al., 2021).

Figure 4 - PRISMA process

It begins with the retrieval of relevant studies by querying databases (e.g. Scopus). This is accomplished by using selected keywords (e.g. "trading stock", "technical analysis"), Boolean operators, and filters to narrow down the search results. In the screening phase, duplicate records are removed, and studies are selected based on publication date and their relevance to the research topic. This is done by analysing the title and abstract of each study. Then, a thorough full text assessment is conducted to ensure remaining studies meet the established research criteria. Finally, in the inclusion phase, the studies that satisfy all requirements are selected and key information (e.g. methods, findings, conclusions) is extracted for further analysis and presentation. The process is summarised on Figure 4.

Financial time series data are highly dimensional and nonlinear, making it challenging to detect market regimes and develop profitable trading strategies. A thorough review of existing methodologies is essential to address these complexities. The PRISMA framework helps ensure a comprehensive and unbiased selection of studies by systematically identifying, screening, and assessing relevant research, reducing the risk of overlooking key methodologies.

3.1.1 Search Queries and Filters

As part of the Identification phase, Scopus and Google Scholar academic databases were queried. The search queries and filters present in Table 1 were used to retrieve relevant studies.

Table 1 - Search Queries and Filters

Database	Search Query	Filters
Google Scholar	("trading stock" OR "stock trading" OR "clustering stock" OR "stock clustering" OR "cluster stock" OR "stock cluster") AND ("unsupervised learning" OR "cluster" OR "SOM" OR "Self Organizing Map" OR "Self-Organizing Map") AND ("technical analysis" OR "technical-analysis" OR "technical indicators") AND ("buy-and-hold" OR "buy and hold")	Publication years: 2022-2025, ordered by relevance
Scopus	("trading stock" OR "stock trading" OR "clustering stock" OR "stock clustering" OR "cluster stock" OR "stock cluster" OR "stock") AND ("unsupervised learning" OR "cluster" OR "SOM" OR "Self Organizing Map" OR "Self-Organizing Map") AND ("buy-and-hold" OR "buy and hold" OR "technical analysis" OR "technical-analysis" OR "technical indicators")	Search within title, abstract, and keywords for publication years: 2020-2025

3.1.2 PRISMA Flow Diagram

The PRISMA flow diagram is a visual representation of the selection process. It includes the number of records identified, the number of records screened, the reasons for exclusion, and the final selection of studies. Figure 5 displays the flow diagram in the context of this thesis.

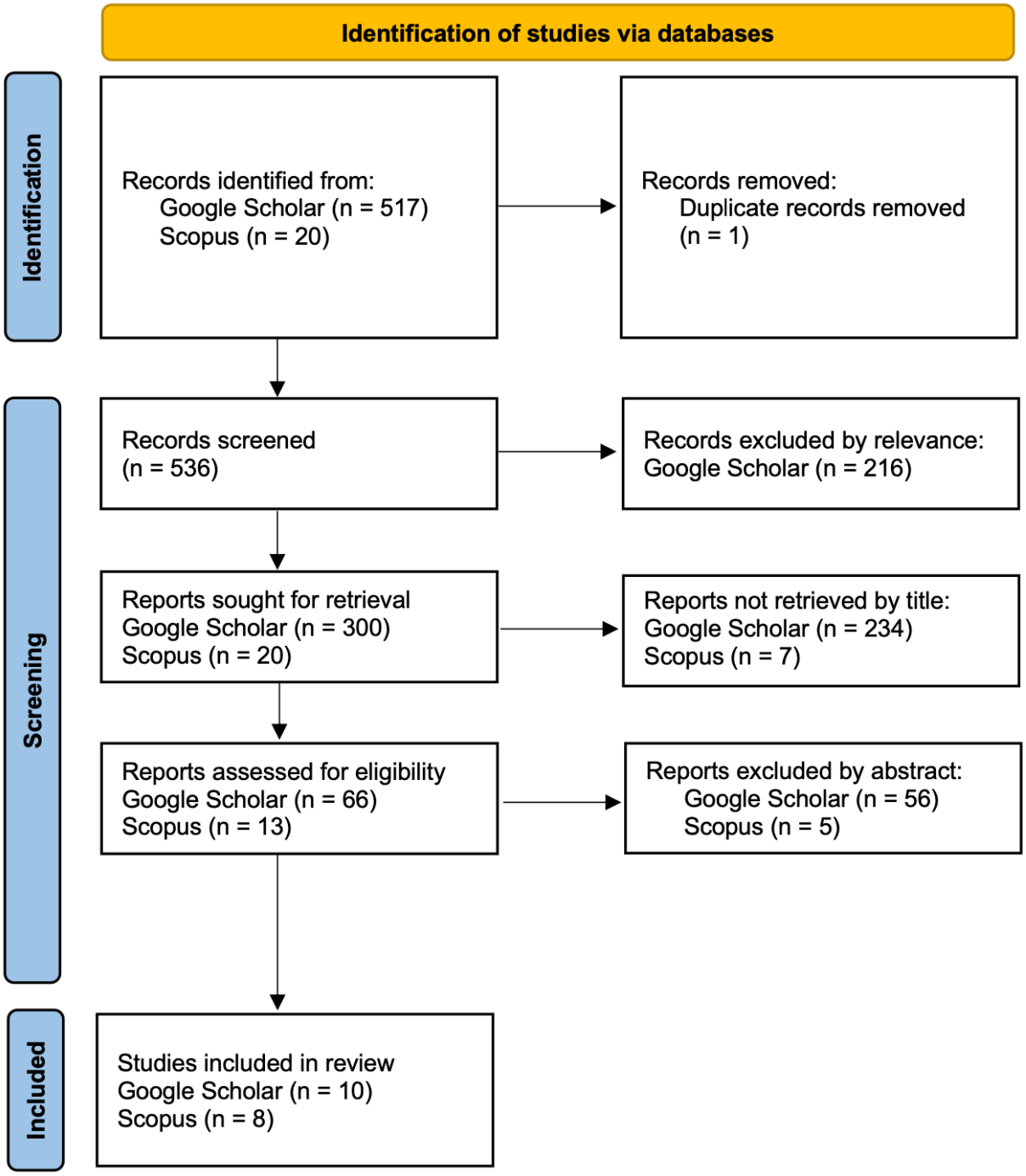

Figure 5 - PRISMA Flow Diagram

During the identification phase, the search queries and filters outlined on Table 1 were applied to the selected databases, resulting in a total of 537 studies, with 517 retrieved from Google Scholar and 20 from Scopus. After removing a duplicate record, the remaining studies proceeded to an evaluation based on their titles. In the Scopus database, out of the initial 20 papers, 13 were selected based on their relevance to the research topic. In contrast, for the Google Scholar database, an additional filtering was applied to order the papers by relevance, meaning that the results were sorted by keywords relevance, journal impact factor, number of citations, and publication recency, ensuring that the most appropriate studies appeared at the beginning of the search results. Consequently, only the first 300 records were analysed,

leading to the selection of 66 papers based on their titles. A considerable number of studies were excluded because they focused only on stock market prediction, a supervised learning approach aimed at forecasting prices or returns. This research, however, focuses on unsupervised learning to identify trend patterns in financial time series using technical analysis, generating trading signals rather than predicting prices directly. After the initial title-based screening, a total of 79 papers were selected for further evaluation. Subsequently, a detailed analysis of the abstracts was conducted to assess their alignment with the research criteria. This process resulted in a final selection of 19 studies, comprising 11 from Google Scholar and 8 from Scopus, as they best meet the research objectives and scope of the thesis. To conclude, an in-depth full text analysis of the 17 papers and 1 master thesis was conducted to ensure they satisfy all requirements, and key information was retrieved and is presented in the next sub-section.

3.1.3 PRISMA Results

As a result of this thorough search process, the final 18 studies are presented in Table 2, highlighting the authors' names, published year, article name, publication type, source and sources' 2023 impact factor when possible. The papers are ordered alphabetically by the authors' names for easier consultation.

Table 2 - Summary of selected studies

Authors	Year	Article Name	Type	Source	Impact Factor
Akbarzadeh, F., & Soleimani, A.	2023	Forecasting financial time series trends by pattern recognition	Paper	International Journal of Nonlinear Analysis and Applications	0.64
Bardi, A., & Takacs, M.	2023	Integrating Technical Analysis and Neural Networks for Optimizing Algorithmic Trading	Paper	IEEE 23rd International Symposium on Computational Intelligence and Informatics	3.4
Bing, H., Zhou, Y., Yuan, Z., Cheng, K., Wang, Y., Liu, H., & Wang, L	2022	A Study on Quantitative Investment Strategies Based on Cluster Analysis	Paper	IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)	3.4
Guan, B., Zhao, C., Yuan, X., Long, J., & Li, X.	2024	Price prediction in China stock market: an integrated method based on time series clustering and image feature extraction	Paper	The Journal of Supercomputing	2.5
Guo, Y.	2020	Stock Trading Based on Principal Component Analysis and Clustering Analysis	Paper	IOP Conference Series: Materials Science and Engineering	0.5
Hu, W., Zhou, J., Hu, W., & Zhou, J.	2024	Building Technical Analysis Strategies Using Multivariate	Paper	Computational Economics	1.9

		Longitudinal and Time-to-Event Data in Stock Markets			
Li, X., & Wu, P.	2022	Stock Price Prediction Incorporating Market Style Clustering	Paper	Cognitive Computation	4.3
Li, X., Liu, Q., Hu, Y., & Liu, H.	2024	The Double-Layer Clustering Based on K-Line Pattern Recognition Based on Similarity Matching	Paper	Information	3.1
Navarro, M. M., Young, M. N., Prasetyo, Y. T., & Taylar, J. V.	2023	Stock market optimization amidst the COVID-19 pandemic: Technical analysis, K-means algorithm, and mean-variance model (TAKMV) approach	Paper	Heliyon	3.4
Pei, D., Luo, C., & Liu, X.	2023	Financial trading decisions based on deep fuzzy self-organizing map	Paper	Applied Soft Computing	7.2
Sagaceta-Mejía, A. R., Sánchez-Gutiérrez, M. E., & Fresán-Figueroa, J. A.	2024	An Intelligent Approach for Predicting Stock Market Movements in Emerging Markets Using Optimized Technical Indicators and Neural Networks	Paper	Economics	0.8
Sarainmaa, O.	2024	Swing Trading the S&P 500 Index with Technical Analysis and Machine Learning Methods with Responsible Way	Master Thesis	Faculty of Social Sciences, Business and Economics, and Law	-
Sáenz, J., V., Quiroga, F. M., & Bariviera, A. F.	2023	Data vs. information: Using clustering techniques to enhance stock returns forecasting	Paper	International Review of Financial Analysis	7.5
Shi, Y., Li, B., Du, G., & Dai, W.	2021	Clustering framework based on multi-scale analysis of intraday financial time series	Paper	Physica A: Statistical Mechanics and its Applications	2.8
Wang, C.	2022	Pattern Classification of Stock Price Moving	Paper	Frontiers in Computing and Intelligent Systems	-
Wu, H., Long, H., Wang, Y., & Wang, Y.	2021	Stock index forecasting: A new fuzzy time series forecasting method	Paper	Journal of Forecasting	3.4
Wu, S.	2020	Application of Cluster Analysis in Stock Selection in United States Stock Market	Paper	ACM International Conference Proceeding Series	0.58
Xu, Y., Yang, C., Peng, S., & Nojima, Y.	2020	A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning	Paper	Applied Intelligence	3.4

Beginning with papers that explored various applications of Self-Organizing Maps (SOM), Akbarzadeh & Soleimani (2023) introduced a Supervised Self-Organizing Map (SSOM) model, that used a correlation coefficient as the distance metric. This model was designed to recognize Elliot Wave patterns, subsequently utilizing them to predict trends in financial time series. It was trained on twelve predefined Elliot Wave patterns. To enhance prediction accuracy, the simple moving average (SMA) of prices was employed, instead of raw price data, to smooth price fluctuations. After testing various time periods, a 40-day SMA was selected as optimal, effectively improving forecasting accuracy, while still retaining sensitivity to price changes. Another feature engineering technique used was the 10-days sliding windows, shifted by one day, for pattern detection. The proposed model, tested on a 13-year period (2001-2014) for the four Forex selected indices, reached an average accuracy of 93.94%. The authors state that it shows significant improvement from others' work.

A distinct SOM approach was proposed by Pei et al. (2023), which consists of a Deep Fuzzy Self-Organizing Map (DFSOM) combined with Gated Recurrent Unit (GRU) networks to analyse intraday time series price movements. These authors used minute-level data to construct the extended candlestick charts, which contain not only the price trend of the stock, but also the corresponding volume. For China's commodity futures (2020) they applied 30-minute intervals and for the forex market (2016), 60-minute intervals were employed. The Histogram of Oriented Gradients (HOG) was applied to reduce redundant information, as it can remove irrelevant information from an image. Sliding windows were used in feature extraction. A HOG description was given to each window, which served as input to the parallel layer of the suggested DFSOM model. The proposed model consists of two layers. The first layer is composed of two parallel FSOMs, and each of them processed different HOG features, extracted from overlapping sliding windows. The second layer is the DFSOM classification layer, where the clustering of combined features from the first layer created representations of movement patterns. In the end, GRU networks were employed to take the DFSOM-clustered features as inputs and predict future prices.

Additionally, comparative studies with six alternative models, including traditional neural networks like Back Propagation (BP), were performed. The results showed that the DFSOM-GRU model outperformed all the other baseline models, producing higher profit rates, greater accuracy and stability, and stronger profit/loss ratio, demonstrating its effectiveness for short-term trading strategies.

Following the same intraday financial time series scope, Shi et al. (2021) developed a clustering framework using multi-scale analysis to identify potential patterns of price movements. Five years data (2015-2020) from the S&P 500, CSI300 and N225 stock indices, representative of America, China and Japan financial markets, respectively, were analysed. The results indicate that linear and nonlinear correlations in the data' intraday time series exist, which suggests that price movements may follow recurring but non-identical patterns. Furthermore, K-Means (KM), SOM and Chameleon clustering algorithms were applied to verify

the recommended framework. Through the evaluation of stability, compactness within clusters and separation between clusters, the authors concluded that SOM is the best performing model, showing the highest stability and cluster separation. Moreover, although the proposed multi-scale similarity measurement did not considerably impact the SOM performance metrics, since the model is inherently effective at handling data dimensionality, it does help in reduce computational cost.

These studies collectively highlight SOM's strength in pattern recognition, clustering and trading decision-making, while also proposing advanced feature extraction techniques to enhance predictive capabilities (Akbarzadeh & Soleimani, 2023; Pei et al., 2023; Shi et al., 2021).

Continuing to unsupervised learning studies employing other techniques besides SOM, Bing et al. (2022) presented a two-stage clustering model and Gray correlation analysis for portfolio selection, avoiding high correlation among selected stocks. A Pearson correlation coefficient-based analysis was performed on 37 technical indicators, which resulted in five categories. This was followed by a two-stage clustering process, beginning with initial categorization via KM clustering, and refined by Gaussian model clustering. Ultimately, Gray correlation analysis was implemented to identify the most representative indicator for each cluster. These five selected indicators, referenced in Table 4, serve as the basis for generating buy and sell signals. The model was tested on the Shanghai Stock Exchange A-shares, which 16 stocks were selected, and returns were computed to evaluate its performance. The authors concluded that this method is appropriate for short-term trading, since, in one month, 14 out of 16 stocks generated positive returns. However, it is not for long-term, where returns were close to 0 in one year period. It is important to acknowledge that the short-term dataset only covers the period from June 1 to June 9, 2015, which may not accurately reflect market conditions.

Following the same goal and based on technical indicators, Guo (2020) investigated how Principal Component Analysis (PCA) and KM clustering could improve stock selection. The authors conducted the experiment using nine technical indicators, referenced also in Table 4, on 10 years of monthly data of 4000 stocks in the HS300 index. PCA was first applied to extract principal components that best explain variance in the stock data, and KM was used for stock grouping. Evaluating the model with Annualized Return rate (AR), Sharpe Ratio (SR), Information Ratio (IR) and Maximum Drawdown (MDD), the study found that the constructed portfolio is significantly better than the HS300 benchmark index. AR is higher than the market (AR=15.2%), SR is 1.12, which indicates a good risk-adjusted performance, IR is 1.21, meaning the strategy outperformed market consistently, and MDD is 0.53, which represents an acceptable level of volatility and risk.

In a contrasting approach, Navarro et al. (2023) addressed the same goal of portfolio selection and optimization, through technical analysis and machine learning techniques. The Technical Analysis, K-Means algorithm, and Mean-Variance model (TAKMV) was proposed. It consists of applying the Moving Average Convergence/ Divergence (MACD) and the MACD with Arnaud

Legoux Moving Average indicators to identify stocks with strong momentum trends. The ones with positive Annual Rate of Return (ROR) were selected for further analysis. KM clustering grouped stocks based on Rate of Return (ROR) and Average Annual Risk (AVGR), using the Elbow method for selection of optimal number of clusters. Finally, portfolio optimization selected up to 10 highest ROR stocks per cluster, with weights determined by minimum variance optimization. The model was assessed on the Philippine Stock Exchange (PSE) market data, in 2018 and 2020, and the model performance was validated through comparison to next year's historical price. This study concluded that, regarding the number of assets with positive ROR, MACD was more effective pre-COVID-19, while MACD-ALMA performed better during COVID-19, a highly volatile market regime. In addition, validation confirmed that MACD-based portfolios outperformed MACD-ALMA in long-term stability.

A similar work introduced by Wu (2020) based on the KM algorithm, aimed to reach the same goal as the previous papers. The authors compared the constructed portfolio to the benchmark S&P 500 index. The experiment used 7-years 5175 stock market United States (US) data and 3 technical indicators: MACD, KDJ and MA. Key findings from this research indicate that cluster-based stock selection significantly outperformed the S&P 500, resulting in an AR of 16.4%, while S&P 500' AR was of 8-10%, SR of 1.11, IR of 1.21, and MDD of 0.42.

In the last paper, on the use of unsupervised learning methods in stock trading, Wang (2022) compared three clustering models' performance in identifying patterns in stock price movement and grouping of similar stocks. The models were the KM, Expectation-Maximization (EM), and Canopy with KM algorithms, and they were tested on S&P 500 time series market data. Several clustering indices, referenced in Table 3, were used for clustering evaluation. Wang (2022) concluded that stock price movements can be effectively classified using clustering approaches, with EM performing best. However, EM was computationally expensive, which makes it less preferred for large datasets or real-time trading, on the other hand, Canopy with KM was a good balance between accuracy and speed, making it more suitable for real-time applications. Regarding KM, it was observed that it is insufficient for detection of complex stock movement patterns, due to its reliance on the Euclidean distance, which does not handle complex time series relationships well.

These five studies highlight the potential of unsupervised clustering methods in accurately predicting stock price movements and constructing efficient portfolios, thereby facilitating the development of optimal trading strategies. However, they also demonstrate the need to be careful when selecting timeframes and periods to test the models, as it should test the adaptability of the model to distinct market conditions (Bing et al., 2022; Guo, 2020; Navarro et al., 2023; Wu, 2020; Wang, 2022).

Proceeding to different types of methods, which combine unsupervised techniques with supervised models, Li & Wu (2022) focused on incorporating market styles into prediction models, considering that stocks exhibit different behaviours under different market conditions. The study recommends a hierarchical clustering approach to categorize market

styles, based on technical indicators and news sentiment features. The data was filtered by market styles, and historical data from similar market styles was selected to train a Support Vector Machines (SVM) model. This approach was compared to a standard SVM. According to the results, incorporating market styles leads to better predictions, as the proposed model improved accuracy and F1-score up to 9% compared to the baseline model.

On the other hand, Guan et al. (2024) developed a model to enhance stock price prediction accuracy. It combined KM with Dynamic Time Wrapping to identify stocks with high correlation to the target, Convolutional Auto Encoder (CAE) for image feature extraction from candlestick charts, and a double-layer long short-term memory (LSTM) network for forecasting, processing both numerical stock data and image features. The authors used Root Mean Squared Error (RMSE), Direction Accuracy (DA) and R-Squared (R^2) to compare their KM-CAE-LSTM hybrid approach with some baseline models, like a single-layer LSTM. The hybrid model outperformed single models, reducing RMSE by 17.2% compared to a standard LSTM. The stock image features extracted by the CAE, and the utilization of highly correlated stocks, led to increased predictive accuracy.

Other researchers, Wu et al. (2021), have suggested a fuzzy time series (FTS) forecasting model that integrates technical analysis, Affinity Propagation (AP) clustering, and Support Vector Regression (SVR) with the same goal of improving stock forecasting. The model, AP-FTS-SVR, was evaluated on TAIEX, S&P 500 and DJIA indices, and compared to a baseline FTS model, AP-FTS, and FTS with technical indicators analysis method. Key findings are that AP clustering reduces forecasting errors by optimizing the partitioning of data into meaningful fuzzy intervals, technical indicators enhance prediction accuracy, and SVR error learning refines predictions. The suggested model outperforms the others, achieving the lowest Root Mean Squared Forecast Error (RMSFE) and Mean Absolute Forecast Error (MAFE) across the datasets.

It can be inferred from these previous studies that incorporating clustering techniques improves stock prediction, and Sáenz et al. (2023) explored this use of clustering methods to enhance stock returns forecasting. KM using distinct distance metrics was firstly applied to group stocks into clusters. Consequently, each stock's model was trained with data from its own cluster rather than the entire market. Four years' financial reports, price movements, and daily returns data of 240 company stocks, from Russel 3000 index, was used. In order to reduce noise, a 5-day moving average was applied. Three forecasting scenarios were tested, single-stock, multi-stock and cluster-based models, which used, respectively, the target stock's historical prices, the prices of all 240 stocks, and only data from stocks within the same cluster. ARIMA and LSTM were trained on these scenarios and evaluated mainly through Mean Absolute Scaled Error (MASE) and Symmetric Mean Absolute Percentage Error (SMAPE). The results indicated that clustering boosts stock price prediction by reducing noise in the data, allowing models to focus on the most relevant information. The best performing model was

the cluster-based LSTM, which generated positive returns, and outperformed traditional strategies like buy-and-hold (B&H) and MACD.

Other authors, such as Xu et al. (2020), also explored the application of clustering techniques but integrated them with ensemble learning. In their study, KM was applied based on technical indicators and historical price data from four Chinese stocks covering 2008-2019. The quality of clusters was measured using the Silhouette Coefficient. Several models were tested, including two-stage baseline models (SVR-SVR and SVR-RF, where SVR stands for Support Vector Regression and RF refers to Random Forest); clustering-based variants of these models (added KM clustering to both two-stage approaches); bagging ensemble models applied to SVR and RF (E-SVR&RF); and the hybrid proposed model, combining KM clustering with ensemble learning (C-E-SVR&RF). This hybrid model used the best clustering results as inputs for the ensemble approach. Evaluation was based on Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Relative Root Mean Squared Error (RRMSE) and Mean Squared Error (MSE). Time series prediction of 1, 5, 10, 20 and 30 days was performed. The outcomes enabled the authors to conclude that ensemble learning significantly enhances predictive power, as the model E-SVR&RF consistently outperformed other models, nevertheless, the hybrid approach C-E-SVR&RF performed best, achieving the highest accuracy on 3 out of 4 stocks. Furthermore, the model was more suitable for short term predictions, as forecasts for 30 days ahead were less accurate than 1-5 days ahead.

Lastly, Hu et al. (2024) suggested a different approach where they use Multivariate Functional Principal Component Analysis (MFPCA) to extract longitudinal informative features that can replace technical indicators. These features serve as inputs to survival models (SA) like Cox Proportional Hazards (Cox PH), Deep Logistic Hazard (DeepLH) and DeepHit, along with historical price data and event occurrence timing data, where the event was a 5% cumulative return (positive or negative) after a trading signal. Historical daily trading data of 385 securities from S&P 500 index was used, with data from 2018 to 2021 for training and 2022 for testing. The proposed MFPCA-SA model was evaluated using 5-fold time series cross-validation. The Concordance Index (C-index) and the Cumulative Annualized Return (CR) were the main performance metrics. In addition, the 2022 testing period was divided into two phases, to represent bull and bear market conditions. Using MFPCA instead of technical indicators, simplified models while maintaining predictive power, as the MFPCA models outperformed both technical indicator-based models and full models (included all technical indicators). Moreover, MFPCA-based strategies outperformed the benchmark models in 95% of evaluated cases. Nevertheless, MFPCA should be applied carefully to ensure it captures relevant stock movement patterns, due to its high dependency on the training process, which means that, for example, if the time windows are improperly selected, the extracted features may be irrelevant.

These six studies collectively highlight the effectiveness of unsupervised techniques in enhancing prediction accuracy (Li & Wu, 2022; Guan et al., 2024; Wu et al., 2021; Sáenz et al.,

2023; Xu et al., 2020; Hu et al., 2024). In each paper, models incorporating clustering methods consistently delivered the best performance. Additionally, there was notable emphasis on the role of ensemble learning in further improving forecasting results. Nonetheless, this hybrid models are computationally expensive, time consuming and produce results that are difficult to interpret. These limitations should be carefully considered when planning their development.

Moving on to the final set of papers, which employ supervised learning models based on technical indicators, Bardi & Takacs (2023) presented an approach integrating technical analysis with neural networks to improve prediction accuracy and enable the development of more effective trading strategies. The proposed simple neural network has three layers, and it was trained using the "Adam" optimizer and the binary cross-entropy loss function. Upon completing experiments using the Tesla stock market data, the best four technical indicators were selected as inputs to the model. Moreover, the Kalman Filter was applied to smooth stock prices, reduce noise, and mitigate overfitting. The target variable indicates whether the closing price on a specified day will increase or decrease the following day, which was used to generate buy and sell signals. To assess the model's performance, the profit made by the neural network, from January 2022 to August 2023, was compared to the profit made by a benchmark strategy, buy-and-hold, which holds the position throughout the entire period. The model strategy outperformed the benchmark by 153.26%, providing accurate buy and sell signals.

In the next year, Sagaceta-Mejía et al. (2024) also suggested a neural network model with technical indicators as inputs, to predict the stock trend direction in emerging markets. However, they proposed a distinct approach to select the most relevant technical indicators and a different neural network. A comprehensive set of 210 technical indicators were computed first, and their selection was based on 8 statistical measures, including Pearson's Correlation, Dispersion Ratio and Principal Feature Analysis, which combined PCA with KM. Class assignment was also applied for the model to learn the relationship between the technical indicators and the next day opening price movement. The neural network consisted of a Multilayer Perceptron (MLP) trained using 10-fold cross validation. Inputs included a subset of selected technical indicators along with stock price data, aiming to predict whether the next day's opening price would increase or decrease. The model was tested on 10 years of data (2010-2020) from three funds: iShares MSCI Chile ETF (ECH), iShares MSCI Brazil ETF (EWZ), and iShares Core S&P 500 ETF (IVV). By assessing the results, the model using only the selected features showed an average accuracy improvement of 13.63%, and an average reduction in training time of 84.68%, compared to the model using all features.

Finally, closely related to the goal of this thesis, Sarainmaa (2024) conducted a master's thesis to evaluate whether a Random Forest (RF) model, combined with technical indicators, could generate higher returns, compared to the traditional buy-and-hold strategy. Historical price data from the S&P 500 was used, with training data covering 2010-2017, and test data

spanning 2018-2019. A limited set of predefined technical indicators was computed, and seasonal patterns were considered too. In addition, the Balanced Random Forest Classifier handled class imbalance, and Grid Search for hyperparameter tuning. Prices were scaled relative to each day's opening price, and buy and sell signals followed a specific criteria: a buy signal was labelled, if the price at a given point was 2% higher than five weeks earlier, and remained 2% higher five weeks later, while a sell signal followed the opposite logic. If neither criteria was met, a hold signal was assigned. The performance of the suggested model was compared against a Logistic Regression (LR) model and further evaluated by backtesting it against a buy-and-hold strategy. Results showed that the RF model using technical indicators outperformed LR. The proposed model also generated higher returns than the B&H strategy, although, these increased returns were not consistent across all periods. Interesting future work proposed by Sarainmaa was the combination of moving averages of different periods to catch the crossing effect of them. The author also suggested the use of more features, especially candlestick chart patterns.

These studies emphasize the power of technical analysis and the potential of feature selection in enhancing prediction accuracy and reduction of computational resources (Bardi & Takacs, 2023; Sagaceta-Mejía et al., 2024; Sarainmaa, 2024).

The following Table 3 presents a summary of the feature engineering techniques, the proposed model and the performance metrics used in each study. All presented models, and some metrics and methods are represented by acronyms, their full translations are available for consultation in the List of Abbreviations and Acronyms.

Table 3 - Summary of Feature Engineering, Proposed Model and Performance Metrics

Reference (Authors, Year)	**Feature Engineering**	**Proposed Model**	**Performance Metrics**
Akbarzadeh & Soleimani, 2023	10 day sliding windows, Min-Max normalization, Elliot Waves pattern definition	SSOM	Accuracy
Bardi & Takacs, 2023	StandardScaler normalization, label target variable (y)	NN	Profit (NN vs B&H)
Bing et al., 2022	Pearson correlation coefficients	KM-GMM-GC	Total Revenue
Guan et al., 2024	StandardScaler for KM, Min-Max for LSTM, candlestick charts	KM-CAE-LSTM	RMSE, R-Square, DA
Guo, 2020	PCA	PCA-KM	AY, MR, SR, IR, QE
Hu et al., 2024	Min-Max normalization, MFPCA, survival data construction	MFPCA-SA	C-INDEX, IBS, CR, MDD, SR, CALMR

Li & Wu, 2022	Time window segmentation, feature representation and normalization, hierarchical clustering	SVM	Accuracy, F1-score
Li et al., 2024	Normalization of candlestick shapes, K-line sequence similarity matching	KM	Accuracy, CR
Navarro et al., 2023	Compute ROR and AVGR, Min-Max normalization	TAKMV	ROR, portfolio Expected Return and Risk
Pei et al., 2023	Extended candlestick charts, HOG, sliding windows, Min-Max normalization	DFSOM-GRU	PR, TN, TN+, TN-, AVGR, AVGP, AVGL, P/L, Accuracy
Sagaceta-Mejía et al., 2024	Feature selection with statistical measures, Min-Max normalization, class assignment	MLP	Accuracy
Sarainmaa, 2024	Prices scaled, Balanced RF Classifier for data imbalance, label the data, seasonal patterns analysis	RF	Confusion matrix, AUC, Precision, Recall, F1-score, Accuracy, AR, SR, MDD
Sáenz et al., 2023	KM, data smoothing MA5	KM-LSTM	MASE, SMAPE, Accuracy, Performance
Shi et al., 2021	Normalization, multi-scale transformation	SOM	Cluster stability, compactness and separation
Wang, 2022	Normalization	KM, EM, Canopy-KM	Indices: Calinski-Harabasz, Davies-Bouldin, Silhouette, Gap, Log-Likelihood; and Time Cost
Wu et al., 2021	Fuzzification through AP, FLRs, Defuzzification, error learning additional feature	AP-FTS-SVR	RMSFE, MAFE, DM
Wu, 2020	Standardization of technical indicators	KM	AR, SR, IR, MDD
Xu et al., 2020	Min-Max normalization, KM, silhouette coefficient analysis	C-E-SVR&RF	MAPE, MAE, RRMSE, MSE

Focusing on the feature engineering techniques used in the reviewed papers, nearly all applied normalization, either Min-Max scaling or StandardScaler, to ensure variables were transformed into a common scale. This is crucial for improving model performance and interpretability.

Two papers employing Self-Organizing Maps (SOM) widely recommended the sliding windows technique as an essential feature engineering approach, due to its effectiveness in recognizing patterns within financial time series data (Akbarzadeh & Soleimani, 2023; Pei et al., 2023).

In addition, some papers highlighted candlestick charts as valuable features, emphasizing their capability of capturing market price movements (Guan et al., 2024; Pei et al., 2023). Following this, Li et al. (2024) proposed a double-layer clustering approach based on K-line sequence similarity matching, calculated using features such as candlestick bodies, shadows and their relative positions. The main goal was to overcome limitations of traditional candlestick pattern

analysis. The model consisted of a KM with two layers to cluster K-line patterns. The first layer identified initial valid patterns, while the second layer filtered out redundant or invalid patterns from those selected in the first stage. The output was a pattern library, which indicated the price data and the predictive capability information of the patterns. Ten stocks from the Shanghai Stock Exchange covering 2019-2023 were used, and out-of-sample testing was performed on stock data from Shanxi Fenjiu during the same period. Three days sliding windows were computed, and validation was performed using four randomly chosen bullish and four randomly selected bearish patterns. According to the results, the presented model effectively distinguishes between bullish and bearish markets, and it is highly reliable and practical in forecasting short-term price movements.

Regarding technical indicators, the most used in the reviewed studies are Moving Averages (MA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Williams %R (WILLR) and Commodity Channel Index (CCI). They are used at least in four out of the thirteen studies that use technical analysis. The remaining five papers applied other approaches such as candlestick charts (Guan et al., 2024; Li et al., 2024; Pei et al., 2023) or only used market data (Shi et al., 2021; Wang, 2022). It is notable the high usage of momentum indicators in the reviewed studies, such as RSI, MACD, WILLR, CCI, Stochastic Oscillators (STOCH), Momentum (MOM) and Rate of Change (ROC). The author Wu (2020) concluded that momentum indicators are effective for clustering stocks into meaningful groups for trading. Additionally, Sarainmaa (2024) stated that the 50-day Simple Moving Average (SMA_50) was the most significant feature, as it achieves higher performance and lower trading period, in comparison to 200-day Simple Moving Average (SMA_200). Volume-based indicators are also used, such as Accumulation/ Distribution Index (AD), On-Balance Volume (OBV), and Volume Change (V%).

It is relevant to acknowledge that from the three papers that used SOM (Akbarzadeh & Soleimani, 2023; Pei et al., 2023; Shi et al., 2021), only one applied technical analysis, being SMA the only indicator applied to smooth prices (Akbarzadeh & Soleimani, 2023). This is a gap in the literature this thesis seeks to address by developing a SOM algorithm to predict market trends, which relies on a thorough technical analysis, which include diverse technical indicators.

In terms of the data sources, the literature predominantly focuses on the stock market, aligning with the primary scope of this thesis. However, notable exceptions include the use of four forex market indices (Akbarzadeh & Soleimani, 2023), as well as Chinese commodity futures and foreign exchange data (Pei et al., 2023). Among stock market datasets, the most frequently employed are Asian stock indices and individual stocks, along with the S&P 500. The timeframe and periods covered by these studies vary considerably. Some studies utilize approximately ten years of data (Akbarzadeh & Soleimani, 2023; Guan et al., 2024; Guo, 2020; Sagaceta-Mejía et al., 2024; Sarainmaa, 2024; Wu et al., 2021; Xu et al., 2020); others use around five years of data (Hu et al., 2024; Li & Wu, 2022; Li et al., 2024; Sáenz et al., 2023; Shi

et al., 2021; Wu, 2020); and additional papers use around 1 year of data (Bardi & Takacs, 2023; Bing et al., 2022; Navarro et al., 2023; Pei et al., 2023). Wang (2022) did not explicitly state the specific timeframe used. Furthermore, all studies utilized quantitative data delivered primarily from price and volume, focusing on technical analysis. However, two papers expanded their scope to include additional analytical methods. Saénz et al. (2023) incorporated company financial reports, representing a shift towards fundamental analysis; and Li & Wu (2022) integrated financial news analysis, employing sentiment analysis.

Table 4 provides a summary of the data, the experiments' timeframe, and the technical indicators used in each study. Almost all indicators are represented by acronyms, their corresponding translations can be found in the List of Abbreviations and Acronyms for reference.

Table 4 - Summary of Data Sources, Timeframe and Technical Indicators

Reference (Authors, Year)	**Data Source**	**Timeframe**	**Technical Indicators**
Akbarzadeh & Soleimani, 2023	4 forex market indices: AUDJPY, AUDUSD, EURGBP, EURJPY	2001-2014	SMA40
Bardi & Takacs, 2023	Tesla stock	1/1/2022-30/8/2023	SMA, RSI, BB, VOL
Bing et al., 2022	Shanghai Stock Exchange A-shares	1/6-9/6, 2015	OBV, DMA, ROC, DMI, WILLR
Guan et al., 2024	China A-shares market stocks	19/9/2012-19/9/2022	-
Guo, 2020	4000 stocks from the HS300 index	2007-2017	ROC, EMA, MACD
Hu et al., 2024	385 securities of S&P 500	2/1/2018-12/12/2022	EMA, KAMA, MACD, DEA, CCI, MOM, RSI, KDJ, WILLR, OBV, TR, ATR
Li & Wu, 2022	Hong Kong Stock Exchange stock price data and financial news	2003-2008	SMA7, SMA14, SMA21, BBUP, BBMD, BBLW, PROC, ADX, CCI, RSI7, RSI14, RSI21, ARDW, ARUP, MACD, MACDSG, MACDHT, AD, NART, DCPERIOD, DCPH, INPH, QUADR, SW, LSW, TRENDMODE
Li et al., 2024	10 Shanghai Stock Exchange stocks	11/11/2019-20/12/2023	-
Navarro et al., 2023	Philippine Stock Exchange	2018 and 2020	MACD, MACD-ALMA
Pei et al., 2023	China's commodity futures 2020, forex market 2016	18/5/2020-26/11/2020,	-

		3/1/2016-1/7/2016	
Sagaceta-Mejía et al., 2024	Three ETFs: ECH, EWZ, IVV	12/12/2009-1/1/2020	EBSW, BOP, STCRSI, CTI, KDJ, WILLR, ZS, DEC, INC, TTM TREND, AOBV, PVR
Sarainmaa, 2024	S&P 500 index price data	2010-2019	SMA50, SMA200, RSI14, V%, month-based patterns
Sáenz et al., 2023	240 company stocks' prices, returns, financial reports from Russell 3000 index	2017-2021	SMA
Shi et al., 2021	Market data from three indices: S&P 500, N225, CSI300	1/6/2015-29/5/2020	-
Wang, 2022	S&P 500 stock price movement data	-	-
Wu et al., 2021	Price data from S&P 500, DJIA, TAIEX stock indices	2010-2019	ROC, MACD, RSI, STC, TRIX, ADMI, CCI, VI+, VI-, ADI, MFI, OBV
Wu, 2020	5175 US stocks monthly market data	1/8/2009-1/8/2016	STMA, LTMA, STKDJ, LTKDJ, MACD
Xu et al., 2020	Four Chinese stocks (SPDB, CITIC, ZTE, LeTV)	1/1/2008-20/1/2019	SMA, EMA, MOM, STCK, STCD, MACD, RSI, WILLR, ADO, CCI

To conclude, this thesis aims to apply Self-Organizing Maps for trading stocks at a daily level, leveraging an ensemble learning approach composed of multiple SOMs. By combining the strengths of multiple SOMs, the research seeks to enhance predictive accuracy and improve overall trading performance. Although some researchers concluded that ensemble learning enhances predictive power (Xu et al., 2020), there is currently limited research on the application of ensemble learning methods using SOMs, representing a gap in the existing literature.

In addition, among the reviewed papers exploring SOM (Akbarzadeh & Soleimani, 2023; Pei et al., 2023; Shi et al., 2021), only Akbarzadeh & Soleimani (2023) employed technical indicators, specially relying solely on the Simple Moving Average (SMA). This indicates a notable gap in the literature, demonstrating a lack of comprehensive technical analysis integrated into SOM-based models. Incorporating a broader range of indicators, such as momentum, trend, volume, or volatility variables, could potentially enhance the performance and robustness of SOM trading strategies.

Furthermore, it is critical to consider distinct market conditions when backtesting any proposed model. This includes simulating and evaluating performance during distinct market regimes, particularly bullish and bearish markets. Such an approach is relatively uncommon in existing research, which typically assesses the models under uniform conditions.

4. METHODOLOGY

This study follows a Cross Industry Standard Process for Data Mining (CRISP-DM) approach. It consists of six iterative stages: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Schröer et al, 2021). The process begins with an understanding of the project' goals in a business perspective, and definition of technical objectives. Following the data understanding stage, where data is retrieved and analysed, the process moves to data cleaning, normalization and transformation to construct the final dataset. This dataset serves as input for the modelling phase, where selected models are applied. Data preparation and modelling are closely interrelated, as model training can provide insights for data transformation, and vice-versa. Moreover, models are evaluated using appropriate metrics, leading to the conclusions. It ends with the deployment stage, outlining the necessary steps and assessments to implement the proposed model in a real-world scenario.

Based on the CRISP-DM methodology, a graphical representation of the methods used in this study, alongside an overview in terms of process and organization are provided in Figure 6.

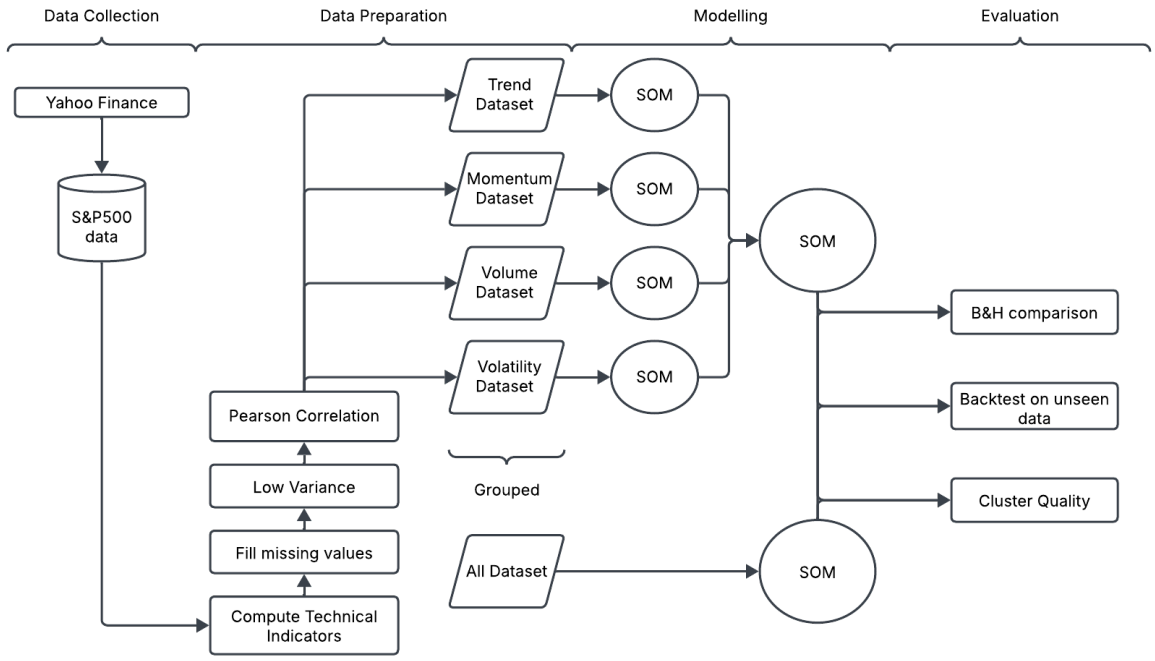

Figure 6 - Methodology Framework

In summary, Figure 6 illustrates the methodology framework used in this thesis, encompassing four main phases: data collection, data preparation, modelling, and evaluation. The process begins with the extraction of historical price and volume S&P 500 data through the yfinance python library. In the data preparation stage, a comprehensive set of technical indicators is computed. Given the inherent lookback periods of these indicators, missing values are addressed to ensure data completeness. Then, feature selection is performed using a two-

step process: first low variance features are removed, followed by the application of Pearson Correlation to eliminate highly correlated variables. The selected indicators are grouped into four categories: trend, momentum, volume, and volatility, each forming a distinct dataset.

In the modelling phase, a Self-Organizing Map (SOM) is trained on each of the four datasets independently. The resulting outputs of these SOMs are then used as inputs for a higher-level SOM, forming a hierarchical SOM architecture. In parallel, a benchmark SOM is trained on a unified dataset that includes all selected technical indicators from the four groups.

The evaluation phase is a comparison between both approaches: the HSOM and the benchmark SOM. Cluster quality metrics are analysed to assess the internal structure and topological coherence of the maps. Backtesting is conducted on unseen data to simulate the real-world performance of the generated trading signals. Finally, results are compared against the traditional buy-and-hold (B&H) strategy. A more detailed explanation of each step, including parameter choices, thresholds, and design decisions, is provided in the following subchapters.

4.1 DATA UNDERSTANDING

The focus of this thesis is stock trading, more precisely a case study on the S&P 500 index. This index, which represents the United States economy, one of the biggest economies in the world, was chosen based on the great availability of data, market coverage, and the high interest among investors. The data was managed using Jupyter Notebook, a web-based interactive computing platform, capable of handling complex data mining tasks and visualization tools. Therefore, the S&P 500 price and volume data was used. The 10-year dataset was retrieved using a library called yfinance (Yahoo Finance), ranging between January 1st, 2015, and January 1st, 2025. It consists of a Date column, which is unique; five price columns: Open, Close, Adjusted Close, High and Low; and a volume column. This daily-level dataset has 2516 trading days. Trading days for this index are regular business days (Monday to Friday), excluding certain holidays, such as Christmas, when the market remains closed.

In financial markets, closing prices are the foundation for decision-making in most trading and investment models, while the other prices provide valuable context. A decision was required about whether to use raw closing prices or adjusted closing prices in this analysis. Adjusted prices account for corporate actions, which can change the classification of market regimes. Upon reflection, raw prices were chosen, as a technical analysis is going to be performed. The technical indicators are calculated based on price data and adjusted prices could introduce distortions in these calculations. Investors trade on actual market prices at the time, not on future-adjusted values.

The closing prices trend during the period in analysis (2015-2025) can be observed in Figure 7. The S&P 500 shows a strong upward trend over the 10-year analysed period, which indicates long-term market growth. It starts at around 2000 USD in 2015 and ends around 6000 USD in 2025. A considerable decline can be seen in the early 2020, which represents COVID-19

pandemic shock, however it has recovered within the same year. Despite some declines, the overall trend remains bullish, representing strong economic growth.

Figure 7 - S&P 500 Closing Price Trend

The volume trend of the S&P 500 index from 2015-2025, illustrated in Figure 8, does not present a notable trend like the closing prices, on the contrary, it is relatively stable over time. However, it reveals a spike in the early 2020, aligning with the COVID-19 market shock, which is consistent with increased investor activity in times of uncertainty. In conclusion, it can be observed market increase during periods of high market volatility.

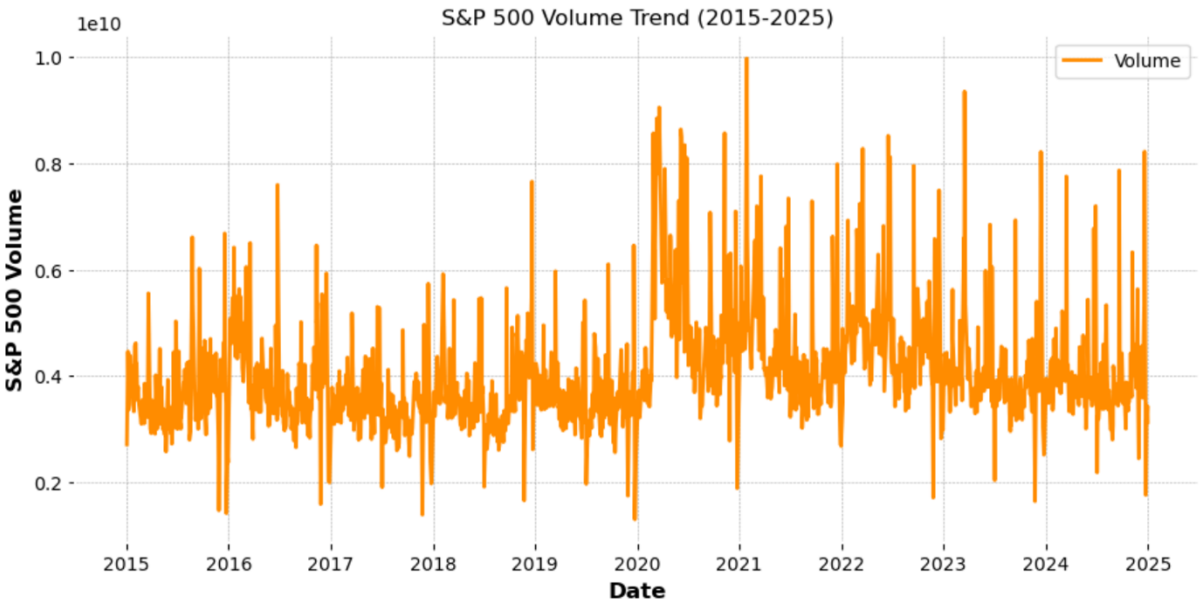

Figure 8 - S&P 500 Volume Trend

Exploring the relationship between volume and price volatility offers valuable insights into market dynamics. Figure 9 shows how volume interacts with the percentage change in closing prices. Most small price changes occur at lower volume levels, while higher volume levels are associated with greater variability in closing prices, both upward and downward. This analysis supports the use of volume as an input feature for SOM, as it may help identify market regimes. The market tends to be more volatile on days with higher trading volume.

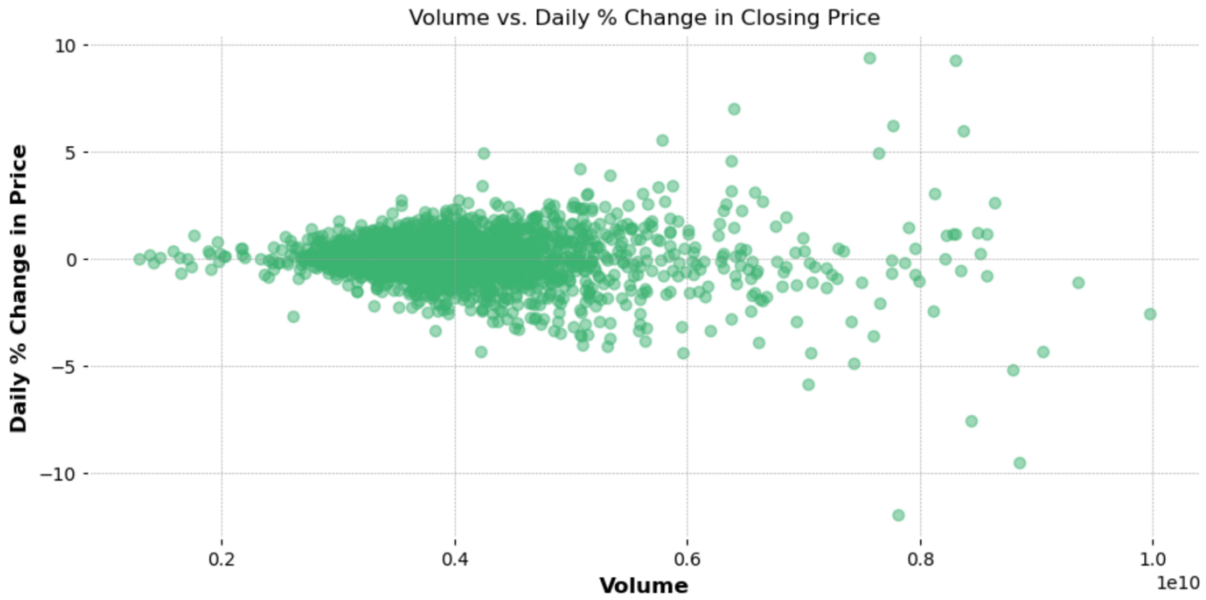

Figure 9 - S&P 500 Volume and % Change in Closing Prices

The data structure and values were thoroughly explored, leading to several conclusions. The dataset consists of 2516 rows and 7 columns, with no duplicate records, null, or missing values. Each row in the dataset represents one trading day, with information regarding the volume and the open, close, highest and lowest price of the index that day.

Statistical information of the data is presented in Table 5, except for the date column which is the unique identifier of the data. Based on Table 5, it can be observed that the Adj Close and the Close columns are identical, as they have the same values across all statistical metrics. This confirms that adjusted closing prices are unnecessary, since S&P 500 is a weighted average of stock prices, so stock splits and dividends do not happen directly. As a result, the Adj Close column was deemed redundant and removed from the dataset. Moreover, the count variable for all columns matches the number of rows in the dataset (2516), indicating no missing values as mentioned, which ensures data completeness and reliability for further analysis. The standard deviation (std) metric provides information into price volatility, with the highest and lowest values presented in the High and Low price columns. Lastly, volume values are much larger compared to the values in the price columns, which supports the need for scaling and normalization before applying the model.

Table 5 - Statistical Information of the data

	count	mean	min	25%	50%	75%	max	std
Adj Close	2516.0	3356.12	1829.08	2432.29	3004.99	4204.6	6090.27	1083.9
Close	2516.0	3356.12	1829.08	2432.29	3004.99	4204.6	6090.27	1083.9
High	2516.0	3373.45	1847.0	2441.47	3016.34	4233.62	6099.97	1089.15
Low	2516.0	3336.47	1810.1	2420.51	2990.94	4185.06	6079.98	1078.18
Open	2516.0	3355.66	1833.4	2431.93	3004.68	4206.07	6089.03	1083.71
Volume	2516	4007033048	1296530000	3428152500	3819815000	4341680000	9976520000	953984062

The resulting dataset incorporates six columns, and data types were changed accordingly to the appropriate ones, presented in Table 6.

Table 6 - Data Structure and Attributes

Column	Unit of Measurement	Definition	Data Type
Date	YYYY-MM-DD	Trading day identifier	datetime
Close	USD ($)	Last price at market close	float
High	USD ($)	Highest price of the day	float
Low	USD ($)	Lowest price of the day	float
Open	USD ($)	First price of the day	float
Volume	Units	Number of shares traded	integer

4.2 Data Preparation

This subchapter consists of all the transformations required to compute the final datasets used for training the Self-Organizing Maps (SOMs). Beginning with the computation of technical indicators, the data was structured with the date column as the index. Indicators were then treated, categorized, and carefully selected to form four distinct datasets, each targeting a specific dimension of market behaviour. Also, an additional dataset grouping all these four datasets was constructed.

This study explores the use of Self-Organizing Maps (SOMs) to develop trading strategies based on technical analysis. The selection of technical indicators was guided by both the literature review, and the availability of indicators within the TA-lib (Technical Analysis Library), a widely used Python library for financial market analytics. Previous research has demonstrated that certain indicators consistently contribute to predictive modelling and market regime detection (Akbarzadeh & Soleimani, 2023; Bardi & Takacs, 2023; Bing et al., 2022; Guo, 2020; Hu et al., 2024; Li & Wu, 2022; Navarro et al., 2023; Sagaceta-Mejía et al., 2024; Sarainmaa, 2024; Sáenz et al., 2023; Wu et al., 2021; Wu, 2020; Xu et al., 2020), serving as the foundation for the selection process.

A total of 38 technical indicators were computed, presented in Table 7 alphabetically. Table 7 displays each indicator's abbreviation, description, calculation period (when applicable), and its assigned category.

Table 7 - Technical Indicators Overview

Indicator	Description	Period (days)	Category
AD	Accumulation/ Distribution Index	-	Volume
ADOSC	Accumulation/ Distribution Oscillator	3, 10	Volume
ADX	Average Directional Movement Index	14	Trend
ATR	Average True Range	14	Volatility
BB_LW	Bollinger Band Lower Line	20	Volatility
BB_MD	Bollinger Band Middle Line	20	Volatility
BB_UP	Bollinger Band Upper Line	20	Volatility
CCI	Commodity Channel Index	14	Momentum
DX	Directional Movement Index	14	Momentum
EMA_20	Exponential Moving Average	20	Trend
EMA_200	Exponential Moving Average	200	Trend
EMA_5	Exponential Moving Average	5	Trend
EMA_60	Exponential Moving Average	60	Trend
KAMA	Kaufman Adaptive Moving Average	30	Trend
MACD	Moving Average Convergence Divergence	12, 26, 9	Momentum
MACD_HIST	Moving Average Convergence Divergence Histogram	12, 26, 9	Momentum
MACD_SIGNAL	Moving Average Convergence Divergence Signal Line	12, 26, 9	Momentum
MFI	Money Flow Index	14	Momentum
MOM	Momentum	10	Momentum
OBV	On-Balance Volume	-	Volume
ROC	Rate of Change	10	Momentum
ROCP	Rate of Change Percentage	10	Momentum
RSI	Relative Strength Index	14	Momentum
SMA_20	Simple Moving Average	20	Trend
SMA_200	Simple Moving Average	200	Trend
SMA_5	Simple Moving Average	5	Trend
SMA_60	Simple Moving Average	60	Trend
STOCH_D	Stochastic Oscillator Smoothed	-	Momentum
STOCH_K	Stochastic Oscillator	-	Momentum
STOCHRSI_D	Stochastic RSI Smoothed	-	Momentum
STOCHRSI_K	Stochastic RSI	-	Momentum
TEMA_20	Triple Exponential Moving Average	20	Trend
TEMA_200	Triple Exponential Moving Average	200	Trend
TEMA_5	Triple Exponential Moving Average	5	Trend
TEMA_60	Triple Exponential Moving Average	60	Trend
TR	True Range	-	Volatility
TRIX	Triple Exponential Average	30	Momentum

WILLR	Williams %R	14	Momentum

The lookback periods were selected based on conventionally accepted standards in financial practice and academic literature. These periods aim to balance responsiveness to short-term market fluctuations and the smoothing necessary for signal stability.

To structure the modelling process and enable modular experimentation, the indicators were categorized into four major groups:

- Trend Indicators (14): Capture directional movements and long-term tendencies.
- Momentum Indicators (16): Measure the rate of price change or acceleration.
- Volume Indicators (3): Reflect buying and selling pressure.
- Volatility Indicators (5): Quantify the extent of price variation.

This categorization was based on the analyse provided by Mostafavi & Hooman (2025), which allows for the training of distinct SOMs specialized in learning structural patterns in each technical dimension. This modular strategy was inspired by approaches in regime detection and hybrid modelling (Li & Wu, 2022; Saénz et al., 2023; Xu et al., 2020), where separating features by nature can improve interpretability and model performance.

Some indicators required historical price windows (lookback periods) to compute their values, resulting in missing entries at the beginning of the datasets. To address this issue, historical data was appended prior to the official analysis period. This ensured stability in indicator computation and completeness in the datasets.

For example, the Triple Exponential Moving Average (TEMA_200) has a lookback period of 200 days, the longest among the selected indicators. Approximately three years of historical data was necessary to compute this indicator. To tackle this challenge, three years previous data was extracted. This process involved incorporating the essential historical data to ensure all technical indicators were fully computed for the 10-year analysis period. Therefore, S&P 500 data from 2012 was extracted and used to calculate the indicators. By applying this approach, missing values were eliminated, ensuring that the first valid row (2015/01/01) contained fully calculated indicators, based on actual historical data instead of estimations or approximations. This strategy secures integrity and reliability to the dataset.

In order to improve the performance and interpretability of machine learning models, feature selection plays a vital role. By eliminating redundant or uninformative variables, the model is able to learn more effectively, reducing computational complexity and the risk of overfitting (Htun et al., 2023; Theng & Bhoyar, 2024). In the context of this study, selecting the appropriate technical indicators is fundamental to capture meaningful market patterns while reducing noise. To this end, two feature selection techniques, Low Variance and Pearson Correlation, were applied. These methods are computationally inexpensive, reduce training

time considerably, and proved to be efficient in selecting relevant features (Bing et al., 2022; Sagaceta-Mejía et al., 2024).

The Low Variance method identifies and removes variables whose values exhibit minimal variation across time. Such features are typically uninformative, as they fail to capture dynamic market behaviour, and offer little discriminatory power in clustering. For instance, an indicator that remains constant or changes very little over a long period is unlikely to contribute meaningfully to the SOM's ability to identify patterns.

In this research, a threshold of 0.01 was set to discard features. As a result, the momentum indicators Triple Exponential Average (TRIX) and Rate of Change Percentage (ROCP) were discarded from the dataset, having a variance lower than 0.01. This step helped reduce dimensionality and improved the computational efficiency of the SOM training process, while preserving features with more pronounced fluctuations reflective of changing market conditions.

Following, the Pearson Correlation technique was applied, which removed a considerable number of technical indicators from the analysis. Four datasets were constructed, each one containing the price and volume initial features, alongside the corresponding technical indicators. There is one dataset for each category of indicators: trend, momentum, volume, and volatility.

Two thresholds were defined due to the nature of the datasets. Momentum indicators behave differently, as they measure both the strength and speed of price movements. Consequently, they tend to show lower correlations, ranging between 0.8 and 0.9, compared to other indicators, which exhibit extreme correlations exceeding 0.9. Taking this into consideration, a threshold of 0.85 was applied to the momentum dataset, to retain the highly correlated pairs of features, while a threshold of 0.9 was employed for the other datasets. A list of highly correlated pairs for each dataset was retrieved and analysed.

Firstly, it was observed that the columns High, Low, and Open were highly correlated with the Close column across all datasets. These columns were primarily used to calculate technical indicators but hold limited value for clustering. For this reason, they were removed from the datasets, retaining only the Close and Volume from the original columns.

Moving Averages, including Simple Moving Average (SMA), Exponential Moving Average (EMA), and Triple Exponential Moving Average (TEMA) are highly correlated between each other, since they all serve as smoothing techniques. To enhance model efficiency only one type remained. The EMA was selected due to its balanced responsiveness, as it reacts faster to price changes than SMA and, at the same time, it is less prone to false signals than TEMA, which overreacts to noise. Figure 10 illustrates the closing price movement along with distinct periods (5, 20, 60, 200) Exponential Moving Averages (EMAs) over time. It confirms that EMAs provide a well-distributed representation of price movement and that it effectively captures

price trends from short-term to long-term perspectives. The exclusion of SMAs and TEMAs improve model performance, without compromising the analysis.

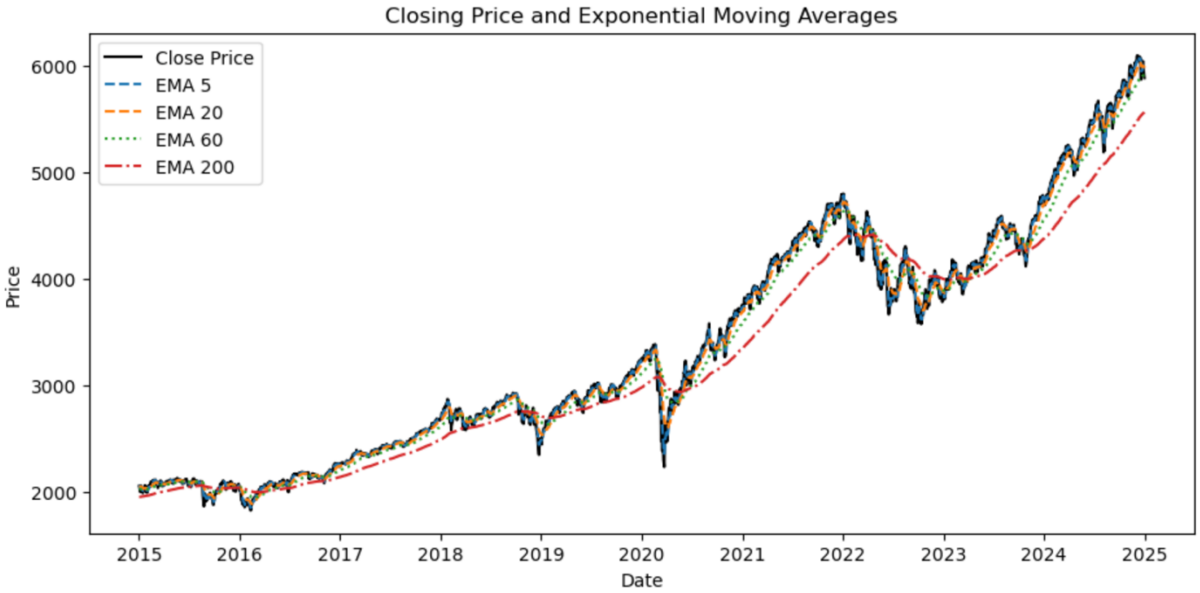

Figure 10 - Close price and EMAs over the period in analysis

After this initial removal of features, a closer look at the remaining variables of each dataset led to the final reduction. Regarding the trend dataset, Table 8 provides the highly correlated pairs of trend features (above the threshold = 0.9) and their level of correlation. Figure 19, in the Appendix A section, presents the correlation heatmap between the remaining features.

Table 8 - Highly Correlated Pairs of Trend Features

Trend x	Trend y	Correlation
Close	EMA_5	0.9996
KAMA	EMA_20	0.9992
KAMA	EMA_60	0.999
EMA_5	EMA_20	0.999
EMA_20	EMA_60	0.9982
Close	EMA_20	0.9978
KAMA	EMA_5	0.997
Close	KAMA	0.9957
EMA_5	EMA_60	0.9955
Close	EMA_60	0.9942
EMA_60	EMA_200	0.994
KAMA	EMA_200	0.99
EMA_20	EMA_200	0.9879
EMA_5	EMA_200	0.9843
Close	EMA_200	0.9831

Through the analyse of Table 8 and Figure 19, Kaufman Adaptive Moving Average (KAMA), Exponential Moving Average with lookback period of 5 days (EMA_5), and Exponential Moving Average with lookback period of 60 days (EMA_60) were deemed unnecessary. Regarding EMAs, EMA_5 is too close to the Close price, as it is the highest correlation pair (0.999572), which is why it was removed. Among the remaining indicators, EMA_20 and EMA_200 are the less correlated (0.987948), making them a suitable pair of moving averages (MA) to capture distinct market trends, with EMA_20 representing short-term and EMA_200 representing long-term trends. This choice aligns with Sarainmaa's (2024) suggestion for incorporating moving averages of different periods to capture their crossing effects. Lastly, KAMA was removed due to its strong correlation with EMA_20, the second highest correlation pair (0.999187).

As for the momentum dataset, Table 9 shows the pairs of features with correlation higher than 0.85, and Figure 20, in the Appendix A, displays the correlation heatmap. Based on these results, Moving Average Convergence Divergence Signal Line (MACD_SIGNAL), Momentum (MOM), Stochastic Oscillator Smoothed (STOCH_D), Stochastic Relative Strength Index Smoothed (STOCHRSI_D) were considered redundant. MOM and Rate of Change (ROC) are the highest correlation pair of features (0.963747), and ROC was the indicator kept due to its

simpler interpretation. After that, Moving Average Convergence Divergence (MACD) and MACD_SIGNAL have a strong correlation, and since MACD_SIGNAL is just a smoothed version of MACD, it does not add new information to the model, hence it was removed. Regarding STOCH_D and STOCHRSI_D, they are both also smoothed versions of STOCH_K and STOCHRSI_K respectively, which makes them discardable due to the high correlation between each other. Ultimately, although Relative Strength Index (RSI) and Williams %R (WILLR) appear in Table 9 as a highly correlated pair, both indicators were kept in the dataset due to their increased presence in the reviewed studies (Bardi & Takacs, 2023; Bing et al., 2022; Hu et al., 2024; Li & Wu, 2022; Sagaceta-Mejía et al., 2024; Sarainmaa, 2024; Wu et al., 2021; Xu et al., 2020).

Table 9 - Highly Correlated Pairs of Momentum Features

Momentum x	Momentum y	Correlation
MOM	ROC	0.9637
MACD	MACD_SIGNAL	0.9457
CCI	WILLR	0.9343
STOCH_K	STOCH_D	0.8912
STOCH_K	STOCHRSI_D	0.8879
MACD_HIST	MOM	0.8634
RSI	WILLR	0.8527

Concerning the volume dataset, as it can be observed in Table 10, Accumulation/ Distribution Index (AD) and On-Balance Volume (OBV) are the highest correlation pair of features. OBV was removed, because it does not account for price position as AD, while both measure cumulative volume flow. A correlation heatmap of the volume dataset can be seen in Figure 21, present in the Appendix A.

Table 10 - Highly Correlated Pairs of Volume Features

Volume x	Volume y	Correlation
AD	OBV	0.9584
Close	AD	0.9538
Close	OBV	0.9355

Regarding the volatility dataset, Table 11 presents the pairs of features with correlation higher than 0.9, and Figure 22, in the Appendix A, provides a visual representation of the correlations through a heatmap. The Bollinger Bands variables are highly correlated between each other

and with the Close price. The correlation levels are extreme, above 0.99, hence Bollinger Bands Upper Line (BB_UP), Bollinger Bands Middle Line (BB_MD), and Bollinger Bands Lower Line (BB_LW) were removed from the analysis.

Table 11 - Highly Correlated Pairs of Volatility Features

Volatility x	Volatility y	Correlation
BB_UP	BB_MD	0.9982
BB_MD	BB_LW	0.998
Close	BB_MD	0.9969
Close	BB_LW	0.9963
Close	BB_UP	0.9939
BB_UP	BB_LW	0.9925

Concerning the proposed model, after carefully selecting the technical indicators, four final datasets: trend, momentum, volume, and volatility serve as inputs to separate SOMs. Moreover, a combined dataset containing all selected indicators across categories is used as input for an additional SOM, which will be used for comparison with the proposed model. Four tables present in the Appendix B (Table 16, Table 17, Table 18, Table 19) display the final features selected for each dataset category, and statistics are included to provide an overview of each feature's distribution before normalization.

4.3 Modelling

Following the feature engineering phase, this study applies the Self-Organizing Map (SOM), a machine learning algorithm, known for its effectiveness in extracting patterns from financial data (Nair et al., 2017). In section 2.2 of this report, you can find a thorough explanation of how the algorithm works. The core contribution of this thesis is the development of a Hierarchical SOM (HSOM) which, when integrated with a predefined trading strategy, aims to generate buy, sell, and hold signals. The objective is to design a trading strategy based on this algorithm, capable of outperforming the traditional buy-and-hold approach. Additionally, a benchmark SOM is implemented to access whether the hierarchical structure provides a meaningful advantage in capturing market behaviour.

The Hierarchical SOM architecture draws inspiration from the Growing Hierarchical Self-Organizing Map (GHSOM) introduced by Dittenbach et al. (2000), and further developed by Rauber et al. (2002). This extension of the traditional SOM model addresses two important limitations. The need to predefine network size, and the inability to capture and represent hierarchical structures in the data. By stacking multiple independent SOMs across layers, the hierarchical SOM enables refined, interpretable, and multi-resolution clustering of complex data, especially beneficial in high-dimensional domains such as finance.

In this study, the HSOM framework is composed of two levels. The bottom-level SOMs, where one SOM is trained for each technical indicator group (Tend, Momentum, Volume, and Volatility), and a top-level SOM, which aggregates the outputs from the bottom-level SOMs into a single map. This structure design allows the model to capture specialized features in each group, while combining them into an integrated market structure view.

The modelling process was developed in Python using the MiniSom library. All datasets were sorted by the date column to preserve temporal consistency. The 10-year data was divided into 80% training (2015-2022) and 20% testing (2023-2024), reserving historical unseen data for backtesting and performance evaluation. Consequently, there are 2012 training instances and 504 testing instances. Log returns were calculated based on the Close prices, offering scale-invariant price change measurement, and both were stored separately for backtesting. Training returns were labelled 0 if negative and 1 if positive, for signal generation. Min-Max normalization was applied after splitting the data to prevent data leakage. It was applied separately to the train and the test data.

All Self-Organizing Maps (SOMs), both at the bottom and top levels of the hierarchy, were configured using a consistent set of parameters to ensure methodological coherence and reproducibility. Each SOM was trained using a 10x10 grid, with a learning rate of 0.9, Principal Component Analysis (PCA) based initialization, and a fixed random seed of 42 to maintain deterministic behaviour across runs. Random weight initialization was tested, however PCA based initialization gave the best results in this context. The number of features in each dataset defined the input dimensionality for the respective SOM. Each model was trained using an epoch-based learning scheme, where the algorithm iteratively processes the entire training dataset a fixed number of times. Additionally, random order was set to true during training, introducing random shuffling of the input data in each epoch. This randomized sequence helps prevent potential learning biases related to temporal ordering of data, while enhancing the model's generalization capability.

To determine the most appropriate training parameters, regarding the sigma, and the number of training epochs, referred to as iterations, a grid search was conducted. The search was performed independently for each bottom-level SOM (Trend, Momentum, Volume, Volatility) and the top-level SOM. The optimal configuration of parameters was selected based on the lowest combined error score, considering the quantization error, topographic error, and distortion measure equally weighted. The resulted number of epochs varied on 5 and 7 depending on the dataset. Precisely, all SOMs had best results with 5 iterations, instead of the trend bottom-level SOM, which performed best with 7 iterations. In relation to the neighbourhood radius, introduced as sigma, 1.0 was the best performer value for all SOMs.

The bottom-level SOMs were trained on the parameters described previously. Once trained, each SOM produced a two-dimensional map where each observation was assigned to its Best Matching Unit (BMU), the neuron whose weight vector most closely matched the input vector. These BMU coordinates served as a representation of the input data and were used as the

input for the top-level SOM, effectively capturing the structural patterns discovered within each group of technical indicators. To assess cluster quality, three commonly used unsupervised evaluation metrics were computed:

- Quantization error: evaluates the average distance between input vectors and their corresponding BMUs.
- Topographic error: measures the preservation of neighbourhood relations.
- Distortion measure: combines aspects of both fit and map structure.

The final stage of the hierarchical modelling process involved training the top-level SOM using the BMU coordinates obtained from the bottom-level maps. This SOM was trained using the the configuration principles discussed earlier. By aggregating the encoded outputs of the different technical categories, this higher-level SOM aimed to uncover global relationships across indicator groups and consolidate them into an integrated market structure. As with the previous models, cluster quality was evaluated using the same three performance metrics. The resulting top-level map is the basis for the trading strategy by enabling consistent classification of market conditions into actionable signals.

To visually assess the quality and structure of the clusters formed by the SOMs, Unified Distance Matrix (U-Matrix) visualizations were generated. The U-Matrix is a widely used tool in SOM analysis, as it highlights the distances between neighbouring neurons in the map (Lötsch & Ultsch, 2014). Figure 11 displays the U-Matrixes plots of the four bottom-level SOMs, each map reveals distinct dark regions suggesting that meaningful internal structures were captured in each group of indicators, supporting the modular modelling approach. Figure 12 shows the U-Matrix of the top-level SOM, that also displays dark regions separating groups of neurons. The presence of these dark zones in both levels confirms that the model identified distinct market conditions, which form the foundation for the trading strategy.

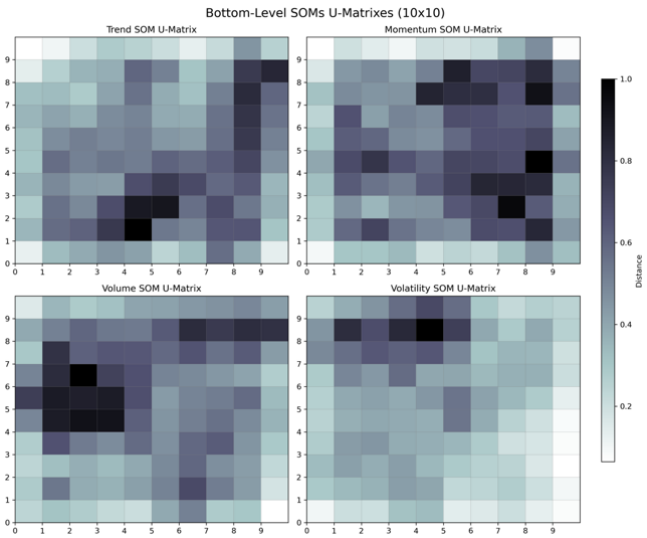

Figure 11 - Bottom-Level SOMs U-Matrixes

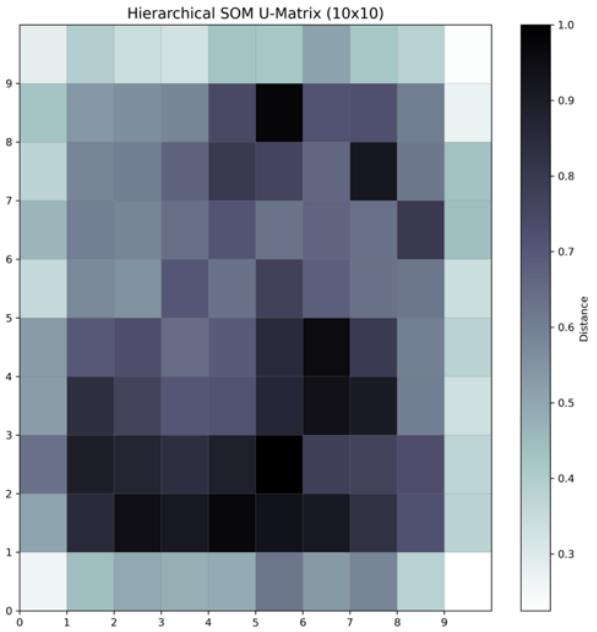

Figure 12 - Top-Level SOM U-Matrix

To evaluate whether the hierarchical structure contributes significantly to capturing and representing underlying patterns in market behaviour, a benchmark Self-Organizing Map (SOM) was developed. This additional model serves as a point of comparison to assess the added value of the hierarchical design in organizing technical information and generating useful trading signals.

The modelling process of this benchmark SOM followed the same methodology as the hierarchical approach, with one key distinction: instead of training separate SOMs for different indicator categories, a single SOM was trained on a comprehensive dataset. This dataset combined the Close price, trading volume, and all final technical indicators from the four previously defined groups (Trend, Momentum, Volume, and Volatility).

The dataset was split in 80% training and 20% testing, and Min-Max normalization was applied independently to each subset to prevent data leakage. Log returns were computed and stored for backtesting purposes, using the same approach described previously. Consistent with the hierarchical model, the training approach was identical, using PCA-based initialization, a learning rate of 0.9, and a 10x10 grid. A grid search was also used to identify the optimal configuration of sigma and training iterations, with the best performance achieved using sigma 1.0 and 5 training epochs, matching the configuration of the top-level SOM in the hierarchical framework. The same performance metrics, quantization error, topographic error, and distortion measure were used to evaluate cluster quality and determine the optimal parameter configuration. Figure 13 presents the resulting U-Matrix, which provides a visual overview of the cluster structure and separation achieved by the benchmark SOM.

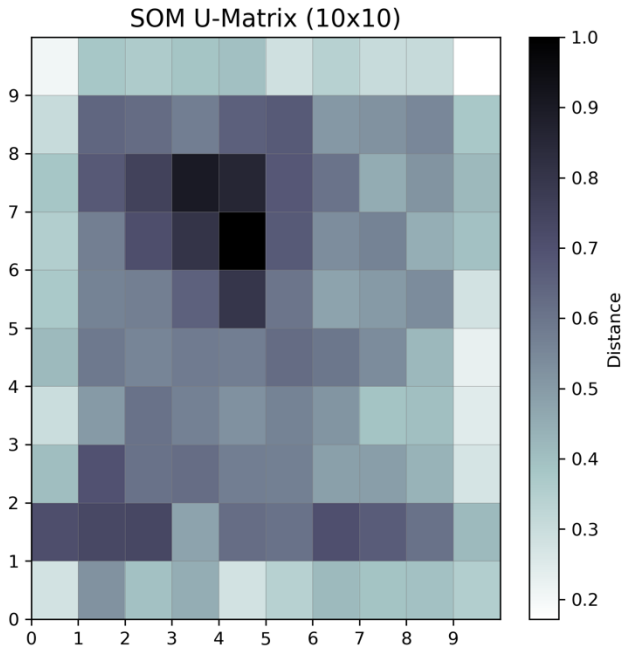

Figure 13 - Benchmark SOM U-Matrix

4.4 EVALUATION

With the final Hierarchical Self-Organizing Map (HSOM) model trained and validated, the next step was to translate its output into practical trading strategies, and benchmark them against the traditional buy-and-hold approach. This subchapter outlines the method used to generate market signals from the HSOM model and formulate trading rules.

The process started by extracting the winning neurons (BMUs) for each instance in the training dataset. Each instance in the training set, labelled as either a positive or negative return, 1 or 0 respectively, was mapped to its corresponding BMU. To determine the dominant signal for each neuron, a majority voting mechanism was applied. Each BMU was assigned the label (0 or 1) that appeared most frequently among its associated instances. This resulted in a dictionary, which maps each neuron of the hierarchical SOM grid, to a single market signal based on the prevailing return direction. To provide further insight into signal confidence, a probability distribution of label occurrences was computed for each BMU. Another dictionary stores the proportion of positive and negative returns mapped to each neuron.

Particularly, six neurons exhibited an equal distribution of positive and negative signals. Rather than arbitrary assigning a direction, a third label, Neutral (label 2), was introduced to reflect market uncertainty. Neurons with this label signal a “hold” action, avoiding trading activity during ambiguous conditions. The final dictionary contains BMU coordinates mapped to a label of 0 (Sell), 1 (Buy), or 2 (Hold).

To visually interpret these results, Figure 14 displays a probability U-Matrix visualization constructed to distinguish neurons by signal and help in identify high-confidence regions in

the map, by showing the positive/ negative signal distribution. In this visualization, neurons associated with label 0 (negative returns) are depicted in dark blue, those corresponding to label 1 (positive returns) are shown in light blue, while neurons representing label 2 (neutral) are illustrated in light grey. In addition, Figure 15 shows a class assignment visualization, which provides the same information in another perspective. The colours used are the same as Figure 14, however each neuron on the map is represented by a pie chart that reflects the proportion of label 0 and 1. Label 2 neurons are represented by a grey pie chart, which are neurons that have the same proportion of label 0 and 1. These visualizations provide insights into the spatial distribution of class labels across the HSOM grid. In this context, a noticeable lack of spatial clustering can be observed. However, this does not undermine the model's purpose, as clustering is not the primary objective. Instead, the model is employed as a mechanism to map unseen instances to specific neurons, their Best Matching Units (BMUs). The outputs are subsequently integrated into the formulated trading strategy, where each signal dictates an investor's action and the capital allocation percentage. Additionally, one specific neuron, coordinate (0,3), is observed to have a 100% probability of label 0. This is explained by the fact that, during training, only a single instance was mapped to this neuron, and it was associated with a negative return.

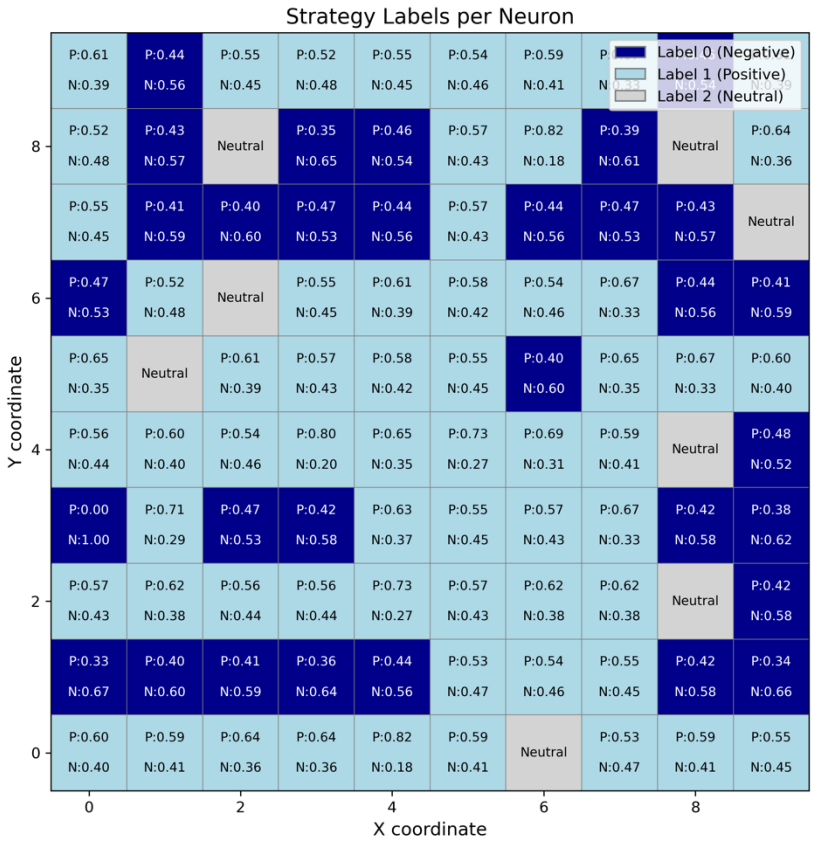

Figure 14 - HSOM Probability U-Matrix

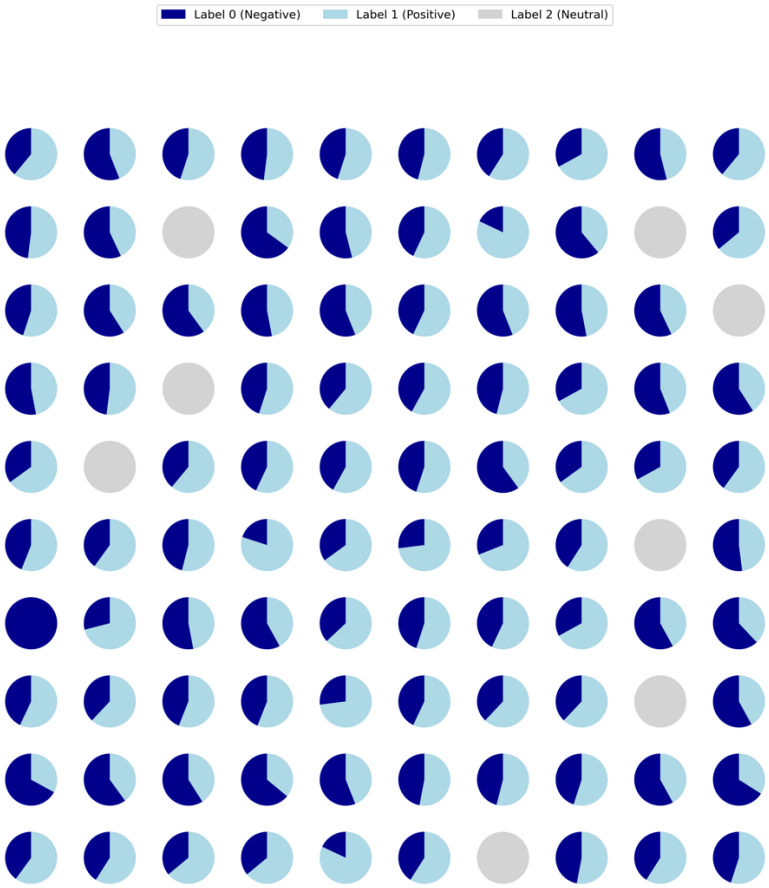

Figure 15 - HSOM Class Assignment Chart

The next critical step is the backtesting phase, which consists of simulating how the model would have performed on unseen historical data to evaluate its practical viability. In this study, backtesting was conducted using the test dataset, which comprises 504 instances corresponding to two years of S&P 500 data (2023-2024). This dataset was not used during the training phase and serves as an out-of-sample period to assess the model's generalization capability.

During backtesting, the trained HSOM was applied to each instance in the test dataset. For every trading day, the model generated the BMU coordinates and assigned a predicted trading signal (0, 1, or 2) based on the final labels attributed to the neurons during training. These signals form the core of the backtesting simulation, as they dictate the sequence of investment decisions that would have been made by following the model's outputs.

To thoroughly assess the effectiveness of these signals, three distinct trading strategies were formulated and backtested. The design of these strategies was intentional, aiming to reflect varying levels of market engagement and risk tolerance, and to provide a comparative perspective on how signal interpretation affects trading performance. Each strategy establishes specific rules on how to act upon the generated signals, defining whether to enter, leave, or hold a market position on a given day. The strategies operate on a daily time window,

meaning that the trading action determined by the signal is executed on the following trading day.

The first approach, the Easy Strategy offers a simple, deterministic approach based on the label assigned to each neuron. It is straightforward, as it does not incorporate model confidence. The second, the Consistent Strategy, incorporates probability thresholds to reflect the model's confidence in each signal, offering a more systematic and data-driven approach. Finally, the Cautious Strategy takes a more conservative and risk averse approach. It only acts when higher-confidence probabilities are present, thereby reducing exposure during periods of uncertainty and aiming to minimize risks.

By backtesting these three strategies, it was possible to evaluate their respective performances across multiple dimensions, including profitability, risk-adjusted returns, and drawdowns. These metrics provided valuable insights into each strategy's effectiveness, stability, and risk profile. In addition, a benchmark comparison was conducted against the traditional Buy-and-Hold strategy, which served as a baseline to assess whether the HSOM-driven approaches could outperform a passive investment strategy. The comparative analysis not only focused on returns, but also on risk mitigation and capital exposure. A detailed overview of the trading rules underlying each strategy is presented in Table 12, which serves as a reference for the decision-making framework applied throughout the backtesting process.

Table 12 - Trading Strategies

Easy Strategy	**Consistent Strategy**	**Cautious Strategy**
BMU = label 0 -> leave the market with 10% of capital (sell)	When Pp <= 0.4 -> leave the market with 20% of capital (sell)	When Pp < 0.5 -> leave the market with 10% of capital (sell)
BMU = label 2 -> hold the market position, do nothing	When 0.4 < Pp < 0.6 -> hold the market position, do nothing	When 0.5 <= Pp < 0.6 -> hold the market position, do nothing
BMU = label 1 -> enter the market with 10% of capital (buy)	When Pp >= 0.6 -> enter the market with 20% of capital (buy)	When Pp >= 0.6 -> enter the market with 20% of capital (buy)

5. RESULTS AND DISCUSSION

This chapter discusses the empirical results obtained from applying Self-Organizing Map (SOM) models to financial market data. The first objective is to assess whether the hierarchical structure contributes meaningfully to cluster quality and overall trading performance, by comparing the Hierarchical SOM (HSOM) model with the benchmark SOM. The second objective is to evaluate the effectiveness of the trading strategies derived from the HSOM and compare them with the traditional buy-and-hold (B&H) benchmark. The results are interpreted through visualizations and quantitative performance metrics, with a particular focus on return value paths, exposure patterns, and metrics such as Sharpe Ratio, Maximum Drawdown, Information Ratio, and Accuracy. In addition, the third objective focuses on the interpretation of the features in the HSOM map, providing insights into the relationships within the data and the organization of the map's neurons.

5.1 Hierarchical SOM vs Benchmark SOM

An essential goal of this research was to assess whether a hierarchical modelling approach provides advantages over a single approach, in the Self-Organizing Map (SOM) model, when applied to the design of trading strategies. The following analysis relies on both cluster quality metrics and strategy performance outcomes.

From a clustering perspective, cluster quality was evaluated through the metrics: quantization error, topographic error, and distortion measure. Table 13 displays their values rounded to two decimal cases. The benchmark SOM produced a lower quantization error (0.34) and distortion measure (6844.93), indicating more precise representation of individual data points. However, the HSOM achieved a lower topographic error (0.25), which suggests a superior ability to preserve the topological relationships in the input space. This is a critical aspect in time-series modelling, where the continuity and structure of market patterns carry essential predictive value. The HSOM's layered architecture, which captures distinct types of technical information through specialized bottom-level SOMs, appears to translate into more coherent and robust cluster boundaries at the top level.

Table 13 - Cluster Quality: SOM vs HSOM

Metric	SOM	HSOM
Quantization Error	**0.34**	2.41
Topographic Error	0.34	**0.25**
Distortion Measure	**6844.93**	59923.28

Entering a trading performance perspective, the advantages of the hierarchical approach become even more pronounced. As shown in Table 14, the HSOM outperformed the benchmark SOM across all three trading strategies in both final capital and return on

investment (ROI). The initial capital value used across this research was 1000€. The Consistent Strategy, which integrates probability-based decision thresholds, achieved the best overall result with the HSOM: a final capital of 1473.54€ and a ROI of 47.4%, compared to a €1268.39 and 26.8% with the benchmark SOM. The Easy Strategy, which was the best performing strategy for the benchmark SOM, reached a 43% ROI with HSOM against a 30.9% with the single approach. Even the most conservative strategy, the Cautious Strategy, yielded a 40% ROI with the hierarchical approach, outperforming the benchmark SOM by almost 20 percentage points.

Table 14 - Capital Return: SOM vs HSOM

Strategies & Metrics		**SOM**	**HSOM**
Easy Strategy	*Final Capital (€)*	1308.75€	1430.48€
	ROI (%)	30.9%	**43%**
Consistent Strategy	*Final Capital (€)*	1268.39€	1473.54€
	ROI (%)	26.8%	**47.4%**
Cautious Strategy	*Final Capital (€)*	1203.53€	1399.74€
	ROI (%)	20.4%	**40%**

These results demonstrate that the hierarchical design not only supports better structural representation of the data, but also translates into increased trading outcomes. The modular separation of indicator categories, followed by their integration into a unified map, enables the model to detect patterns and make more informed trading decisions. While the benchmark SOM offers simplicity and computational efficiency, it lacks the architectural nuance required to fully capture the multi-dimensional nature of market dynamics.

5.2 Hierarchical Trading Strategies vs Buy-and-Hold Traditional Strategy

Focusing on the Hierarchical SOM approach, the backtesting results present in Figure 16 show that all three HSOM-based trading strategies outperformed the traditional B&H strategy, both in terms of return and risk-adjusted performance. It illustrates not only the quantitative superiority of the trading strategies derived from the HSOM model, but also the qualitative advantage of integrating signal confidence in trading decisions. Although the trajectories of the Easy, Consistent, and Cautious strategies follow a similar overall path, notable divergences arise during periods of market volatility. The Consistent Strategy (orange line) proves superior adaptability, consistently outperforming the others from the intermediate of the test period onward. Its ability to recover faster from drawdowns and to capture upward trends with greater precision leads to the highest final capital of 1473.54€, which represents a 38.28% increase over the B&H strategy during the same period. On the contrary, the Easy Strategy (blue line) exhibits a more stable and less aggressive path, failing to capitalize as effectively on

upward market movements. Lastly, the Cautious Strategy (green line) is visibly more conservative, reflected in its flatter trajectory in bullish periods. This illustrates that, although it is more protective in uncertain conditions, it also sacrifices potential gains in favourable environments.

The B&H trading strategy (dashed red line) relies on the initial action of buying an asset and sticking with it for long periods of time, believing that markets generate increased value over time. It is a benchmark strategy that has been followed over decades, and it has had positive results in the S&P 500 context. However, this thesis' results reinforce the practical value of active, model-driven strategies, which can dynamically respond to market conditions rather than remain fully exposed.

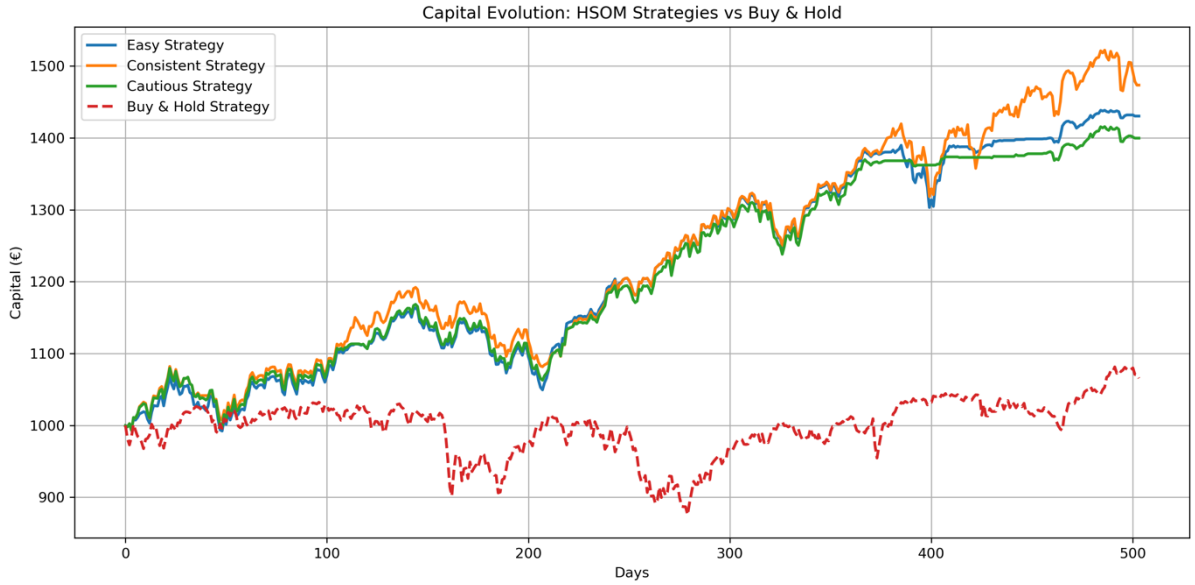

Figure 16 - Capital Evolution: HSOM Strategies vs Buy-and-Hold

Diving into more detail on the performance metrics, Table 15 summarizes their values for each Strategy. The Consistent Strategy clearly stands out as the best overall performer. As mentioned, it achieved the highest final capital (1473.54€) and accuracy (52.29%), naturally followed by the highest percentage capital difference in relation to the B&H approach (38.28%). It also reached the highest Information Ratio (0.883), which measures excess return over the benchmark adjusted for volatility, reinforcing its superior relative performance. These results suggest that incorporating signal probabilities into the decision-making process allows the model to capitalize on favourable conditions while avoiding exposure during periods of increased uncertainty. On the other hand, the B&H strategy, which remained fully exposed throughout the period, had the worst performance, producing the lowest final capital (1065.59€) and Sharpe Ratio (0.221), and the highest Maximum Drawdown (15.16%). These results highlight the vulnerability of static investment strategies, and confirm the advantage of using adaptive, model-driven trading strategies, relying solely on technical analysis.

Nevertheless, when comparing the Consistent Strategy with the other two HSOM-based trading strategies, its performance reflects balanced risk-return trade-off. While it achieved the highest return, it also recorded the lowest Sharpe Ratio among the three (1.648), indicating that it delivered slightly lower risk-adjusted returns compared to the others. The Cautious Strategy attained the highest Sharpe Ratio (1.696) and the lowest maximum drawdown (9.04%), reflecting its strength in capital preservation, but it lagged in total return due to its more selective market exposure, which makes it suitable to more risk-averse investors. The Easy Strategy, on the contrary, achieved a higher final capital than the Cautious Strategy, but with the highest Maximum Drawdown among the three (9.81%), and slightly lower risk-adjusted performance. These contrasts reveal how different interpretations of model signals can significantly impact trading behaviour and outcomes. The Consistent Strategy, by maintaining a middle ground, demonstrated strong returns while keeping risk within acceptable boundaries, making it the winning strategy in this evaluation.

Overall, this analyse underscores the value of probabilistic modelling within the HSOM framework. The Consistent Strategy was able to better differentiate between strong and uncertain signals by integrating probabilistic signal interpretation in the trading logic, leading to more effective capital allocation. The trade-off between assertiveness and cautious becomes especially apparent in the risk-adjusted metrics, revealing that superiors' returns can be achieved without compromising on risk, provided that the signals are interpreted and acted upon with the right level of confidence.

Table 15 - Performance Metrics per Strategy

Metric	Buy & Hold	Easy	Consistent	Cautious
Final Capital (€)	1065.59	1430.48	**1473.54**	1399.74
Relative Difference (%)	--	+34.24%	**+38.28%**	+31.36%
Sharpe Ratio	0.221	1.691	1.648	**1.696**
Max Drawdown (%)	15.16%	9.81%	9.26%	**9.04%**
Information Ratio	--	0.835	**0.883**	0.791
Accuracy (%)	100%	49.30%	**52.29%**	45.13%

Additionally, exposure levels were analysed to understand how each strategy allocated capital across time. Figure 17 illustrates the capital exposure levels for each trading day during the test period, offering insight into the behavioural patterns of the three HSOM-based strategies.

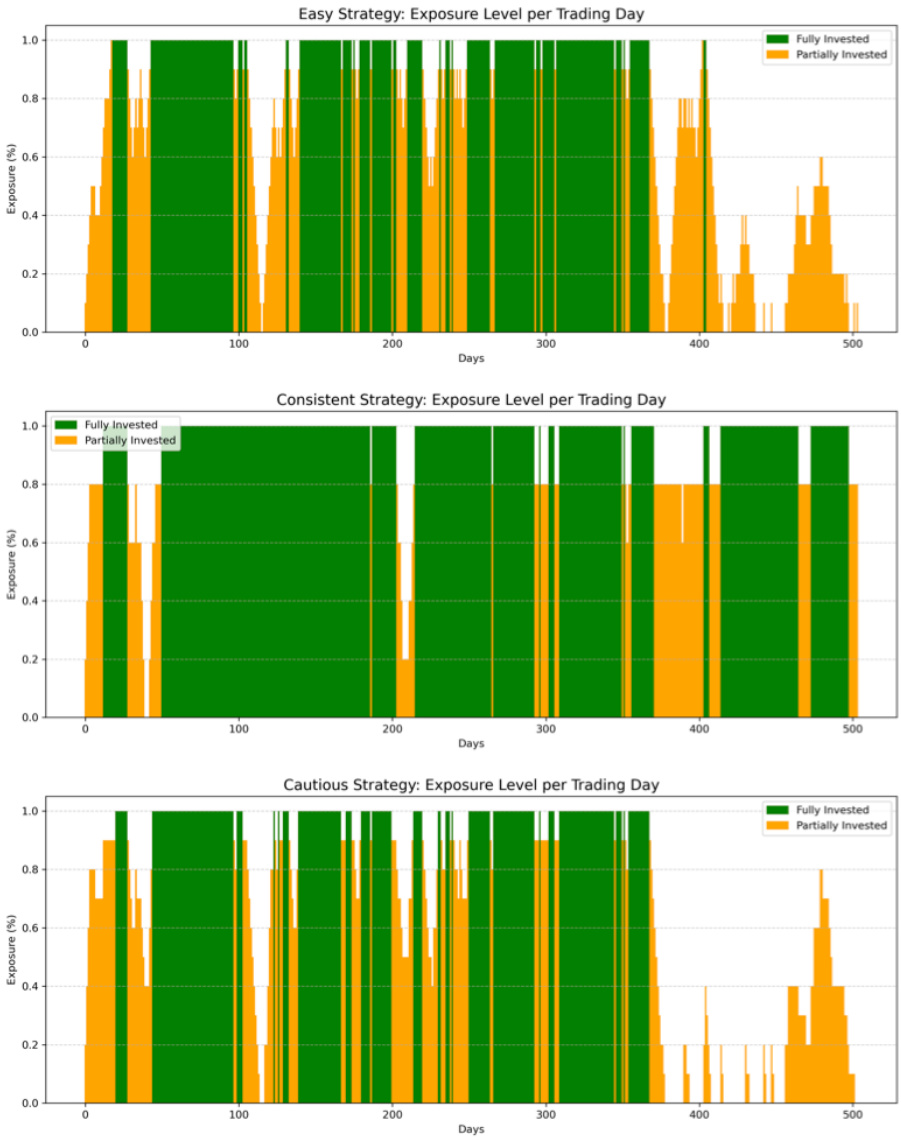

Figure 17 - HSOM-based Strategies Exposure Levels

The Consistent Strategy stands out for maintaining an overall higher exposure to the market, particularly during favourable periods, while still demonstrating the capacity to withdraw capital in less confident scenarios. Its exposure shifts are measured and strategic, reflecting the influence of signal probability thresholds and reinforcing the strategy's ability to balance assertiveness with caution. On the contrary, the Cautious Strategy exhibits a distinctly more conservative profile. Exposure levels are frequently below full investment, especially in the higher stages of the test period. This confirms its principle of only investing when the model's confidence is strongest, resulting in reduced risk but also missed opportunities for profit. The Easy Strategy shows greater volatility in its exposure pattern, especially on the second half of the test period. Capital is frequently withdrawn and re-entered successively, which leads to less stability, and potentially increasing transaction costs. This observed differences in exposure level align with the overall performance results.

5.3 Feature Analysis

To gain further insight into how the model organizes and interprets market dynamics, a component plane analysis was conducted on the trained Hierarchical SOM (HSOM). The resulting visualization, Figure 18, illustrates the distribution of each feature across the HSOM's neurons, offering a two-dimensional representation of the internal structure learned during training.

Each subplot in the component plane map corresponds to a single input feature, such as the Close Price, Volume, or technical indicators. Warmer colours (reds) indicate higher normalized values of that feature, while cooler colours (blues) indicate lower values. This allows us to observe how each variable is spatially organized and whether it aligns with specific regions of the map.

From the visual analysis, several important patterns emerge. Regarding volatility indicators, both the Average True Range (ATR) and True Range (TR) exhibit mainly blue areas across the map, indicating low normalized values. This suggests that the training data predominantly reflects periods of low market volatility, where price changes are smoother and more gradual. The Volume and Directional Movement Index (DX) component planes also show mostly blue regions, indicating low trading volume and weak trend strength on average. Moving on to trend indicators, the Exponential Moving Averages (EMA_20 and EMA_200), alongside the Close price and the volume indicator Accumulation/ Distribution (AD), display component planes with higher values concentrated on the upper region of the map. This overlap indicates that HSOM has grouped together neurons associated with upward price trends, higher moving averages, and accumulation phases. Neurons in this region likely reflect bullish market regimes, characterized by steadily rising prices and buying pressure.

On the other hand, regarding momentum indicator convergence, Commodity Channel Index (CCI), Relative Strength Indicator (RSI), Money Flow Index (MFI), Rate of Change (ROC), Moving Average Convergence Divergence (MACD), and Moving Average Convergence Divergence Histogram (MACD_HIST) share similar mid-range activation patterns, marked by moderate values spread across the centre of the map. This indicates a consistent but not exaggerated presence of momentum, representing moderate bullish phases, where price movements have already developed but are not yet exhausted. In addition, it is relevant to analyse the Stochastic Oscillator (STOCH_K), Stochastic Relative Strength Index (STOCHRSI_K), and the Williams %R (WILLR), which present component planes with high contrasting colours: dark reds in the upper right regions and dark blues predominantly located in the lower left regions. This behaviour is typical of oscillators that rapidly fluctuate between overbought and oversold conditions. It suggests that SOM has captured more reactive patterns, which could indicate volatile market segments. These regions might be particularly important for short-term strategies or cautionary signals.

Although these findings do not provide clear distinctions of bullish, bearish, or volatile markets, they reveal how the HSOM organizes the input space based on shared feature behaviour. The upper region of the map appears to cluster trend-following and accumulation behaviour, while other regions differentiate between moderate momentum and oscillator conditions, with overall low volatility. The interpretation of these component planes evidence that the HSOM is learning financially meaningful structure in the data. These neurons are organized in a way that reflects relationships among indicators and their association with market conditions. It validates the model as a strategy building tool that is not only predictive but interpretable.

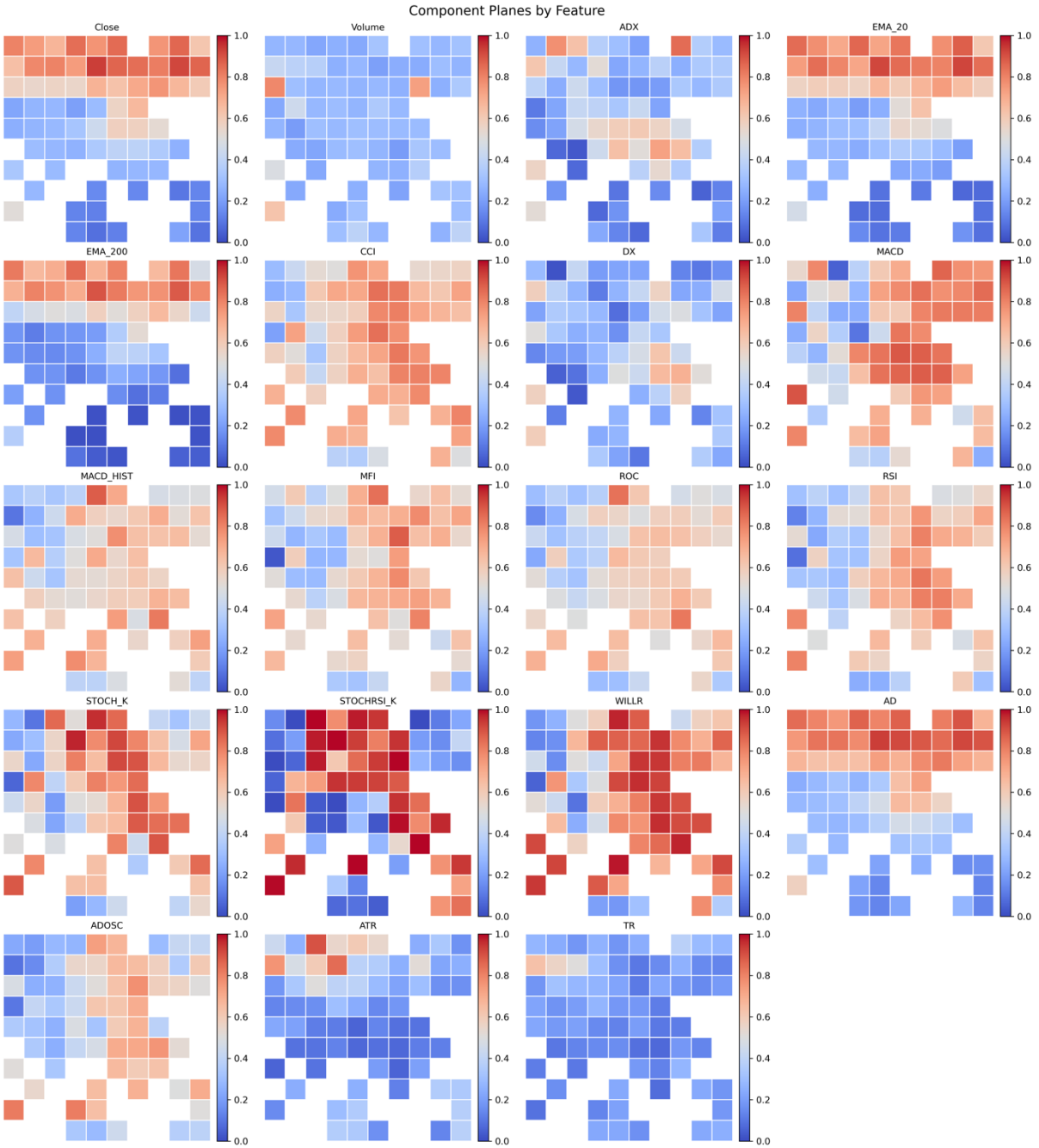

Figure 18 - HSOM Component Planes

6. CONCLUSIONS AND FUTURE WORK

This final chapter aims to provide a clear overview of the research findings and conclusions. It will also refer some limitations and topics that can be further explored. This thesis has the primary goal of exploring the use of Self-Organizing Maps (SOMs) in the development of adaptive trading strategies, that potentially outperform traditional strategies such as buy-and-hold. Through the integration of machine learning, strategy simulation, and financial strategy analysis, the research contributes to the growing field of machine learning and quantitative finance.

Answering the first sub-question this thesis seeks to address, a major contribution of this study is the comparative analysis between the benchmark Self-Organizing Map (SOM) model and the Hierarchical Self-Organizing Map (HSOM) approach. The HSOM model adopts a modular architecture, grouping technical indicators by category, and consistently delivered superior performance across all three trading strategies. While the Easy Strategy was the top performer under the benchmark SOM, the Consistent Strategy achieved the highest returns using the HSOM approach. Even when comparing the same strategy, the HSOM version of the Easy Strategy outperformed its benchmark counterpart by a margin of 12 percentage points in Return on Investment (ROI). Beyond return metrics, the HSOM also exhibited a lower topographic error, suggesting a more faithful preservation of the input data's structural relationships. These findings support the hypothesis that a layered architecture improves SOM learning and interpretability in financial applications.

Regarding the second sub-question, this thesis successfully demonstrated that Hierarchical Self-Organizing Maps (HSOMs) can effectively organize and cluster high-dimensional technical indicator data from the S&P 500. By grouping and selecting technical indicators into four distinct categories: trend, momentum, volume, and volatility, the model was able to group similar market behaviours in an unsupervised way. The visualization of component planes revealed some meaningful patterns in the map's neuron structure, which validates the model as a suitable tool for market regime detection.

Lastly, this research tackled the third sub-question by designing and testing three distinct trading strategies: Easy, Consistent, and Cautious, each interpreting HSOM output signals differently. The Consistent Strategy, incorporating probability thresholds in its decision rules, demonstrated the best overall performance. These strategies were compared to the buy-and-hold (B&H) traditional approach in terms of returns and risk-adjusted metrics. All strategies have outperformed the traditional benchmark, with the best-performing strategy achieving a capital difference of 38.28% in relation to the B&H strategy. This proves the ability of the HSOM serving as a core signal generator for practical trading strategies. Furthermore, the HSOM strategies demonstrated lower drawdowns and higher Sharpe Ratio and Information Ratios, showing that the approach is not only profitable but also more robust under changing market conditions.

While the findings of this thesis are promising, several limitations must be acknowledged, which also suggests directions for future research and development. First, the current model does not account for transaction costs or market liquidity constraints, which could impact real-world profitability. Incorporating these factors into the framework would enable a more realistic assessment of the strategies's performance. Furthermore, the trading strategies were backtested on a separate, two-year out-of-sample test period. Despite providing valuable insights, it may not fully capture the long-term complexity and unpredictability of financial markets. Future work should therefore consider validating the model in real-time or live trading environment to assess its practical viability. In addition, performing a comprehensive robustness test is strongly recommended for future studies to evaluate the model's stability across distinct market regimes.

Further enhancements could also involve the integration of alternative data sources, such as sentiment or macroeconomic indicators, which may improve both market regime classification and strategy performance. Another valuable proposal would be the addition of a bottom-level SOM focused on candlestick pattern recognition, potentially improving the model's responsiveness to short-term price movements.

BIBLIOGRAPHICAL REFERENCES

Akbarzadeh, F., & Soleimani, A. (2023). Forecasting financial time series trends by pattern recognition. *International Journal of Nonlinear Analysis and Applications*, *14*(1), 2587–2600. https://doi.org/10.22075/IJNAA.2022.27602.3660

Ayunku, P. E. (2020). The Efficient Market Hypothesis: A Review of Precise Literatures. *IOSR Journal of Economics and Finance*, *11*(1), 50–56. https://doi.org/10.9790/5933-1101055056

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. *LNCS*, *3516*, 476–483.

Bardi, A., & Takacs, M. (2023). Integrating Technical Analysis and Neural Networks for Optimizing Algorithmic Trading. *In Proceedings of the IEEE 23rd International Symposium on Computational Intelligence and Informatics*, 303–308. https://doi.org/10.1109/CINTI59972.2023.10382019

Bing, H., Zhou, Y., Yuan, Z., Cheng, K., Wang, Y., Liu, H., & Wang, L. (2022). A Study on Quantitative Investment Strategies Based on Cluster Analysis. *IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, *June*, 956–960. https://doi.org/10.1109/ITAIC54216.2022.9836859

Chiang, T. C., & Zheng, D. (2010). An empirical analysis of herd behavior in global stock markets. *Journal of Banking & Finance*, *34*(8), 1911–1921. https://doi.org/10.1016/J.JBANKFIN.2009.12.014

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *Journal of Finance and Data Science*, *2*(1), 42–57. https://doi.org/10.1016/j.jfds.2016.03.002

Deboeck, G. & Kohonen, T. (2000). *Visual Explorations in Finance*. 2nd edition, Springer Science & Business Media, New York.

Dittenbach, M., Merkl, D., & Rauber, A. (2000). *The growing hierarchical self-organizing map*. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2000)*, *6*, 15–19. https://doi.org/10.1109/IJCNN.2000.859416

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, *38*(1), 34–105.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383-417. https://doi.org/10.2307/2325486

Fama, E. F., Booth, D., Bradley, M., Brennan, M., Buser, S., Campbell, J., Chen, N.-F., Cochrane, J., Constantinides, G., Ferson, W., French, K., Harvey, C., Ippolito, R., Jensen, M., Kaul, G.,

Lakonishok, J., Mcdonald, B., Merton, R., Mitchell, M., … Warner, J. (1991). Efficient Capital Markets: II. *The Journal of Finance*, *46*(5), 1575–1617. https://doi.org/10.1111/J.1540-6261.1991.TB04636.X

Gervais, S. and Odean, T. (2001). Learning to be overconfident. *Review of Financial Studies, 14*(1), 1–27.

Grossman, S. J. & Stiglitz, E., J. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review, 70*(3), 393-408. https://papers.ssrn.com/abstract=228054

Guan, B., Zhao, C., Yuan, X., Long, J., & Li, X. (2024). Price prediction in China stock market: an integrated method based on time series clustering and image feature extraction. The *Journal of Supercomputing*, *80*(7), 8553–8591. https://doi.org/10.1007/S11227-023-05562-Z/TABLES/7

Guo, Y. (2020). Stock Trading Based on Principal Component Analysis and Clustering Analysis. *IOP Conference Series: Materials Science and Engineering*, *740*(1). https://doi.org/10.1088/1757-899X/740/1/012129

Han, X., & Wang, L. (2008). Stock company comprehensive assessment model based on kohonen network. *Proceedings - 2nd International Conference on Genetic and Evolutionary Computing, WGEC 2008*, 185–188. https://doi.org/10.1109/WGEC.2008.97

Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, *9*(1), 1–25. https://doi.org/10.1186/S40854-022-00441-7/FIGURES/3

Hu, W., Zhou, J., Hu, W., & Zhou, J. (2024). Building Technical Analysis Strategies Using Multivariate Longitudinal and Time-to-Event Data in Stock Markets. *Computational Economics*, 1–32. https://doi.org/10.1007/S10614-024-10782-3

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern*, *43*, 59–69.

Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of the IEEE*, 78(9), 1464-1480.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Kossakowski, P., & Bilski, P. (2017). Analysis of the Self-Organizing Map-based Investment Strategy. *International Journal of Computing*, *16*(1), 10–17. https://doi.org/10.47839/ijc.16.1.866

Li, X., & Wu, P. (2022). Stock Price Prediction Incorporating Market Style Clustering. *Cognitive Computation*, *14*(1), 149–166. https://doi.org/10.1007/S12559-021-09820-1

Li, X., Liu, Q., Hu, Y., & Liu, H. (2024). The Double-Layer Clustering Based on K-Line Pattern Recognition Based on Similarity Matching. *Information, 15*(12), 821. https://doi.org/10.3390/INFO15120821

Li, Z., Li, R., & Xiao, B. (2021). A Literature Review on the Evidence and Limitation for the AMH Theory. *Proceedings of the 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)*, *203*, 2886–2891. https://doi.org/10.2991/ASSEHR.K.211209.468

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129-137.

Lo, Andrew W. (2007). Efficient Markets Hypothesis. In: THE NEW PALGRAVE: A DICTIONARY OF ECONOMICS, L. Blume, S. Durlauf, eds., 2nd Edition, Palgrave Macmillan Ltd., Available at SSRN: https://ssrn.com/abstract=991509

Lötsch, J., Ultsch, A. (2014). Exploiting the Structures of the U-Matrix. In: Villmann, T., Schleif, FM., Kaden, M., Lange, M. (eds) Advances in Self-Organizing Maps and Learning Vector Quantization. Advances in Intelligent Systems and Computing, *295*, 249-257 https://doi.org/10.1007/978-3-319-07695-9_24

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, *17*(1), 59–82. https://doi.org/10.1257/089533003321164958

Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, *174*(3), 1742–1759. https://doi.org/10.1016/J.EJOR.2005.03.039

Mohamed, A. (2019). *Artificial Intelligence in investing: Stock clustering with Self-organizing map and return prediction with model comparison*. Master thesis. School of Business and Management, Lappeenranta University of Technology.

Mostafavi, S. M., & Hooman, A. R. (2025). Key technical indicators for stock market prediction. *Machine Learning with Applications*, *20*, 100631. https://doi.org/10.1016/J.MLWA.2025.100631

Nair, B. B., Kumar, P. K. S., Sakthivel, N. R., & Vipin, U. (2017). Clustering stock price time series data to generate stock trading recommendations: An empirical study. *Expert Systems with Applications*, *70*, 20–36. https://doi.org/10.1016/j.eswa.2016.11.002

Navarro, M. M., Young, M. N., Prasetyo, Y. T., & Taylar, J. V. (2023). Stock market optimization amidst the COVID-19 pandemic: Technical analysis, K-means algorithm, and mean-variance model (TAKMV) approach. *Heliyon*, *9*(7), e17577. https://doi.org/10.1016/J.HELIYON.2023.E17577

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, *9*(1). https://doi.org/10.1186/2251-712X-9-1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews, 10*(89). https://doi.org/10.1186/s13643-021-01626-4

Pei, D., Luo, C., & Liu, X. (2023). Financial trading decisions based on deep fuzzy self-organizing map. *Applied Soft Computing*, *134*, 109972. https://doi.org/10.1016/J.ASOC.2022.109972

Pölzlbauer, G., Dittenbach, M., & Rauber, A. (2005). A visualization technique for Self-Organizing Maps with vector fields to obtain the cluster structure at desired levels of detail. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1558–1563. https://doi.org/10.1109/IJCNN.2005.1556110

Rauber, A., Merkl, D., & Dittenbach, M. (2002). The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, *13*(6), 1331–1341. https://doi.org/10.1109/TNN.2002.804221

Rieger, M. O. (2022). Uncertainty avoidance, loss aversion and stock market participation. *Global Finance Journal*, *53*, 100598. https://doi.org/10.1016/J.GFJ.2020.100598

Sáenz, J., V., Quiroga, F. M., & Bariviera, A. F. (2023). Data vs. information: Using clustering techniques to enhance stock returns forecasting. *International Review of Financial Analysis*, *88*, 102657. https://doi.org/10.1016/J.IRFA.2023.102657

Sagaceta-Mejía, A. R., Sánchez-Gutiérrez, M. E., & Fresán-Figueroa, J. A. (2024). An Intelligent Approach for Predicting Stock Market Movements in Emerging Markets Using Optimized Technical Indicators and Neural Networks. *Economics*, *18*(1), 20220073.

Salehpour, A., & Samadzamini, K. (2023). Machine Learning Applications in Algorithmic Trading: A Comprehensive Systematic Review. *International Journal of Education and Management Engineering*, *13*(6), 41–53. https://doi.org/10.5815/ijeme.2023.06.05

Samuelson, Paul A. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review*, *6*(2), 41-49.

Sarainmaa, O. (2024). *Swing Trading the S&P500 Index with Technical Analysis and Machine Learning Methods with Responsible Way*. Master thesis. Faculty of Social Sciences, Business and Economics, and Law. https://www.doria.fi/handle/10024/189661

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Shi, Y., Li, B., Du, G., & Dai, W. (2021). Clustering framework based on multi-scale analysis of intraday financial time series. *Physica A: Statistical Mechanics and Its Applications*, *567*, 125728. https://doi.org/10.1016/J.PHYSA.2020.125728

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/S10115-023-02010-5/TABLES/6

Wang, C. (2022). Pattern Classification of Stock Price Moving. *Frontiers in Computing and Intelligent Systems*, *2*(2), 32–41. https://doi.org/10.54097/FCIS.V2I2.3754

Wilhelmina Afua Addy, Adeola Olusola Ajayi-Nifise, Binaebi Gloria Bello, Sunday Tubokirifuruar Tula, Olubusola Odeyemi, & Titilola Falaiye. (2024). Machine learning in financial markets: A critical review of algorithmic trading and risk management. *International Journal of Science and Research Archive*, *11*(1), 1853–1862. https://doi.org/10.30574/IJSRA.2024.11.1.0292

Wu, H., Long, H., Wang, Y., & Wang, Y. (2021). Stock index forecasting: A new fuzzy time series forecasting method. *Journal of Forecasting*, *40*(4), 653–666. https://doi.org/10.1002/FOR.2734

Wu, S. (2020). Application of Cluster Analysis in Stock Selection in United States Stock Market. *Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning*, 310–313. https://doi.org/10.1145/3377571.3377628

Xu, Y., Yang, C., Peng, S., & Nojima, Y. (2020). A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning. *Applied Intelligence*, *50*(11), 3852–3867. https://doi.org/10.1007/S10489-020-01766-5

APPENDIX A – CORRELATION HEATMAPS

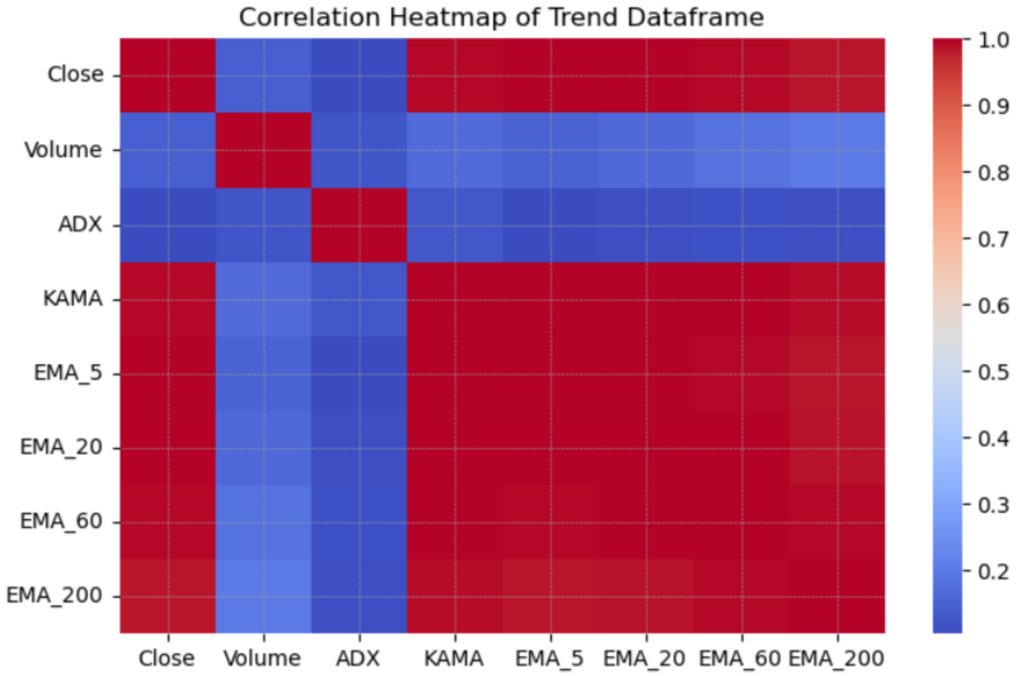

Figure 19 - Trend Feature Correlation Heatmap

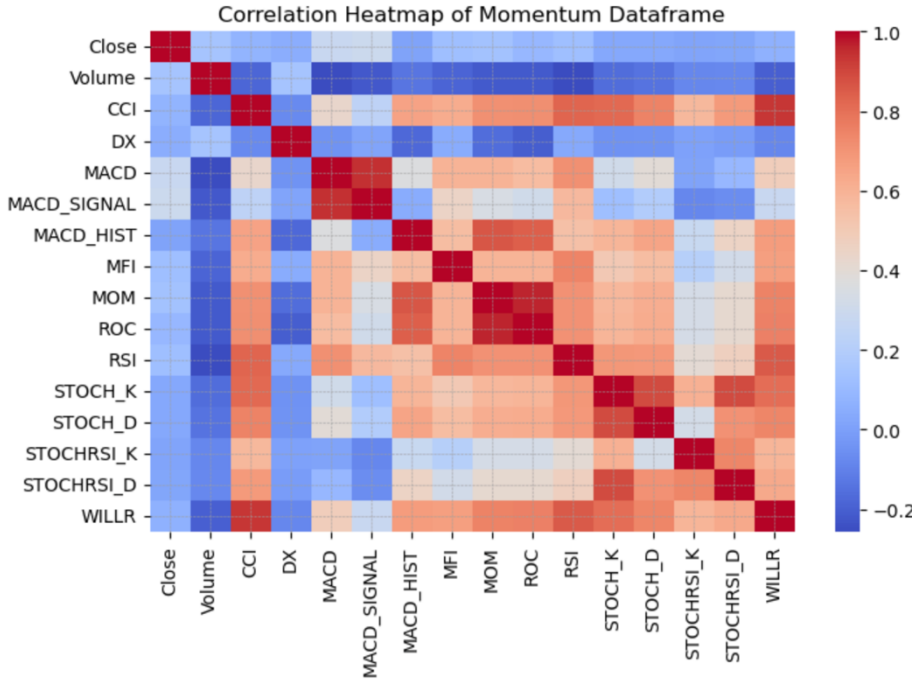

Figure 20 - Momentum Feature Correlation Heatmap

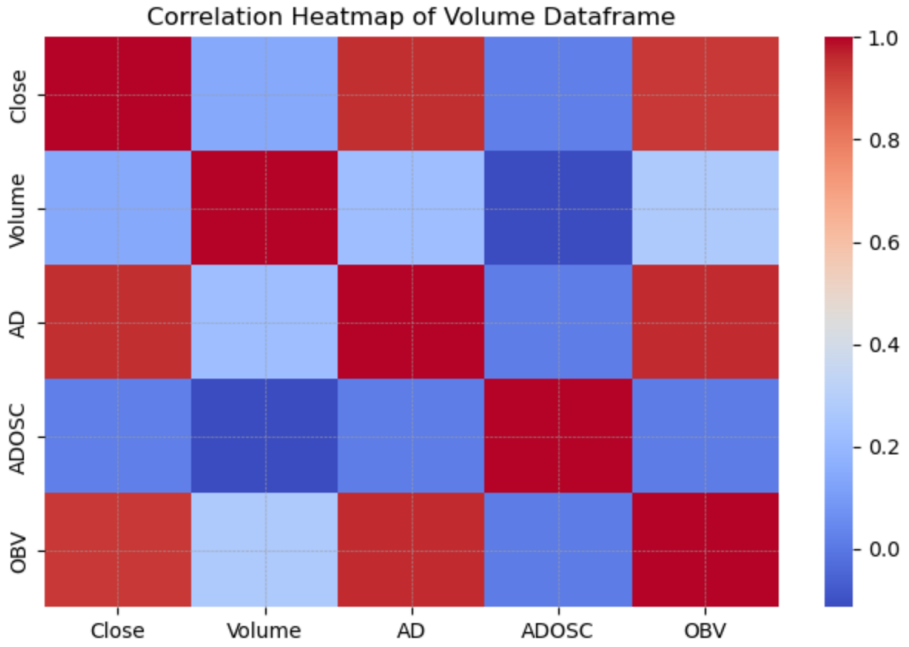

Figure 21 - Volume Feature Correlation Heatmap

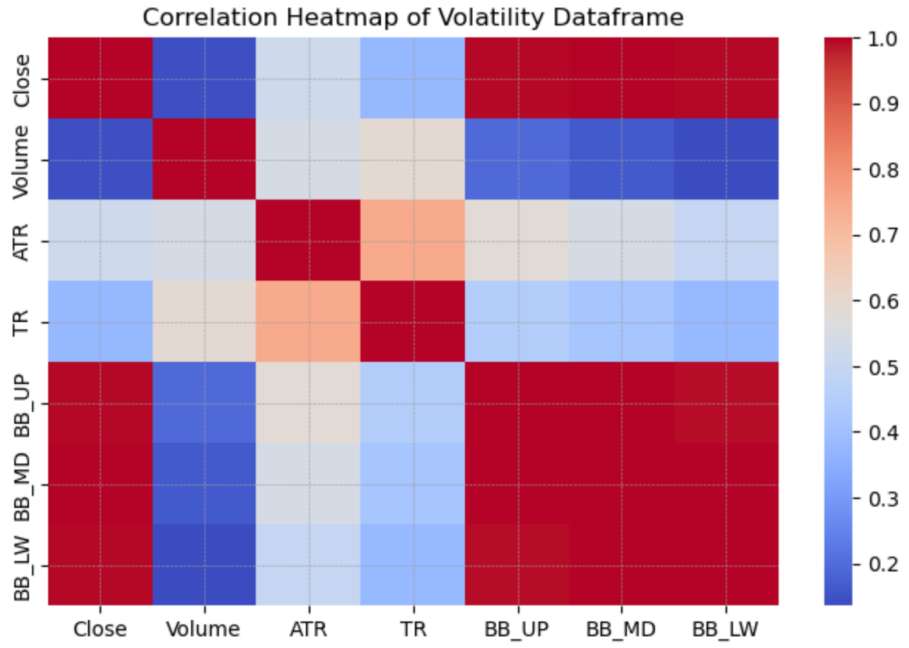

Figure 22 - Volatility Feature Correlation Heatmap

APPENDIX B – FINAL DATASETS STATISTICAL TABLES

Table 16 - Trend Final Dataset Statistics

	count	mean	std	min	25%	50%	75%	max
Close	2516.00	3356.12	1083.90	1829.08	2432.29	3004.99	4204.60	6090.27
Volume	2516	4007033048	953984062	1296530000	3428152500	3819815000	4341680000	9976520000
ADX	2516.00	22.94	8.24	9.45	16.79	21.30	27.86	55.05
EMA_20	2516.00	3341.33	1071.96	1892.72	2428.58	2974.27	4191.74	6023.63
EMA_200	2516.00	3213.53	986.58	1955.51	2311.93	2944.54	4121.01	5560.26

Table 17 - Momentum Final Dataset Statistics

	count	mean	std	min	25%	50%	75%	max
Close	2516.00	3356.12	1083.90	1829.08	2432.29	3004.99	4204.60	6090.27
Volume	2516	4007033048	953984062	1296530000	3428152500	3819815000	4341680000	9976520000
CCI	2516.00	27.98	106.51	-355.74	-48.02	54.87	108.67	318.65
DX	2516.00	22.94	15.71	0.02	10.18	20.94	32.81	78.17
MACD	2516.00	10.95	36.63	-237.02	-3.49	14.84	33.41	92.58
MACD_HIST	2516.00	0.00	11.91	-61.22	-4.38	0.34	5.13	60.90
MFI	2516.00	54.84	14.33	12.66	44.39	55.87	64.98	93.19
ROC	2516.00	0.47	3.15	-23.39	-0.86	0.75	2.26	19.05
RSI	2516.00	55.64	11.33	16.77	47.55	57.06	63.93	86.69
STOCH_K	2516.00	60.49	25.66	1.21	39.74	63.75	83.52	99.81
STOCHRSI_K	2516.00	55.02	41.70	0.00	2.40	61.58	100.00	100.00
WILLR	2516.00	-34.97	30.41	-100.00	-58.44	-25.26	-8.01	-0.00

Table 18 - Volume Final Dataset Statistics

	count	mean	std	min	25%	50%	75%	max
Close	2516.00	3356.12	1083.90	1829.08	2432.29	3004.99	4204.60	6090.27
Volume	2516	4007033048	953984062	1296530000	3428152500	3819815000	4341680000	9976520000
AD	2516.00	940751177748.39	311908087712.76	410079434726.77	686153788639.99	942783092504.94	1216210877307.03	1467775794795.58
ADOSC	2516.00	1441653595.36	2611353940.84	-7531260938.26	-367921450.24	1496915056.96	3375177847.29	9330119610.61

Table 19 - Volatility Final Dataset Statistics

	count	mean	std	min	25%	50%	75%	max
Close	2516.00	3356.12	1083.90	1829.08	2432.29	3004.99	4204.60	6090.27
Volume	2516	4007033048	953984062	1296530000	3428152500	3819815000	4341680000	9976520000
ATR	2516.00	41.27	23.71	9.31	20.96	38.27	55.01	152.76
TR	2516.00	41.52	32.83	4.48	18.31	32.08	54.80	330.08

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa