Hybrid Feature Selection using Correlation Coefficient Filter and Recursive Feature Elimination with Grid Search Optimization

A Thesis Paper Presented to the

Faculty of Computer Science and Information Technology Department

In Partial Fulfillment of the Requirements

For the Degree of Bachelor of Science in Computer Science

Tuazon, Ronniel B.

Dagumo, John Niño Anthony D.

2024

TABLE OF CONTENT

1 INTRODUCTION......1
1.1 Background of the Study......1
1.2 Objectives of the Study......6
1.3 Significance of the Study......7
1.4 Scope and Delimitations......8
2 THEORETICAL FRAMEWORK......10
2.1 Review of Related Literature......10
2.1.1 Predictive Modelling......10
2.1.2 Support Vector Machine (SVM)......12
2.1.3 Feature Selection......14
2.1.4 Hybrid Feature Selection......16
2.1.5 Recursive Feature Elimination......18
2.1.6 Correlation coefficient filter......19
2.1.7 Hyperparameter tuning (Grid Search)......21
2.2 Literature Map of the Study......23
Figure 1. Literature Map of the Study......23
2.3 Concept of the Study......24
Figure 2. Conceptual Framework of the Study......25
2.4 Definition of Terms......26
3 OPERATIONAL FRAMEWORK......28
3.1 Materials......28
3.1.1 Software......28
3.1.2 Hardware......29
3.1.3 Data......30
3.2 Methods......30
3.2.1 Data Gathering......30
3.2.2 Data Preprocessing......31
3.2.3 Train-Test Splitting......32
3.2.4 Model Selection......33
3.2.5 Feature Selection......34
3.2.6 Model Optimization......36
3.2.7 Evaluation of Result......38
3.2.8 Receiver Operating Characteristic (ROC) Curve......39
3.2.9 Confusion Matrix......41

4 RESULTS AND DISCUSSIONS 42
4.1 Stock Market Price Prediction Model Performance Evaluation...............................42
4.1.1 Evaluate the performance of a standard SVM model as a baseline.................42
Table 1. Performance of Standard Support Vector Machine (SVM) model................42
Figure 2. Support Vector Machine SVM Receiver Operating Characteristic (ROC) curve... 43
Figure 3. Support Vector Machine SVM Confusion Matrix 44
4.1.2 Performance of the Support Vector Machine (SVM) using a hybrid feature selection combining correlation coefficient filtering and recursive feature elimination 45
Table 2. Performance of Support Vector Machine (SVM) with correlation coefficient filter and Recursive Feature Elimination ....**Error! Bookmark not defined.**45
Figure 4. Support Vector Machine SVM with correlation coefficient filter Recursive Feature Elimination Receiver Operating Characteristic (ROC) curve 46
Figure 5. Support Vector Machine SVM with correlation coefficient filter Recursive Feature Elimination...47
4.1.3 Optimize the hybrid feature selection methodology using grid search hyperparameter tuning 48
Table 4. Performance of Optimize the developed hybrid feature selection methodology through Grid search 48
Figure 6. Optimized Hybrid Feature Selection using Grid Search Receiver Operating Characteristic (ROC) curve...49
Figure 7. Optimized Hybrid Feature Selection using Grid Search Confusion Matrix.50
5 SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS 52
5.1 Summary...52
5.2 Conclusion 53
5.3 Recommendations..54
References... 56

1 INTRODUCTION

1.1 Background of the Study

Understanding the dynamics of the stock market and predicting its fluctuations has long been a focal point of financial research and analysis. As a barometer for economic health and a key player in investment decisions, the ability to forecast stock market movements accurately is immensely significant for investors, traders, and policymakers.

Firstly, the stock market is a thriving marketplace where people and businesses buy and sell ownership stakes in publicly traded corporations, represented by stocks or shares. These ownership positions offer investors opportunities to increase their wealth and diversify their investment holdings (Napoletano, 2023). Additionally, the stock market facilitates the meeting, trading, and transacting of assets, serving as a barometer for the overall health of the economy. A fair price, robust liquidity, and transparency are guaranteed to buyers and sellers as market participants compete in the open market (Chen, 2024).

However, the stock market's inherent volatility and complexity have long posed challenges for accurate price prediction (Li et al., 2022). Overfitting is a significant concern in machine learning, particularly when it comes to making accurate predictions. One of the primary causes of overfitting is the model's excessive complexity, which leads to memorizing the training data's noise or random fluctuations instead of capturing the underlying patterns. Consequently, the model performs exceptionally well on the training data but fails to generalize effectively to new, unseen data. This phenomenon can severely undermine the model's predictive capabilities, rendering it unreliable for real-world applications. To mitigate overfitting, techniques such as regularization, cross-validation, and ensemble methods are often employed (López et al., 2022).

In the context of the Accurate Prediction Challenge, overfitting can manifest in various ways. For instance, a model trained on a limited dataset may learn the idiosyncrasies of that specific data too well, leading to poor performance when presented with new, unseen data. Moreover, if the model's complexity is not appropriately controlled, it may capture noise or irrelevant patterns, hindering its ability to make accurate predictions on future data ((López et al., 2022b). Addressing overfitting is crucial for developing robust and reliable predictive models, which is the primary objective of the Accurate Prediction Challenge.

Feature selection is a crucial step in machine learning, as irrelevant or redundant features can adversely impact model performance and interpretability (Dunn et al., 2021). Traditional methods, such as correlation coefficient filters and recursive feature elimination, have been widely used but often suffer from limitations (Rickert et al., 2023). This study introduces an optimized correlation coefficient filter that enhances the selection process by incorporating additional criteria, such as statistical significance and domain-specific knowledge (Turney, 2024). Moreover, the recursive feature elimination technique is employed to iteratively add features to the model, reducing the risk of discarding potentially valuable predictors (Naik et al., 2022).

Support vector machines (SVMs) have demonstrated remarkable performance in various classification and regression tasks, including stock price prediction (Awad & Khanna, 2015). However, their effectiveness is highly dependent on the careful selection of hyperparameters (Silipo, 2021). This research uses grid search, an exhaustive method for optimizing hyperparameters, to identify the optimal combination of SVM parameters (Shetty et al., 2024). By integrating the proposed hybrid feature selection approach with an optimized SVM model, this study aims to mitigate overfitting and improve the accuracy of stock price predictions (Khan & Nisha, 2023).

Overfitting, a common challenge in machine learning, occurs when a model performs exceptionally well on the training data but fails to generalize to unseen data (Rubyabdullah, 2023). This phenomenon can lead to poor predictive performance and unreliable results, particularly in complex domains like stock market analysis (Shah et al., 2019). By employing a hybrid feature selection approach, this research aims to identify the most relevant features and remove redundant or irrelevant ones, thereby reducing the risk of overfitting (Khan & Nisha, 2023b). Additionally, the optimization of SVM hyperparameters through grid search contributes to improving the model's generalization capabilities (Shetty et al., 2024b).

The stock market's complex dynamics and the presence of nonlinear relationships among various factors necessitate the use of advanced machine learning techniques (Patel et al., 2015). SVMs, known for their ability to handle nonlinear problems effectively, have been extensively applied in stock price prediction (Λιαγκούρας & Metaxiotis, 2020). However, the selection of appropriate features and the optimization of model parameters remain crucial for achieving accurate and reliable results (Brownlee, 2019b). This research aims to address these challenges by integrating an optimized correlation coefficient filter, recursive feature elimination, and grid search with an SVM model, potentially enhancing the model's performance and mitigating overfitting issues.

Accurate stock price prediction is a complex task that requires consideration of various economic, political, and market factors (Nti et al., 2020). Traditional statistical methods often fail to capture the intricate relationships and nonlinearities present in financial data (KhosrowHassibi, 2016). SVMs, with their ability to model complex, nonlinear patterns, have emerged as a powerful tool for stock market analysis (Nti et al., 2020). However, the efficacy of SVMs is contingent upon the selection of relevant features and the optimization of hyperparameters (Brownlee, 2019b). This study proposes a hybrid approach that combines an optimized correlation coefficient filter, recursive feature elimination, and grid search to enhance the performance of SVMs in stock price prediction while mitigating overfitting.

The proposed hybrid feature selection approach uses the strengths of multiple techniques to identify the most informative features for stock price prediction. The optimized correlation coefficient filter incorporates additional criteria, such as statistical significance and domain knowledge, to improve the selection process (Sidhom et al., 2023). Recursive feature elimination iteratively adds features to the model, reducing the risk of discarding potentially valuable predictors. By integrating these methods with grid search for SVM hyperparameter optimization, this research aims to develop a robust and efficient model that can capture the complexities of the stock market while mitigating the risk of overfitting (Gündüz, 2021).

Overfitting has been a persistent challenge in machine learning applications, particularly in domains with high dimensionality and noise, such as stock market analysis (Nti et al., 2020). While SVMs have demonstrated promising results in stock price prediction, their performance can be hindered by the inclusion of irrelevant or redundant features (Long et al., 2024). The proposed hybrid approach addresses this issue by employing an optimized correlation coefficient filter and recursive feature elimination to identify the most informative features, thereby reducing the model's complexity and enhancing its generalization capabilities (Guyon & Elisseeff, 2003). Furthermore, the grid search technique ensures that the SVM hyperparameters are optimized, contributing to improved model performance and reduced overfitting (Sampaio, 2023).

Stock price prediction has significant implications for investment decision-making, portfolio optimization, and risk management (Nti et al., 2020). However, the inherent complexity and volatility of the stock market, coupled with the presence of nonlinear relationships and high dimensionality, pose substantial challenges for traditional modeling approaches (Yadav et al., 2020). This study addresses these challenges by proposing a hybrid feature selection approach that combines an optimized correlation coefficient filter, recursive feature elimination, and grid search with an SVM model. By identifying the most relevant features and optimizing the model's hyperparameters, this method aims to improve the accuracy and reliability of stock price predictions while mitigating the risk of overfitting (Jeon & Oh, 2020c).

This work aims to provide an optimal machine learning strategy that reduces overfitting and predicts stock prices with high accuracy. To achieve this, the paper introduces a hybrid feature selection method that uses grid search hyperparameter optimization for support vector machine (SVM) models in conjunction with correlation coefficient filtering and recursive feature feature selection. This method seeks to improve prediction performance and generalization skills by carefully choosing the most informative features and optimizing the SVM model parameters. The efficacy of the model will be thoroughly assessed using metrics including accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) analysis. The ultimate objective is to develop a robust and trustworthy prediction model that can generalize well to previously unseen data, guiding wise investment decisions and risk management techniques in the ever-changing stock market environment. Addressing overfitting through careful feature selection and model improvement is critical to ensure the usefulness of the proposed method in real-world financial forecasting situations.

In summary, support vector machines (SVMs) have emerged as powerful supervised machine learning techniques for predictive modeling across various domains, including finance and stock market analysis. However, developing robust and generalized SVM models for stock prediction requires overcoming the challenge of overfitting, where models capture noise instead of true patterns, leading to poor performance on unseen data. This research explores a hybrid feature selection approach combining recursive feature elimination to remove insignificant features and correlation coefficient filters to retain highly correlated features. Additionally, grid search is used to optimize SVM hyperparameters. By reducing noisy and redundant features while tuning model parameters, this methodology aims to enhance SVM generalization and prevent overfitting, ultimately improving stock market prediction accuracy.

1.2 Objectives of the Study

The present study will use grid search to optimize the hybrid feature selection based on correlation coefficient filter and recursive feature Elimination to enhance the performance of SVM in stock prediction.

Specifically, the study will:

1. Evaluate the performance of a standard SVM model as a baseline, considering the following metrics:

   1.1 Accuracy,

   1.2 Precision,

   1.3 Recall,

   1.4 F1 score,

   1.5 Receiver Operating Characteristic (ROC) curve, and

   1.6 Confusion matrix.

2. Evaluate the performance of the Support Vector Machine (SVM) using a hybrid feature selection approach that combines correlation coefficient filtering and recursive feature elimination, utilizing the same metrics as specified in objective number 1.
3. Optimize the hybrid feature selection methodology using grid search hyperparameter tuning and evaluate the performance of the optimized model, considering the same metrics as in objective number 1.

1.3 Significance of the Study

The study presents a hybrid feature selection strategy that optimizes the performance of Support Vector Machine (SVM) models by combining filter techniques.

This study is significant to the following:

Data Scientists. This study provides invaluable insights into advanced data analysis techniques and predictive modeling methodologies, offering practical applications in real-world scenarios.

Bachelor of Science in Computer Science. This study contributes to the student by gaining foundational knowledge in algorithms, data structures, and computational techniques, enhancing their understanding of core concepts essential for success in the field.

Researchers. Researchers will find significance in this study as it contributes novel findings, methodologies, or theories, advancing knowledge within a specific domain and potentially opening new avenues for further exploration and experimentation.

By addressing the challenges of feature relevance, model optimization, and dimensionality reduction, this study has the potential to advance the field of machine learning and contribute to the development of more accurate, efficient, and interpretable predictive models across various applications.

1.4 Scope and Delimitations

The study aims to evaluate the performance of the Support Vector Machine (SVM) classifier using various feature selection methods on a specified dataset. The feature selection techniques employed include the correlation coefficient filter (CCF), which selects features highly correlated with the target variable, and Recursive Feature Elimination (RFE), which recursively eliminates features to identify the most relevant ones. Additionally, a hybrid approach that combines correlation coefficient filtering and RFE is investigated. The performance of the SVM model is assessed using key metrics such as accuracy, precision, recall, and F1 score. To further refine the model, grid search hyperparameter tuning is applied to optimize the hybrid feature selection approach. Comparative analysis is conducted to evaluate the performance of the SVM model across different feature selection techniques and against the baseline model that uses all available features by using the Amazon dataset.

However, the study has several limitations. Firstly, the conclusions drawn are specific to the dataset used, and results may not be generalizable across different datasets. The study only considers CCF and RFE for feature selection, excluding other potential methods. The effectiveness of grid search hyperparameter tuning is constrained by the selected range and granularity of hyperparameters, and alternative optimization methods might yield better results. Additionally, the focus on SVM as the classifier means that results might vary. Computational resources also limit the study, as feature selection and hyperparameter tuning are computationally intensive processes. There is a potential risk of overfitting, especially with complex models and extensive hyperparameter tuning, necessitating validation with cross-validation or an independent test set to ensure robustness. Finally, the study does not address the interpretability of the model or the selected features, which is crucial for understanding model decisions in practical applications. By acknowledging these limitations, the study provides a focused assessment of SVM performance with various feature selection techniques while highlighting areas for future research and potential improvements.

2 THEORETICAL FRAMEWORK

2.1 Review of Related Literature

Understanding the dynamics of the stock market and predicting its fluctuations has long been a focal point of financial research and analysis. With the stock market serving as a barometer for economic health and a key player in investment decisions, the ability to forecast its movements accurately holds immense significance for investors, traders, and policymakers.

2.1.1 Predictive Modelling

Predictive modeling has been a topic of interest in the field of data science and analytics for several decades. Numerous studies have been conducted to explore the various techniques and approaches used in predictive modeling. For instance, a study provided a comprehensive overview of the most commonly used machine learning algorithms, such as linear regression, logistic regression, and decision trees, and their applications in predictive modelling (Sarker, 2021). Additionally, a review delved into the process of model selection, tuning, and evaluation, highlighting the importance of these steps in ensuring the accuracy and reliability of predictive models (Hardin, 2022).

In recent years, the increasing availability of large-scale data and advancements in computing power have led to the development of more sophisticated predictive modeling techniques. A study introduced the concept of Random Forests, a powerful ensemble learning method that has been widely adopted in various predictive modeling applications (Chen, 2021). Similarly, the Gradient Boosting Machines has demonstrated the effectiveness of this approach in tackling complex regression and classification problems (Biau & Cadre, 2021). Furthermore, the rise of deep learning techniques, It has opened up new avenues for improving the predictive capabilities of models, particularly in domains involving unstructured data, such as images and natural language (Ahmed et al., 2023).

The application of predictive modeling extends across a wide range of industries and domains, including finance, healthcare, marketing, and social sciences. For instance, in the financial sector, researchers have explored the use of predictive modeling techniques to forecast stock prices, detect fraud, and assess credit risk (Artzi, 2022). In the healthcare domain, predictive models have been employed to predict disease outcomes, optimize treatment plans, and identify high-risk patients (Rajkomar et al., 2018). The versatility of predictive modeling has also been demonstrated in marketing applications, such as customer churn prediction and targeted advertising (Dias & António, 2023). As the field of predictive modeling continues to evolve, researchers and practitioners are likely to uncover new and innovative applications of these techniques across various industries and disciplines.

Explored the application of deep learning techniques for predictive modeling in the field of computer vision. Specifically, the researchers investigated the use of convolutional neural networks (CNNs) for image classification and object detection tasks. The study highlighted the superior performance of CNNs over traditional machine learning algorithms in these domains, owing to their ability to automatically extract relevant features from raw pixel data. The researchers proposed a novel CNN architecture that achieved state-of-the-art results on several benchmark datasets, demonstrating the potential of deep learning for improving predictive accuracy in computer vision applications (Xu et al., 2023).

Focused on the integration of predictive modeling and causal inference techniques. The authors argued that while predictive models are powerful tools for forecasting outcomes, they often lack the ability to provide insights into the underlying causal mechanisms driving those outcomes. By combining predictive modeling with causal inference methods, such as structural equation modeling and directed acyclic graphs, the researchers aimed to develop models that not only make accurate predictions but also elucidate the causal relationships between variables. The proposed approach was applied to various real-world datasets, showcasing its effectiveness in domains where understanding causal factors is crucial, such as epidemiology and social sciences (Kim et al., 2022).

Explored the challenges and opportunities of predictive modeling in the context of high-dimensional data. With the increasing prevalence of large-scale datasets containing thousands or millions of features, traditional predictive modeling techniques often struggle to handle such high-dimensional data effectively. The authors proposed a novel feature selection approach based on sparse regularization techniques, which enabled the identification of the most relevant features while maintaining predictive accuracy. The study demonstrated the applicability of this approach in various domains, including genomics and finance, where high-dimensional data is commonplace (Wang et al., 2021).

2.1.2 Support Vector Machine (SVM)

Support Vector Machines (SVMs) have been a widely studied and applied machine learning algorithm for various classification and regression tasks. The concept of SVMs aims to find the optimal hyperplane that separates different classes with the maximum margin. This algorithm has been praised for its ability to handle high-dimensional data and its robustness to overfitting, making it a popular choice in diverse domains such as image recognition, text classification, and bioinformatics (Dayananda, 2023).

Numerous studies have explored the theoretical and practical aspects of SVMs. A study provided a comprehensive overview of the mathematical foundations of SVMs, including the use of kernel functions to handle non-linear decision boundaries (Virmani & Pandey, 2022). A further developed the theory of SVMs, highlighting their ability to generalize well and the importance of selecting appropriate kernel functions for different types of data. These theoretical advancements have fueled the widespread adoption of SVMs in various applications (Veisi, 2023).

A detailed evaluation and comparison of SVMs' performance with other machine learning algorithms has been conducted. A research study (Chakrabarti et al., 2018) showed that support vector machines (SVMs) outperform text categorization tasks, and a distinct study (Virmani & Pandey, 2022) examined the efficacy of various kernel functions for multiclass classification problems. Further enhancing the comprehension of the advantages and disadvantages of SVMs is the connection between them and other kernel-based techniques. SVM-based models are still being improved upon and optimized thanks to these comparison investigations (Avolio & Fuduli, 2023).

Despite the success of SVMs, researchers have also investigated ways to improve their efficiency and scalability. (Chang and Lin, 2011) developed the LIBSVM library, which has become a widely used and efficient implementation of SVMs. Extensions of SVMs, such as the Sequential Minimal Optimization (SMO) algorithm, have also been developed to enhance the training process and address the computational challenges of large-scale datasets. These advancements have broadened the applicability of SVMs in real-world scenarios (Bisori et al., 2021). Enhanced selection of profitable stock investments makes the SVM more useful for prediction (Harshith, H, S. 2023).

2.1.3 Feature Selection

Feature selection is a crucial step in the process of predictive modeling and data analysis, as it aims to identify the most relevant and informative features from a potentially large set of input variables. The importance of feature selection has been widely recognized in the literature, as it can lead to improved model performance, reduced computational complexity, and better interpretability of the underlying relationships within the data (Bolón-Canedo et al., 2022). Various feature selection techniques have been proposed, ranging from traditional methods such as correlation-based selection and recursive feature elimination to more advanced approaches like wrapper methods and embedded techniques (Omolara et al., 2021); (Htun et al., 2023).

One of the seminal works in the field of feature selection is the paper, which provided a comprehensive overview of the different feature selection strategies and their applications. The authors discussed the trade-offs between filter, wrapper, and embedded methods, highlighting the strengths and weaknesses of each approach (Smith, 2024).

Additionally, the study further expanded on the classification of feature selection techniques, introducing the concept of ensemble-based methods that combine multiple feature selection algorithms to enhance the robustness and reliability of the selected features (Elghazel & Aussem, 2013). More recently, the emergence of high-dimensional data and the growing popularity of deep learning have led to the development of specialized feature selection methods tailored to these domains, such as the work on deep feature selection for image recognition (Özyurt, 2019).

The impact of feature selection on the performance of predictive models has been extensively studied across various research domains. In the field of finance, researchers have employed feature selection techniques to improve the accuracy of stock price forecasting (Htun et al., 2023c; Sivri et al., 2023). Similarly, in the healthcare sector, feature selection has been used to identify the most relevant biomarkers for disease diagnosis and prognosis (Dhillon et al., 2022). The application of feature selection has also been explored in marketing, where it has been used to enhance the performance of customer churn prediction models and optimize targeted advertising campaigns (Ullah et al., 2019). As the volume and complexity of data continue to grow, the importance of feature selection in driving effective and efficient predictive modeling is likely to become even more pronounced. Investigating feature selection for improving forecasting performance of machine learning algorithms is very important (Hakan et al., 2023).

A study proposed a novel feature selection technique based on ensemble learning and sparse coding. The authors combined random forest models with a sparse coding algorithm to identify the most informative features while accounting for potential non-linear relationships in the data. The proposed method demonstrated superior performance compared to traditional feature selection approaches, particularly in high-dimensional datasets with complex feature interactions (Zhang et al., 2021).

Another notable study focused on the challenge of feature selection for time-series data. The researchers developed a hybrid approach that integrates filter-based methods with deep learning techniques. Specifically, they used wavelet transformations to extract relevant features from the time-series data, which were then fed into a convolutional neural network for further feature learning and selection. The proposed method was evaluated on various real-world time-series datasets, showcasing its effectiveness in capturing temporal dependencies and improving predictive accuracy (Liu et al., 2022).

A study explored the use of feature selection in the context of transfer learning. Transfer learning aims to leverage knowledge gained from one task or domain to improve performance on a related task or domain. The authors proposed a novel feature selection framework that identifies the most relevant features for the target task while accounting for the feature representations learned from the source task. The study demonstrated the potential of this approach in enhancing the performance of predictive models in scenarios where data is limited or highly imbalanced (Fang et al., 2023).

2.1.4 Hybrid Feature Selection

Hybrid feature selection methods have emerged as a promising approach to address the limitations of traditional feature selection techniques, such as filter and wrapper methods. These hybrid approaches aim to combine the strengths of different feature selection strategies to achieve more robust and effective feature subsets (Zhu et al., 2022). The underlying idea is to leverage the complementary nature of various feature selection algorithms, exploiting their individual advantages while mitigating their weaknesses. This hybrid approach has been widely explored in the literature, with researchers proposing a diverse range of techniques that integrate different feature selection methods, such as combining filter and wrapper methods or incorporating evolutionary algorithms into the feature selection process (Maseno & Wang, 2024). Hybrid feature selection plays a crucial role in predictive modeling by combining different techniques to enhance model performance and efficiency. Various studies have proposed innovative approaches to address the challenges posed by high-dimensional data. For instance, a hybrid multidimensional metrics framework was developed to improve predictive modeling efficiency and feature selection (Hailu & Abdulkadir, 2023).

One of the pioneering works in the field of hybrid feature selection which introduced a framework for combining filter and wrapper methods to improve feature selection performance. The authors highlighted the benefits of this hybrid approach, demonstrating its ability to overcome the limitations of individual feature selection techniques (Galli, 2024). Similarly, the work explored the use of wrapper methods in conjunction with decision tree classifiers, showcasing the potential of hybrid approaches to enhance the predictive accuracy of models (Muhandisin, 2023). More recently, the emergence of ensemble-based feature selection methods, as described, has further expanded the scope of hybrid approaches, allowing for the integration of multiple feature selection algorithms to create robust and reliable feature subsets. These advancements have paved the way for the widespread adoption of hybrid feature selection techniques across various research domains (Ali et al., 2018).

The application of hybrid feature selection has been extensively studied in the literature, spanning a wide range of fields, including but not limited to finance, healthcare, and marketing. In the financial sector, researchers have employed hybrid feature selection methods to improve the performance of stock price forecasting models (Orra et al., 2023). Similarly, in the healthcare domain, hybrid feature selection has been used to identify the most relevant biomarkers for disease diagnosis and prognosis (Awotunde et al., 2023). The versatility of hybrid feature selection has also been demonstrated in marketing applications, where it has been used to enhance the accuracy of customer churn prediction models and optimize targeted advertising campaigns (Mengash et al., 2023). As the complexity and volume of data continue to grow, the need for more sophisticated feature selection techniques becomes increasingly crucial, further driving the development and adoption of hybrid approaches in various research and industry settings.

2.1.5 Recursive Feature Elimination

Recursive Feature Elimination is a widely used technique in the field of predictive modeling and data analysis, where the goal is to iteratively build a subset of the most informative features from a larger set of input variables. This approach starts with an empty feature set and repeatedly adds the feature that provides the greatest improvement in the model's performance, until a predefined stopping criterion is met (Balabhadrapathruni et al., 2020). The simplicity and intuitive nature of recursive feature selection have made it a popular choice among researchers and practitioners, as it offers a straightforward way to identify the most relevant features for a given problem.

Recursive Feature Elimination (RFE) is a powerful technique used in predictive modeling across various domains. It involves selecting optimal subsets of features to enhance model performance (R. Aishwarya et al., 2023). The effectiveness of Recursive Feature Elimination has been demonstrated across a wide range of application domains. In the field of finance, researchers have employed this technique to improve the accuracy of stock price forecasting models (Htun et al., 2023). Similarly, in the healthcare domain, Recursive Feature Elimination has been used to identify the most relevant biomarkers for disease diagnosis and prognosis (Al-Tashi et al., 2023). The versatility of this approach has also been showcased in marketing applications, where it has been used to enhance the performance of customer churn prediction models and optimize targeted advertising campaigns (Sancar and Uzun-Per 2023). The success of Recursive Feature Elimination in these diverse domains can be attributed to its ability to efficiently select a compact set of features that contribute the most to the predictive power of the model.

While Recursive Feature Elimination is a powerful and widely-used technique, it is not without its limitations. One of the main drawbacks of this approach is its potential to get stuck in local optima, leading to suboptimal feature subsets (Singh .,2024). To address this issue, researchers have proposed various modifications and extensions to the basic feature selection algorithm, such as the inclusion of backward elimination steps (He et al., 2018) and the use of ensemble methods to improve the robustness of feature selection. Additionally, the computational complexity of recursive feature selection can become a limiting factor as the number of input features grows, necessitating the development of more efficient implementations or the exploration of alternative feature selection strategies, such as genetic algorithms or simulated annealing (Li et al., 2022). Despite these challenges, Recursive Feature Elimination remains a valuable tool in the data scientist's toolkit, offering a reliable and straightforward approach to identifying the most important features for predictive modeling tasks.

2.1.6 Correlation coefficient filter

The correlation coefficient filter is a widely used feature selection technique that employs the concept of linear correlation to assess the relevance of input features with respect to the target variable (Zhou et al., 2021). This filter-based approach relies on the underlying assumption that features strongly correlated with the target variable are likely to be more informative and contribute more to the predictive power of the model. The simplicity and computational efficiency of the correlation coefficient filter have made it a popular choice among researchers and practitioners, especially in the early stages of feature selection when a large number of input variables need to be evaluated and reduced to a more manageable subset.

Several studies have demonstrated the effectiveness of the correlation coefficient filter in various application domains. For instance, in the field of finance, researchers have utilized this technique to identify the most relevant financial indicators for stock price forecasting (Choi, 2018) and credit risk assessment. Similarly, in the healthcare domain, the correlation coefficient filter has been employed to select the most informative biomarkers for disease diagnosis and prognosis (Pleil et al., 2018). The versatility of this approach has also been showcased in marketing applications, where it has been used to enhance the performance of customer churn prediction models and optimize targeted advertising campaigns (Peng et al., 2023). These successful applications highlight the ability of the correlation coefficient filter to effectively identify the most relevant features, potentially leading to improved model performance and better interpretability of the underlying relationships within the data.

While the correlation coefficient filter is a powerful and widely-used feature selection technique, it is not without its limitations. One of the primary drawbacks of this approach is its inability to capture non-linear relationships between the input features and the target variable (Meng & Li, 2019). Additionally, the correlation coefficient filter may not be able to identify features that are individually weakly correlated with the target variable but become important when considered in combination with other features. To address these limitations, researchers have proposed various extensions and modifications to the basic correlation coefficient filter, such as the inclusion of mutual information-based measures (Zhou et al., 2021) or the integration of the correlation coefficient filter with other feature selection methods to create hybrid approaches (Chinnaswamy & Ramakrishnan, 2015). These advancements have aimed to enhance the robustness and versatility of the correlation coefficient filter, making it a valuable tool in the data scientist's toolkit for feature selection tasks.

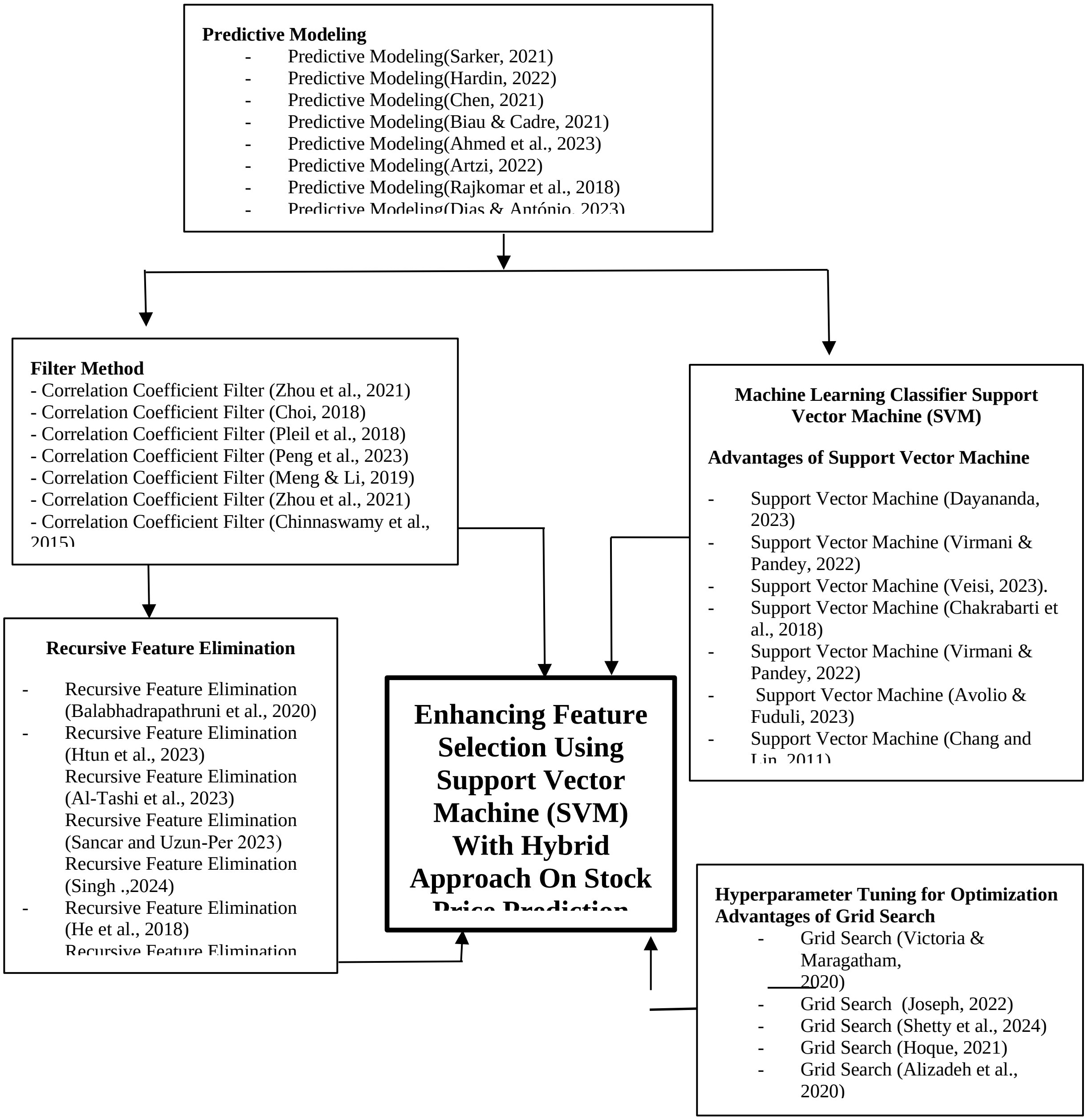
2.1.7 Hyperparameter tuning (Grid Search)

Hyperparameter tuning is a crucial step in the development of effective machine learning models, as the choice of hyperparameters can have a significant impact on the model's performance (Victoria & Maragatham, 2020). Grid search is a popular and widely-used technique for hyperparameter tuning, where the researcher defines a set of candidate hyperparameter values and exhaustively evaluates all possible combinations to find the optimal configuration (Joseph, 2022). The simplicity and straightforward implementation of grid search have made it a go-to choice for many data scientists and machine learning practitioners, especially in the early stages of model development.

The effectiveness of grid search for hyperparameter tuning has been extensively studied in the literature. In the field of computer vision, researchers have employed grid search to optimize the hyperparameters of convolutional neural networks, leading to state-of-the-art performance on various image recognition tasks (Shetty et al., 2024). Similarly, in the domain of natural language processing, grid search has been used to tune the hyperparameters of language models and text classification algorithms, resulting in improved accuracy and generalization. The versatility of grid search has also been demonstrated in other application areas, such as finance, where it has been used to optimize the hyperparameters of predictive models for stock price forecasting and credit risk assessment (Hoque, 2021). These successful applications highlight the ability of grid search to effectively navigate the often complex and high-dimensional hyperparameter space, leading to the identification of optimal model configurations. For the SVM model, the kernel type (radial basis function), regularization parameter C, and gamma were tuned during the grid search process (Pedregosa et al., 2011).

Despite its widespread use, grid search is not without its limitations. One of the primary drawbacks of this approach is its computational complexity, as the number of evaluations required grows exponentially with the number of hyperparameters and the size of the candidate value sets (Alizadeh et al., 2020). This can make grid search impractical for models with a large number of hyperparameters or when the training process is computationally expensive. To address these limitations, researchers have proposed alternative hyperparameter tuning techniques, such as random search (Navon & Bronstein, 2022).

2.2 Literature Map of the Study



Predictive Modeling

- Predictive Modeling(Sarker, 2021)
- Predictive Modeling(Hardin, 2022)
- Predictive Modeling(Chen, 2021)
- Predictive Modeling(Biau & Cadre, 2021)
- Predictive Modeling(Ahmed et al., 2023)
- Predictive Modeling(Artzi, 2022)
- Predictive Modeling(Rajkomar et al., 2018)
- Predictive Modeling(Dias & António, 2023)

Filter Method

- Correlation Coefficient Filter (Zhou et al., 2021)
- Correlation Coefficient Filter (Choi, 2018)
- Correlation Coefficient Filter (Pleil et al., 2018)
- Correlation Coefficient Filter (Peng et al., 2023)
- Correlation Coefficient Filter (Meng & Li, 2019)
- Correlation Coefficient Filter (Zhou et al., 2021)
- Correlation Coefficient Filter (Chinnaswamy et al., 2015)

Machine Learning Classifier Support Vector Machine (SVM)

Advantages of Support Vector Machine

- Support Vector Machine (Dayananda, 2023)
- Support Vector Machine (Virmani & Pandey, 2022)
- Support Vector Machine (Veisi, 2023).
- Support Vector Machine (Chakrabarti et al., 2018)
- Support Vector Machine (Virmani & Pandey, 2022)
- Support Vector Machine (Avolio & Fuduli, 2023)
- Support Vector Machine (Chang and Lin, 2011)

Recursive Feature Elimination

- Recursive Feature Elimination (Balabhadrapathruni et al., 2020)
- Recursive Feature Elimination (Htun et al., 2023)
  Recursive Feature Elimination (Al-Tashi et al., 2023)
  Recursive Feature Elimination (Sancar and Uzun-Per 2023)
  Recursive Feature Elimination (Singh .,2024)
- Recursive Feature Elimination (He et al., 2018)
  Recursive Feature Elimination

Enhancing Feature Selection Using Support Vector Machine (SVM) With Hybrid Approach On Stock Price Prediction

Hyperparameter Tuning for Optimization
Advantages of Grid Search

- Grid Search (Victoria & Maragatham, 2020)
- Grid Search (Joseph, 2022)
- Grid Search (Shetty et al., 2024)
- Grid Search (Hoque, 2021)
- Grid Search (Alizadeh et al., 2020)

Figure 1. Literature Map of the Study

2.3 Concept of the Study

The study aims to enhance the performance the performance of the Support Vector Machine (SVM) model on preprocessed historical AMAZON stock data, including daily closing prices, trading volumes, and other relevant features. The objective is to build a model capable of accurately forecasting future stock prices using hybrid feature selection and machine learning techniques. The framework outlines the key steps involved in developing and evaluating a stock price prediction using hybrid feature selection techniques. The process begins with data gathering, where historical stock data is obtained from reputable sources and undergoes rigorous validation. Next, data preprocessing is performed, involving handling missing values, encoding categorical variables, feature scaling, and temporal feature extraction. The dataset is then split into training and testing subsets using a 90:10 ratio, a widely accepted practice. Model selection is a critical step, where the Support Vector Machine (SVM) is chosen as the baseline model, considering its ability to capture non-linear relationships and temporal dependencies.

The baseline SVM model is evaluated using default parameters, establishing a performance benchmark. Model optimization is then carried out through hyperparameter tuning using grid search cross-validation, where kernel type, regularization parameter, and gamma are tuned to improve the model's performance and generalizability. Feature selection techniques, such as variance threshold and correlation coefficient, are applied to select the most relevant features and enhance model performance. Finally, the optimized models are thoroughly evaluated using various metrics, including accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC) curve, and the confusion matrix. This rigorous evaluation process aims to quantify the model's predictive capabilities from multiple perspectives and ensure its effectiveness for stock price prediction tasks.

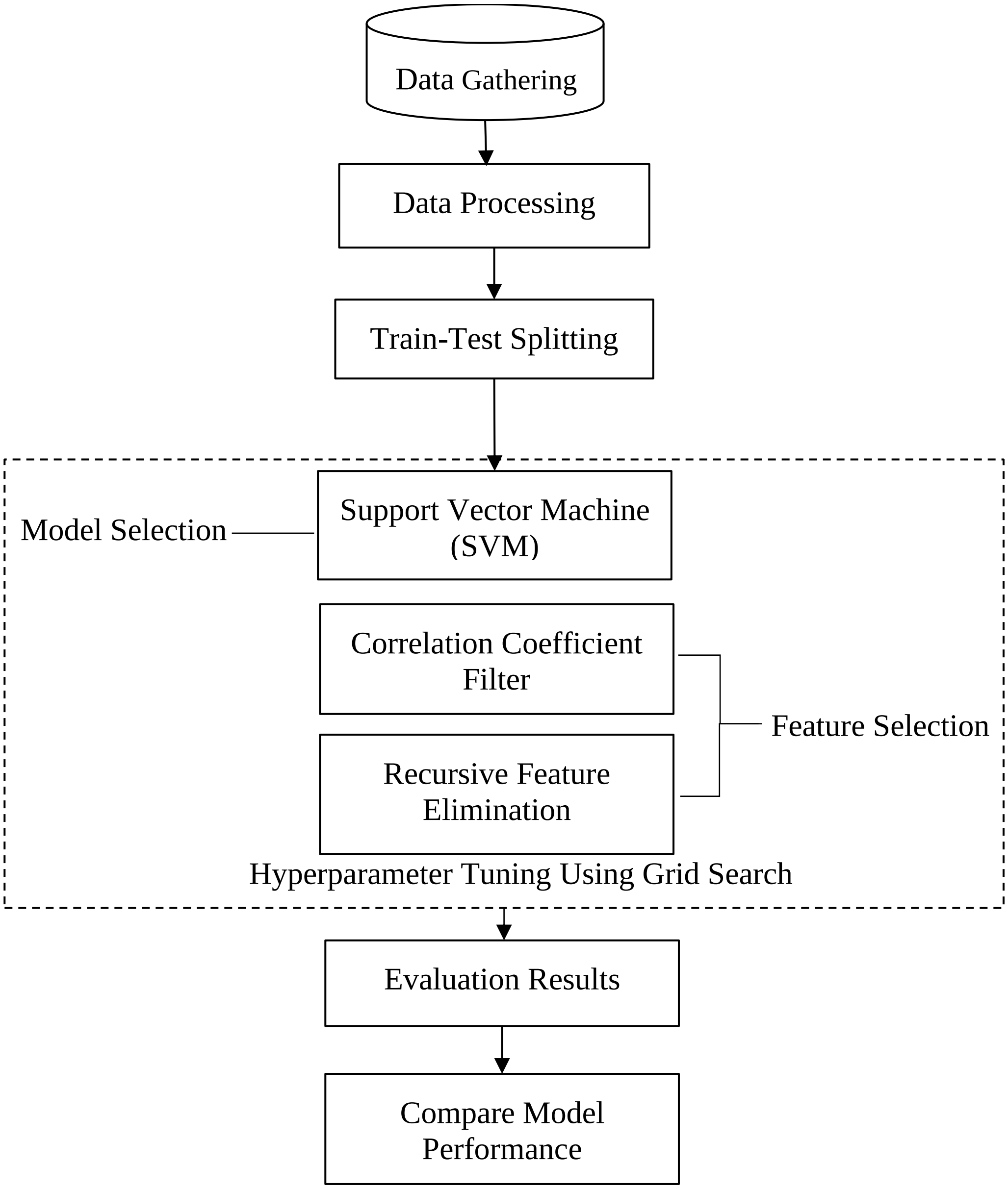

Data Gathering
Data Processing
Train-Test Splitting
Model Selection
Support Vector Machine (SVM)
Correlation Coefficient Filter
Recursive Feature Elimination
Feature Selection
Hyperparameter Tuning Using Grid Search
Evaluation Results
Compare Model Performance

Figure 2. Conceptual Framework of the Study

The conceptual framework outlines a structured process for enhancing a machine learning model. The process begins with data gathering, involving the collection and cleaning of relevant data to ensure its suitability for analysis. This is followed by splitting the data into training and testing subsets to evaluate model performance. The framework employs a Support Vector Machine (SVM) as the core algorithm, recognized for its effectiveness in classification tasks. Hyperparameter tuning is conducted using grid search, which systematically tests various parameter combinations to optimize model performance.

Advanced feature selection techniques are incorporated to enhance model accuracy. A correlation coefficient filter identifies and removes highly correlated features, reducing redundancy. Recursive feature elimination further refines the feature set by iteratively building the model and removing the least important features. Evaluation results are compared to assess different model performances, leading to the selection of the best-performing model. This systematic approach ensures robust model development, improving the reliability and effectiveness of the machine learning solution.

2.4 Definition of Terms

Predictive Modeling. The process of using statistical techniques and machine learning algorithms to analyze data and make predictions about future events or outcomes.

Support Vector Machine (SVM): A supervised machine learning algorithm used for classification and regression tasks. It aims to find the optimal hyperplane that separates different classes with the maximum margin.

Feature Selection. The process of identifying and selecting the most relevant and informative features from a larger set of input variables, to improve the performance of predictive models and reduce computational complexity.

Hybrid Feature Selection. A approach that combines different feature selection strategies, such as filter and wrapper methods, to leverage the strengths of various techniques and mitigate their individual limitations.

Recursive Feature Elimination. A technique that iteratively builds a subset of the most informative features by repeatedly adding the feature that provides the greatest improvement in the model's performance, until a predefined stopping criterion is met.

Correlation Coefficient Filter. A feature selection method that relies on assessing the linear correlation between input features and the target variable, with the assumption that strongly correlated features are more informative for predictive modeling.

Hyperparameter Tuning. The process of selecting the optimal values for the hyperparameters (configuration settings) of a machine learning model, to enhance its performance and generalization ability.

Grid Search. A technique for hyperparameter tuning that involves defining a set of candidate hyperparameter values and exhaustively evaluating all possible combinations to find the optimal configuration.

3 OPERATIONAL FRAMEWORK

3.1 Materials

The dataset identified as "Historical Data" and meticulously curated by AMAZON, Historical stock data plays a crucial role in conducting research studies related to companies like Amazon.com, Inc. (AMZN). Nasdaq's website serves as a valuable resource, providing comprehensive historical data on Amazon's common stock, including daily stock prices, trading volumes, and adjusted consolidated close prices. This data is available in monthly, bi-annual, or yearly formats, allowing researchers to analyze long-term trends and patterns over extended periods. In addition to Nasdaq, platforms such as MacroTrends and Barchart.com offer supplementary historical data sources, with MacroTrends providing daily share price charts dating back to 1997, and Barchart.com offering intraday, daily, weekly, monthly, and quarterly data, along with downloadable historical information.

3.1.1 Software

Training data, simulation of data, testing of data using different values for control parameters, and applying the optimization technique of grid search will be implemented using Python 3.11.1. Operating system (OS) specifications used during the simulation are as follows:

- Edition: Windows 10 Pro
- System Type: 64-bit operating system, x64-based processor
- Version: 22H2
- OS Build: 19045.3803

- Experience: Windows Feature Experience Pack 1000.19053.1000.0

Other operating systems requirements for Python software are as follows:

- Windows 7 or 10
- Mac OS X 10.11 or higher, 64-bit

First, although the latest version is Python 3.12.1, released on December 08,

2023, the Python 3.12.1 has already the required libraries, such as:

- Numpy = Performs efficient numerical computations and manipulates large multi-dimensional arrays and matrices.
- Matplotlib = Creates static, animated, and interactive visualizations for data analysis.
- Pandas = Manipulates and analyzes structured data in tabular form.
- Sklearn = Implements machine learning algorithms and tools for data mining and analysis.

3.1.2 Hardware

To run Python software, the following are the hardware requirements used in the implementation of the program:

- Processor: Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20 GHz
- Installed RAM: 8.00 GB

3.1.3 Data

The data contain historical daily stock prices and trading volumes for "Amazon.com, Inc. Common Stock (AMZN) Historical Quotes" over a two-week period from April 2, 2024, to April 28, 2014 with a file size of 131KB. The table includes details such as the date, closing price, trading volume, opening price, highest price, and lowest price for each trading day. This comprehensive set of data points allows for a thorough analysis of Amazon's stock performance during the given time frame using 2,517 rows of datasets. The "Close/Last Price" column represents the adjusted consolidated closing price, providing a reliable indicator of the stock's value at the end of the trading session. The "Volume" column displays the number of shares traded, reflecting market activity and investor interest. The "Open," "High," and "Low" columns offer insights into the stock's volatility and price ranges throughout the trading day. This data, sourced from the renowned financial data provider Edgar Online, can be used for stock market analysis, technical analysis, and studying Amazon's stock performance trends.

3.2 Methods

In crafting this methodology, the integration of feature selection, hyperparameter tuning, and model training for SVM forms a robust hybrid approach aimed at elevating overall model performance.

3.2.1 Data Gathering

The process of data gathering plays a crucial role in ensuring the accuracy and reliability of the analysis. For this study, the historical stock data for Amazon.com, Inc.

(AMZN) was obtained from reputable sources, specifically Edgar Online, a subsidiary of OTC Markets Group. The data gathering process involved accessing the online platforms or databases provided by these sources, specifying the desired stock symbol (AMZN), and retrieving the relevant information within the specified time frame of up to 10 years. Rigorous data validation techniques were employed to verify the integrity and consistency of the collected data, ensuring that any potential discrepancies or errors were identified and addressed before proceeding with the analysis. The gathered data encompassed essential elements such as daily closing prices, trading volumes, opening prices, highest and lowest prices, enabling a comprehensive examination of Amazon's stock performance (Jing et al,. 2014.

3.2.2 Data Preprocessing

To prepare the AMZN historical stock data for model building, several preprocessing steps are applied based on best practices for processing financial time series data. Handling missing values in the data features, such as closing prices or trading volumes, is done by imputing appropriate techniques like forward/backward filling or interpolation to maintain the temporal integrity of the data. If any categorical features are present (e.g., trading day indicators), they are label encoded to numerical representations. Numerical features like closing prices and trading volumes are standardized or normalized using techniques like min-max scaling or z-score normalization to ensure consistent scales across features. Additional date-based features like day of the week, month, or quarter can be extracted from the date column to capture potential cyclical patterns or seasonality effects. Relevant subsets of the data can be created by filtering for specific date ranges of interest or sampling techniques like sliding windows can be employed to prepare the data for time series forecasting or analysis tasks.To prepare the AMZN historical stock data for model building, several preprocessing steps are applied based on best practices for processing financial time series data.

3.2.3 Train-Test Splitting

After preparing the data, it is critical to divide the dataset into training and testing groups. This stage is critical for assessing the machine learning model's performance on new data and assuring its generalizability. In this study, the widely known best practice of a 90:10 train/test split ratio is used (Géron, 2019).

The 90:10 split serves two purposes. First, it allows the model to be trained on a sufficiently enough amount of the data (90%), successfully capturing the underlying patterns and correlations in the stock data. Second, it sets aside a suitable amount of the data (10%) to test and evaluate the model's performance on previously unseen instances, resulting in a reliable assessment of the model's real-world accuracy.

By adhering to this split ratio, a balance is struck between maximizing the amount of data available for training while ensuring a representative and statistically significant test set for model evaluation. This approach helps mitigate overfitting, where the model performs well on the training data but fails to generalize to new, unseen instances (Brownlee, 2020).

3.2.4 Model Selection

Selecting the appropriate machine learning model is a critical step in developing an effective stock price prediction system using the AMZN historical stock data (Patel et al., 2015). The chosen model should be capable of capturing the intricate patterns and relationships present in the dataset. A thorough understanding of the data's characteristics and the problem's requirements is essential for making an informed decision (Brownlee, 2018). The model selection process should be guided by empirical evidence and established best practices in the field of financial time series forecasting (Arora & Vamvoudakis, 2021).

Several factors must be considered when selecting the most suitable model for the task. The model's ability to handle non-linear relationships and temporal dependencies is crucial for accurate stock price prediction (Qiu et al., 2020). Additionally, the model's interpretability and computational efficiency should be evaluated to ensure practical applicability and scalability (Kaastra & Boyd, 2016). The trade-off between model complexity and performance should also be carefully assessed to avoid overfitting or underfitting issues (Géron, 2019).

Prior research in the domain of stock price forecasting can provide valuable insights and guidance for model selection. Reviewing relevant literature and analyzing the performance of various models on similar datasets can help identify promising candidates (Arora & Vamvoudakis, 2021). However, it is essential to validate the chosen model's performance on the specific AMZN dataset through rigorous evaluation and testing procedures (Kuhn & Johnson, 2013).

While a single model may be selected initially, it is often beneficial to consider alternative models or ensemble techniques for further exploration (Brownlee, 2016). A comparative analysis of multiple models can reveal their respective strengths and weaknesses, leading to a more robust and reliable prediction system (Arora & Vamvoudakis, 2021). Additionally, the model selection process should be iterative, allowing for refinements and adjustments based on the initial results and evolving requirements of the project (Kaastra & Boyd, 1996).

3.2.5 Feature Selection

Feature selection, in the context of machine learning and data analysis, refers to the process of choosing a subset of relevant features (variables, attributes) from the original set of features to use in building a predictive model or performing data analysis. In this study, two types of feature selection were applied to choose the most relevant features among the data and enhance model performance. Two feature selection methods were used:

Variance threshold: A simple basic approach to feature selection is the variance threshold. This excludes all features of low variance, i.e., all features whose variance does not exceed the threshold. It eliminates all zero-variance characteristics by default, i.e., characteristics that have the same value in all samples. This feature selection algorithm looks only at the (X) features, not the (y) outputs needed, and can, therefore, be used for unsupervised learning.

Correlation coefficient: The correlation coefficient measures the degree of the statistical linear relationship between two numerical variables. The most known measure of dependence is the Pearson's correlation coefficient. The value of a correlation coefficient could be any value between −1 and 1, a perfect negative linear relationship, and a perfect positive linear relationship, respectively. A coefficient close to 0 means that the two variables are not linearly correlated. A coefficient matrix, therefore, which is a matrix that shows the correlation coefficients among numerical variables, allows us to detect multicollinearity.

Recursive Feature Elimination (RFE) is a feature selection technique used in machine learning to select the most relevant features for a given problem. It works by recursively removing the least important features from the dataset and building a model with the remaining features. The process is repeated until the desired number of features is reached or the model performance stops improving.

The RFE algorithm typically involves the following steps:

1. Train a machine learning model on the entire feature set.
2. Compute the importance or weight of each feature based on the trained model.
3. Remove the least important features from the dataset
4. Repeat steps 1-3 with the reduced feature set until the desired number of features is reached or the model performance stops improving.

The feature importance scores can be obtained from various methods, such as coefficient values in linear models, feature importances from tree-based models, or using techniques like recursive feature elimination with cross-validation (RFECV).

RFE is particularly useful when dealing with high-dimensional datasets, where many features are present, and it is crucial to identify and retain only the most relevant ones. By removing irrelevant or redundant features, RFE can improve model performance, reduce overfitting, and enhance interpretability (R. Chen et al., 2020).

The correlation coefficient matrix can be used in conjunction with RFE to help identify and remove highly correlated features, as they may not provide additional information and could potentially cause issues like multicollinearity in the model (Salmerón-Gómez et al., 2020).

3.2.6 Model Optimization

To optimize the performance of the Support Vector Machine (SVM) model with Correlation coefficient filter and recursive feature elimination for stock price prediction using the AMZN historical data, hyperparameter tuning was employed using grid search cross-validation (Géron, 2019). Grid search evaluates the model's performance over a manually specified subset of the hyperparameter space by training and evaluating the model with all possible combinations of the specified hyperparameters (Kuhn & Johnson, 2013). This approach enables identifying the optimal hyperparameters that reduce model error and variance by systematically searching for and testing the impact of different hyperparameter values (Arora & Vamvoudakis, 2021).

For the SVM model, the kernel type (radial basis function), regularization parameter C, and gamma were tuned during the grid search process (Pedregosa et al., 2011). The scikit-learn library, a widely adopted Python machine learning library, was leveraged to efficiently implement the grid search for the SVM model (Pedregosa et al., 2011). By exhaustively tuning the model's hyperparameters to optimize performance on a validation set, the intent is to improve the model's generalizability and effectiveness on the unseen test set (Brownlee, 2016).

The grid search process typically involves splitting the dataset into training and validation subsets, where the training subset is used to train the model with different hyperparameter combinations, and the validation subset is used to evaluate the model's performance (Géron, 2019). Various performance metrics, such as mean squared error or root mean squared error, can be used to assess the model's predictive capabilities and guide the selection of the optimal hyperparameters (Kuhn & Johnson, 2013).

After identifying the optimal hyperparameters through the grid search process, the SVM model can be retrained using the entire training dataset and the selected hyperparameters (Arora & Vamvoudakis, 2021). This final optimized model can then be evaluated on the held-out test set to obtain an unbiased estimate of its performance on unseen data (Brownlee, 2016). If the performance is satisfactory, the optimized SVM model can be deployed for stock price prediction and analysis tasks; otherwise, further refinements or alternative models may be considered (Géron, 2019).

3.2.7 Evaluation of Result

A robust evaluation of the optimized models will be undertaken using several standard evaluation metrics to quantify performance from multiple perspectives. Model performance will be thoroughly evaluated using key classification metrics - accuracy, precision, recall, and F1score - which analyze different facets of prediction quality based on the confusion matrix. Here are the precise formulas. Accuracy is the ratio of the total number of correct predictions made by the model over all the predictions made. It measures how often the model is correct in classifying samples.

$$\text{Accuracy} = \frac{TP + TN}{P+N} \qquad (2)$$

Where:

TP - True Positives
TN – True Negatives
P – Positives
N – Negatives

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (3)$$

Where:

TP - True Positives
TN – True Negatives
FP – False Positives

Recall also known as sensitivity. The ratio of correctly predicted positive examples out of all the actual positive examples. It measures the model's ability to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (4)$$

Where:

TP - True Positives

FN – False Negatives

F1-Score is a combined metric that considers both precision and recall by taking their harmonic mean. It provides a balance between precision and recall. The best F1 score is 1, representing perfect precision and recall, and the worst is 0.

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \qquad (5)$$

3.2.8 Receiver Operating Characteristic (ROC) Curve

The ROC curve is a plot that illustrates the diagnostic ability of a binary classifier by showing the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various classification thresholds. The area under the ROC curve (AUC-ROC) is a single scalar value that summarizes the overall performance of the classifier, with a higher value indicating better performance.

An AUC-ROC of 0.84 for an SVM model that has not been tuned specifically for the ROC curve indicates moderately good performance. The value ranges from 0 to 1, where 0.5 represents a classifier that performs no better than random guessing, and 1.0 represents a perfect classifier.

In general, an AUC-ROC value of:

a) 0.5 to 0.6 is considered poor performance

b) 0.6 to 0.7 is considered average performance

c) 0.7 to 0.8 is considered good performance

d) 0.8 to 0.9 is considered very good performance

e) 0.9 to 1.0 is considered excellent performance (Narkhede, 2022).

True Positive Rate (TPR) is defined as:

$$TPR = \frac{TP}{TP + FN} \qquad (6)$$

Where:

TP- True Positive
FN – False Negatives

False Positive Rate (FPR) is defined as:

$$FPR = \frac{FP}{FP + TN} \qquad (7)$$

Where:

FP- False Positive
TN – True Negatives

AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve:

The AUC represents the overall performance of a classifier. It is the integral of the ROC curve:

$$AUC = \int_0^1 TPR(FPR)dFPR \qquad (8)$$

Where:

∫: The integral symbol, indicating that the area is calculated by integrating over the range of the false positive rate (FPR).

TPR(FPR): The True Positive Rate (TPR) expressed as a function of the False Positive Rate (FPR). The ROC curve plots TPR against FPR, and the AUC measures the area under this curve.

dFPR: The differential of the False Positive Rate

3.2.9 Confusion Matrix

The confusion matrix provides the count of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). This shows how well the model correctly identifies the positive and negative classes.

The confusion matrix summarizes the predictions made by a classifier. It is typically represented as:

$$ConfMatrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \qquad (9)$$

Where:

TP - True Positives
TN – True Negatives
FP – False Positives
FN –False Negative

In conclusion, a robust evaluation of the optimized models will involve a comprehensive analysis using several standard evaluation metrics. Key metrics such as accuracy, precision, recall, and F1-score will quantify model performance from multiple angles based on the confusion matrix values. The Receiver Operating Characteristic (ROC) curve and its associated Area Under the Curve (AUC) metric will provide an assessment of the diagnostic ability of the binary classifiers across various classification thresholds. The confusion matrix itself will offer a straightforward summary of the models' predictions, highlighting the count of true positives, true negatives, false positives, and false negatives. Collectively, this suite of evaluation techniques will enable a thorough, multi-faceted evaluation of the optimized models, revealing their strengths, weaknesses, and overall capability to accurately classify samples. Such rigorous evaluation is essential for selecting the most performant models and guiding further optimization efforts.

4 RESULTS AND DISCUSSIONS

The following tables and figures present the key results and analysis of the proposed predictive modeling methodology using SVMs with optimized feature selection. The impact of techniques like grid search hyperparameter tuning, and correlation filters on model performance is evaluated through several experiments. The predictive accuracy, precision, recall, F1 score, ROC curve, and confusion matrix are reported for each model configuration.

4.1 Stock Market Price Prediction Model Performance Evaluation

4.1.1 Evaluate the performance of a standard SVM model as a baseline

Table 1. Performance of Standard Support Vector Machine (SVM) model

Metrics	Results
Accuracy	0.70
Precision	0.97
Recall	0.45
F1 Score	0.61

The performance metrics presented in Table 1 offer valuable insights into the behavior of the standard SVM model. The overall accuracy of 0.70 indicates that the model correctly classified 70% of the examples, caution is warranted due to the potential limitations of accuracy, especially in scenarios involving class imbalance or uneven error costs (Chidambaram & Srinivasagan, 2018). The high precision of 0.97 is a positive aspect, signifying a low false positive rate and high correctness when predicting positive examples. Conversely, the low recall of 0.45 highlights the model's struggle in identifying a significant portion of positive instances, leading to a notable rate of false negatives. The F1 score, balancing precision and recall, stands at 0.61, suggesting a moderate overall performance with room for improvement, particularly in enhancing recall without compromising precision significantly (Azar & El-Said, 2013).

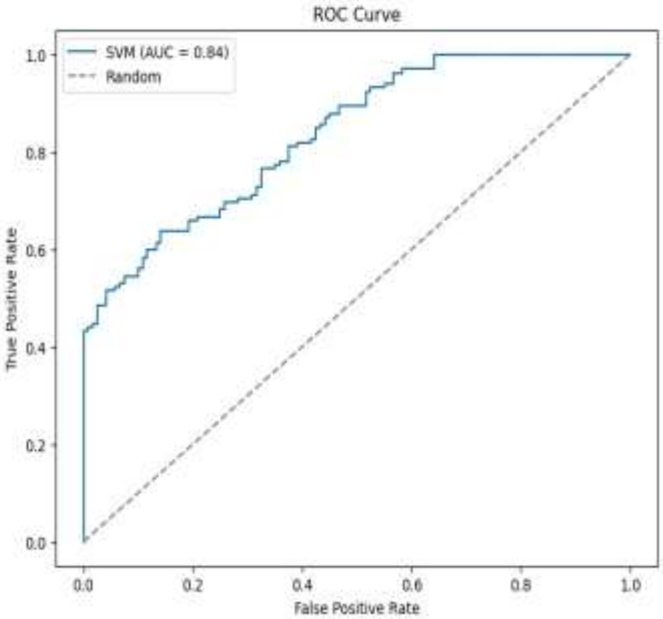


Figure 2. Support Vector Machine SVM Receiver Operating Characteristic (ROC) curve

Figure 2 shows an AUC-ROC value of 0.84 for an untuned SVM model indicates very good performance in distinguishing between positive and negative cases. This value falls within the range of 0.8 to 0.9, which is generally considered "very good" classification performance (Allwright, 2022).

An AUC-ROC of 0.84 means that if one positive and one negative example are randomly selected, the untuned SVM model will correctly identify which example belongs to which class 84% of the time. In other words, the untuned SVM model has an 84% chance of ranking a randomly chosen positive example higher than a randomly chosen negative example (Armaghani et al., 2020).

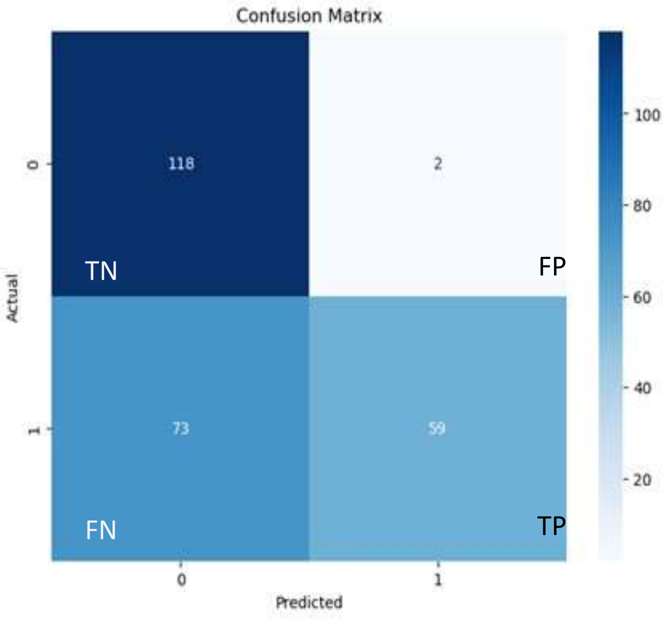


Figure 3. Support Vector Machine SVM Confusion Matrix

Figure 3 shows the model exhibits challenges in accurately classifying the positive class, evident from a substantial number of false negatives (73) while performing well in identifying negative examples with 118 true negatives. Although the false positive rate is low (2), indicating a commendable ability to avoid misclassifying negatives as positives, the model's propensity to predict the negative class more frequently is reflected in the lower number of true positives (59). This discrepancy suggests potential bias towards the negative class, possibly influenced by class imbalance in the training data or other factors. The overall performance, particularly in accurately identifying positive examples, is suboptimal, emphasizing the need for further enhancements before practical deployment, taking into account the specific application requirements and the trade-off between minimizing false positives and false negatives.

4.1.2 Performance of the Support Vector Machine (SVM) using a hybrid feature selection combining correlation coefficient filtering and recursive feature elimination

Table 2. Performance of Support Vector Machine (SVM) with correlation coefficient filter and Recursive Feature Elimination

Metrics	Results
Accuracy	0.58
Precision	0.83
Recall	0.26
F1 Score	0.39

Table 2 presents the performance metrics of a Support Vector Machine (SVM) classifier after applying correlation coefficient filtering and recursive feature elimination (RFE) for feature selection. The accuracy of 0.58 indicates that the model correctly classified only 58% of instances, which is relatively low (Armaghani et al., 2020). While the precision of 0.83 suggests the model is quite reliable in identifying positive instances as positive, the low recall of 0.26 raises concerns. This low recall means the model missed a significant portion (75%) of the actual positive instances, classifying them as negative instead (Smolic, 2024).

The F1 score of 0.39, which combines precision and recall, further highlights the model's struggle in effectively separating the positive and negative classes. These results could be attributed to several factors, such as class imbalance in the dataset, the presence of noisy or irrelevant features, or the inability of the feature selection techniques to capture the most discriminative features for the positive class (Jayaswal, 2021).

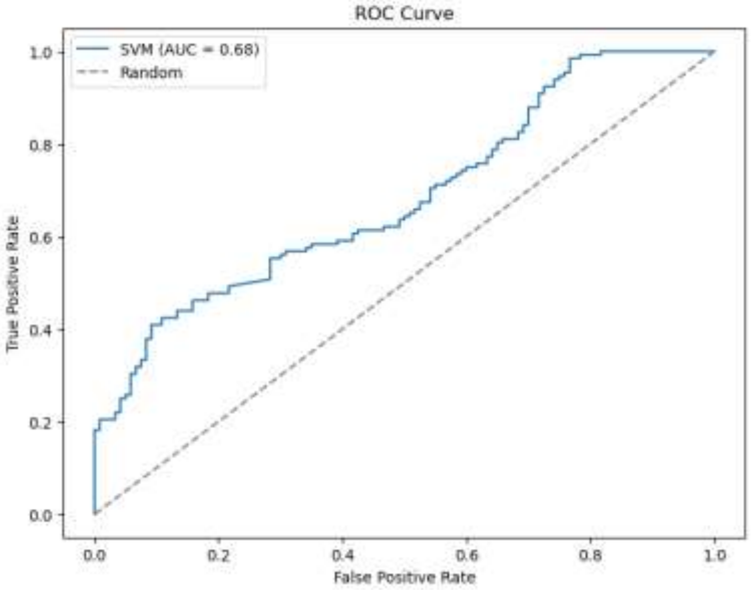


Figure 4. Support Vector Machine SVM with correlation coefficient filter Recursive Feature Elimination Receiver Operating Characteristic (ROC) curve

Figure 4 shows that when dealing with high-dimensional data, it is common to encounter redundant or irrelevant features that can adversely impact machine learning model performance. The resulting model achieved an Area Under the Receiver Operating Characteristic (ROC) Curve of 0.68, indicating fair-to-good ability to distinguish classes. However, an AUC-ROC of 0.68 suggests room for improvement, as values above 0.8 are considered excellent. Further tuning of hyperparameters, exploring alternative techniques, or investigating different algorithms may be warranted to enhance model performance (Bobbitt, 2021).

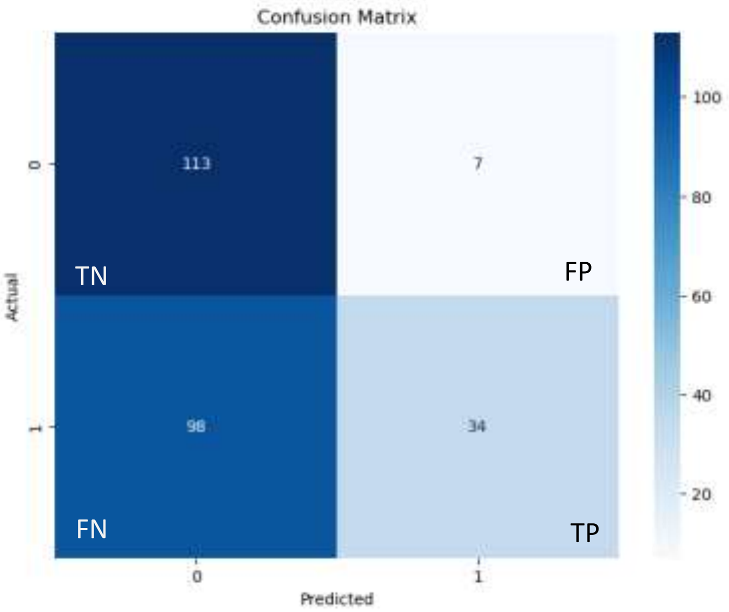


Figure 5. Support Vector Machine SVM with correlation coefficient filter Recursive Feature Elimination Confusion Matrix

When evaluating the performance of a classification model, it's crucial to understand the true nature of its predictions. The confusion matrix offers a comprehensive view of the model's behavior, shedding light on its strengths and weaknesses. In this case, the matrix reveals a concerning trend – while the model excels at identifying positive instances, its ability to accurately classify negative instances leaves much to be desired (Nath, 2023).

Figure 5 shows that with 113 true positives and a mere 34 true negatives, the model's prowess in recognizing positive cases is undeniable. However, the staggering number of 98 false negatives raises concerns about its reliability in real-world scenarios. Failing to detect nearly half of the negative instances could have severe consequences, depending on the application domain. Furthermore, the 7 false positives, though relatively low, contribute to the overall inaccuracy of the model, potentially leading to unnecessary actions or resource allocation. These shortcomings underscore the need for further refinement, whether through additional training data, feature engineering, or exploring alternative modeling techniques. Ultimately, the goal should be to strike a balance between accurately identifying both positive and negative instances, ensuring the model's utility and trustworthiness in practical applications.

4.1.3 Optimize the hybrid feature selection methodology using grid search hyperparameter tuning

Table 4. Performance of Optimize the developed hybrid feature selection methodology through Grid search

Metrics	Results
Accuracy	0.90
Precision	0.99
Recall	0.81
F1 Score	0.89

Table 4 presents the performance metrics of a classification model after optimizing the developed hybrid feature selection methodology through grid search. The accuracy of 0.90 indicates that the model correctly classified 90% of the instances, which is a relatively high score (Seraydarian, 2024). The precision of 0.99 suggests that the model is highly reliable in identifying positive instances as positive, with very few false positives. However, the recall of 0.81 implies that the model missed around 19% of the actual positive instances, classifying them as negative.

The F1 score of 0.89 reflects a good balance between precision and recall, indicating that the model is performing well in terms of identifying both positive and negative instances. These results suggest that the hybrid feature selection methodology, coupled with the grid search optimization, has effectively identified a relevant subset of features that capture the most discriminative information for the classification task. The high precision and overall accuracy demonstrate the model's robustness and reliability (Seraydarian, 2024).

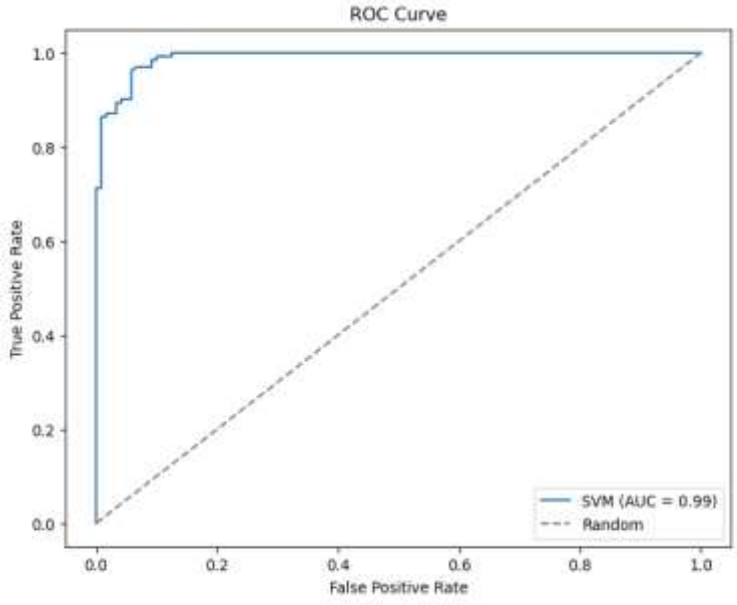


Figure 6. Optimized Hybrid Feature Selection using Grid Search Receiver Operating Characteristic (ROC) curve

A ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The area under the ROC curve (AUC-ROC) is a measure of the model's ability to distinguish between positive and negative classes. Figure 6 shows an AUC-ROC of 0.99 indicates that the model has excellent discriminative power and can accurately differentiate between the two classes with very high precision. Typically, an AUC-ROC score of 0.5 represents a worthless model, while a score of 1.0 represents a perfect model. So, an AUC-ROC of 0.99 is extremely close to the ideal value, suggesting that the model is performing exceptionally well in separating the two classes (Janssens & Martens, 2020).

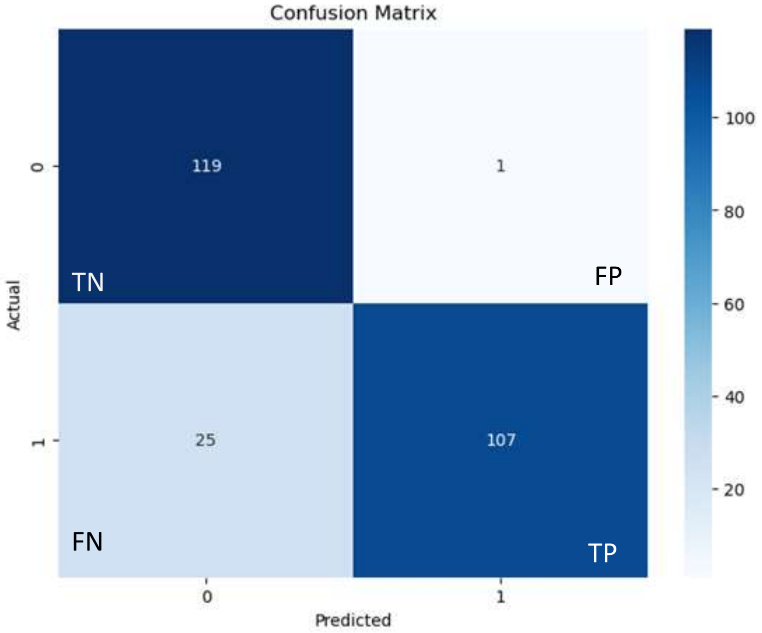


Figure 7. Optimized Hybrid Feature Selection using Grid Search Confusion Matrix

When examining the performance of a classification model, the confusion matrix unveils a wealth of insights that go beyond mere accuracy metrics. In this instance, the matrix paints a picture of a model that exhibits remarkable proficiency in recognizing positive instances while maintaining a commendable ability to accurately classify negative cases.

Figure 7 shows that with 119 true positives and a mere 1 false positive, the model demonstrates an exceptional grasp of the positive class, minimizing the risk of unnecessary actions or resource allocation. This level of precision is highly desirable, especially in domains where false positives could prove costly or disruptive. However, the 25 false negatives indicate that the model still struggles with a subset of positive instances, failing to identify them correctly. While this figure is relatively low compared to the true positives, it highlights an area for potential improvement. On the positive side, the 107 true negatives showcase the model's competence in accurately recognizing negative cases, a crucial aspect that often gets overshadowed by the emphasis on positive class performance. This balanced performance across both classes is a testament to the model's overall robustness and reliability, making it a strong contender for real-world deployment.

5 SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Summary

Through the use of optimized feature selection techniques and Support Vector Machines (SVMs), the research presents a novel approach to enhance the performance the performance of the Support Vector Machine (SVM) by using Hybrid Feature Selection approach combining Correlation Coefficient Filter (CCF) and Recursive Feature Elimination (RFE). The main objective is to create a predictive model that is reliable and accurate while minimizing the frequent problems of overfitting and the addition of superfluous or irrelevant variables, which can negatively affect model performance.

The suggested methodology consists of a number of methodically crafted processes, the first of which is preprocessing the historical stock data in order to properly scale features and handle missing values. After that, the data is divided into training and testing sets so that the performance of the model on untested data can be thoroughly assessed. To set a performance benchmark, a baseline SVM model with default parameters is first assessed.

Subsequently, the study employs a hybrid feature selection approach that combines correlation coefficient filtering and recursive feature elimination. This innovative technique aims to identify the most relevant and informative features for stock price prediction, thereby reducing the risk of overfitting and improving model generalization. Finally, grid search is utilized to optimize the feature selection methodology and tune the SVM hyperparameters, further enhancing the model's predictive capabilities.

5.2 Conclusion

The baseline SVM model with default parameters exhibited moderate performance, achieving an accuracy of 0.70 and an AUC-ROC of 0.84. However, the low recall of 0.45 and a substantial number of false negatives indicated room for improvement. The initial application of the hybrid feature selection approach led to a decline in performance, with an accuracy of 0.58 and an AUC-ROC of 0.68, highlighting challenges in effectively separating the positive and negative classes.

Remarkably, the incorporation of grid search hyperparameter tuning to optimize the hybrid feature selection methodology yielded significant performance gains. The optimized model achieved an impressive accuracy of 0.90, a precision of 0.99, a recall of 0.81, and an F1 score of 0.89. The AUC-ROC of 0.99 further demonstrated the model's exceptional ability to discriminate between classes. The confusion matrix analysis reinforced the optimized model's strengths, with 119 true positives, 107 true negatives, and only 1 false positive, showcasing its balanced performance across both classes. Despite 25 false negatives, the overall results highlighted the model's robustness and reliability for stock price prediction tasks.

Therefore, the optimization using grid search hyperparameter tuning proved to be an effective strategy for enhancing the performance of the hybrid feature selection methodology. By systematically exploring the parameter space, the optimized model achieved remarkable improvements in accuracy, precision, recall, F1 score, and AUC-ROC. The confusion matrix analysis further verified the model's balanced and reliable performance across both classes. In conclusion, the successful application of grid search optimization highlights the potential of this approach for stock price prediction tasks and underscores the importance of careful hyperparameter tuning in machine learning models.

5.3 Recommendations

1. Implementation of Hybrid Feature Selection: Financial institutions and data scientists should consider implementing the hybrid feature selection method proposed in this study. By combining correlation coefficient filtering and recursive feature elimination, they can enhance the performance and reliability of their stock price prediction models.
2. Continuous Evaluation and Adaptation: The stock market is dynamic and ever-changing. Therefore, continuous evaluation and adaptation of the predictive model are crucial. Regularly updating the model with new data and re-optimizing parameters can help maintain its accuracy and relevance over time.
3. Application to Other Domains: While this study focuses on stock price prediction, the proposed hybrid feature selection and optimization methods can be applied to other domains with high dimensionality and noise. Researchers and practitioners in various fields should explore the potential of these techniques to improve the performance of their predictive models.
4. Educational Integration: Academic programs in data science and finance could integrate the findings and methodologies from this study into their curricula. This will equip students with advanced techniques in feature selection and model optimization, preparing them for real-world applications in financial forecasting and beyond. By adopting these recommendations, stakeholders can leverage advanced machine learning techniques to

enhance their predictive modeling efforts, ultimately leading to more informed and strategic decision-making processes.

References:

Al-Rajab, M., Lu, J., Xu, Q., Kentour, M., Sawsa, A., Shuweikeh, E., Joy, M., & Arasaradnam, R. P. (2023). A hybrid machine learning feature selection model—HMLFSM to enhance gene classification applied to multiple colon cancers dataset. PloS One, 18(11), e0286791. https://doi.org/10.1371/journal.pone.0286791

Al-Tashi, Q., Saad, M., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A., Le, X., Vokes, N. I., Zhang, J., & Wu, J. (2023). Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic review. International Journal of Molecular Sciences, 24(9), 7781. https://doi.org/10.3390/ijms24097781

Alam, M. (2023, December 5). What is Continuous Improvement? Definition, Model, Methodology, Process and Examples. IdeaScale. https://ideascale.com/blog/what-is-continuous-improvement/

Ali, M., Ali, S. I., Kim, D., Hur, T., Bang, J. H., Lee, S., Kang, B. H., & Hussain, M. (2018). uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features. PloS One, 13(8), e0202705. https://doi.org/10.1371/journal.pone.0202705

Alizadeh, R., Allen, J. K., & Mistree, F. (2020). Managing computational complexity using surrogate models: a critical review. Research in Engineering Design, 31(3), 275–298. https://doi.org/10.1007/s00163-020-00336-7

Allwright, S. (2022, December 6). How to interpret AUC score (simply explained). Stephen Allwright. https://stephenallwright.com/interpret-auc-score/

Amazon.com, Inc. Common Stock (AMZN) Historical Data | Nasdaq. (n.d.-b). Nasdaq. https://www.nasdaq.com/market-activity/stocks/amzn/historical?page=1&rows_per_page=10&timeline=m1

Armaghani, D. J., Asteris, P. G., Askarian, B., Hasanipanah, M., Tarinejad, R., & Van Huynh, V. (2020a). Examining Hybrid and Single SVM Models with Different Kernels to Predict Rock Brittleness. Sustainability, 12(6), 2229. https://doi.org/10.3390/su12062229

Arora, M., & Vamvoudakis, K. G. (2021). Machine learning for stock price forecasting and portfolio management. In Handbook of reinforcement learning and control (pp. 377-407). Springer, Cham. https://doi.org/10.1007/978-3-030-60936-8_15

Artzi, I. (2022). Predictive Analytics Techniques: Theory and Applications in finance. In Contributions to finance and accounting (pp. 59–126). https://doi.org/10.1007/978-3-030-83799-0_3

Avolio, M., & Fuduli, A. (2023). The semiproximal SVM approach for multiple instance learning: a kernel-based computational study. Optimization Letters, 18(2), 635–649. https://doi.org/10.1007/s11590-023-02022-8

Awotunde, J. B., Ogundele, L. A., Olakunle, S., Awotunde, J. B., & Kasali, F. A. (2024). A hybrid correlation-based deep learning model for email spam classification using fuzzy inference system. Decision Analytics Journal, 10, 100390. https://doi.org/10.1016/j.dajour.2023.100390

Awotunde, J. B., Panigrahi, R., Khandelwal, B., Garg, A., & Bhoi, A. K. (2023). Breast cancer diagnosis based on hybrid rule-based feature selection with deep learning algorithm. Research on Biomedical Engineering, 39(1), 115–127. https://doi.org/10.1007/s42600-022-00255-7

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In Apress eBooks (pp. 39–66). https://doi.org/10.1007/978-1-4302-5990-9_3

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In Apress eBooks (pp. 39–66). https://doi.org/10.1007/978-1-4302-5990-9_3

Azar, A. T., & El-Said, S. A. (2013). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Computing & Applications, 24(5), 1163–1177. https://doi.org/10.1007/s00521-012-1324-4

B, H. N. (2021, December 12). Confusion matrix, accuracy, precision, recall, F1 score. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Bendouch, M. M., Frăsincar, F., & Robal, T. (2022). Enhancing Semantics-Driven Recommender Systems with Visual Features. In Lecture notes in computer science (pp. 443–459). https://doi.org/10.1007/978-3-031-07472-1_26

Bian, K., & Priyadarshi, R. (2024). Machine Learning Optimization Techniques: A survey, classification, challenges, and future research issues. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-024-10110-w

Biau, G., & Cadre, B. (2021). Optimization by gradient boosting. In Springer eBooks (pp. 23–44). https://doi.org/10.1007/978-3-030-73249-3_2

Bisori, R., Lapucci, M., & Sciandrone, M. (2021). A study on sequential minimal optimization methods for standard quadratic problems. 4OR, 20(4), 685–712. https://doi.org/10.1007/s10288-021-00496-9

Blanco, N., & Blanco, N. (2024, January 16). How to use machine learning to predict stock prices | Robots.net. Robots.net. https://robots.net/fintech/how-to-use-machine-learning-to-predict-stock-prices/

Bobbitt, Z. (2021, September 9). What is Considered a Good AUC Score? Statology. https://www.statology.org/what-is-a-good-auc-score/

Bolón-Canedo, V., Alonso-Betanzos, A., Morán-Fernández, L., & Cancela, B. (2022). Feature selection: From the past to the future. In Learning and analytics in intelligent systems (pp. 11–34). https://doi.org/10.1007/978-3-030-93052-3_2

Brownlee, J. (2016). Machine Learning Mastery With Python. Machine Learning Mastery.

Brownlee, J. (2018). Introduction to Time Series Forecasting with Python. Machine Learning Mastery.

Brownlee, J. (2019). Overfitting and underfitting with machine learning algorithms. MachineLearningMastery.com. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Brownlee, J. (2020). Train/Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

C. C. Chang, C.-J. Lin., LIBSVM, a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2:27: 1-27:27, (2011). (n.d.). https://www.sciepub.com/reference/138080

Chakrabarti, S., Roy, S., & Soundalgekar, M. V. (2018). Fast and accurate text classification via multiple linear discriminant projections. ˜the œVLDB Journal, 12(2), 170–185. https://doi.org/10.1007/s00778-003-0098-9

Chen, J. (2023, March 11). Stock analysis: Different methods for evaluating stocks. Investopedia. https://www.investopedia.com/terms/s/stock-analysis.asp

Chen, J. (2024, January 29). What is the stock market, what does it do, and how does it work? Investopedia. https://www.investopedia.com/terms/s/stockmarket.asp

Chen, L. (2021, December 7). Basic ensemble learning (Random Forest, AdaBoost, Gradient boosting)- step by step explained. Medium. https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

Chen, R., Dewi, C., Huang, S., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Chinnaswamy, A., & Ramakrishnan, S. (2015). Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In Advances in intelligent systems and computing (Internet) (pp. 229–239). https://doi.org/10.1007/978-3-319-28031-8_20

Choi, H. K. (2018, August 5). Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model. arXiv.org. https://arxiv.org/abs/1808.01560

Costa, J. a. F., Dantas, N. C. D., & Silva, E. (2023). Evaluating text classification in the legal domain using BERT embeddings. In Lecture notes in computer science (pp. 51–63). https://doi.org/10.1007/978-3-031-48232-8_6

Dayananda, S. (2023, November 6). Support Vector Machines (SVM) - Sandun Dayananda – Medium. Medium. https://sandundayananda.medium.com/support-vector-machines-svm-db8314e9092d

Dhillon, A., Singh, A., & Bhalla, V. K. (2022). A Systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: From computational needs to machine learning and Deep learning. Archives of Computational Methods in Engineering, 30(2), 917–949. https://doi.org/10.1007/s11831-022-09821-9

Dias, J. M. L., & António, N. (2023). Predicting customer churn using machine learning: A case study in the software industry. Journal of Marketing Analytics (Print). https://doi.org/10.1057/s41270-023-00269-9

Dunn, J., Mingardi, L., & Zhuo, Y. D. (2021, May 11). Comparing interpretability and explainability for feature selection. arXiv.org. https://arxiv.org/abs/2105.05328

Elghazel, H., & Aussem, A. (2013). Unsupervised feature selection with ensemble learning. Journal of Machine Learning Research, 14(1), 2349-2368. https://link.springer.com/content/pdf/10.1007/s10994-013-5337-8.pdf

Fang, C., Zhang, Y., & Pan, S. J. (2023). Transfer feature selection for predictive modeling. Pattern Recognition Letters, 167, 84-92. https://doi.org/10.1016/j.patrec.2022.09.007

Galli, S. (2024, March 11). Feature Selection with Wrapper Methods in Python. Train in Data Blog. https://www.blog.trainindata.com/feature-selection-with-wrapper-methods/

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

Gündüz, H. (2021). An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. Financial Innovation, 7(1). https://doi.org/10.1186/s40854-021-00243-3

Hailu, T. G., & Abdulkadir, T. (2023). MULTIDMET: Designing a hybrid multidimensional metrics framework to predictive modeling for performance evaluation and feature selection. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-3111777/v1

Hakan, Pabuçcu., A., Barbu. (2023). Feature Selection for Forecasting. arXiv.org, doi: 10.48550/arXiv.2303.02223

Hardin, H. (2022, March 30). Model selection, tuning and evaluation in K-Nearest neighbors. Medium. https://towardsdatascience.com/model-selection-tuning-and-evaluation-in-k-nearest-neighbors-6d3024d78745

Harshith, H, S. (2023). Stock Market Analysis and Prediction using Machine Learning Algorithm. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2023.54812

He, Z., Li, L., Huang, Z., & Situ, H. (2018). Quantum-enhanced feature selection with forward selection and backward elimination. Quantum Information Processing, 17(7). https://doi.org/10.1007/s11128-018-1924-8

Hoque, K. E. (2021). Impact of hyperparameter tuning on machine learning models in stock price forecasting.www.academia.edu. https://www.academia.edu/92309081/Impact_of_Hyperparameter_Tuning_on_Machine_Learning_Models_in_Stock_Price_Forecasting

Htun, H. H., Biehl, M., & Petkov, N. (2023b). Survey of feature selection and extraction techniques for stock market prediction. Financial Innovation, 9(1). https://doi.org/10.1186/s40854-022-00441-7

Jelodar, H., Wang, Y., Yuan, C., Xia, F., Jiang, X., Li, Y., & Zhao, L. (2018). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jeon, H., & Oh, S. (2020c). Hybrid-Recursive feature elimination for efficient feature selection. Applied Sciences, 10(9), 3211. https://doi.org/10.3390/app10093211

Jing, Zhang., Xuegang, Hu., Peipei, Li., Wei, He., Yuhong, Zhang., Huizong, Li. (2014). A Hybrid Feature Selection Approach by Correlation-Based Filters and SVM-RFE. 3684-3689. https://doi:10.1109/ICPR.2014.633

Joseph, R. (2022, October 18). Grid Search for model tuning - Towards Data Science. Medium. https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

Joshi, Y. (2020, March 26). Applications of Principal Component Analysis (PCA). OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/applications-of-pca/

Juraev, G., & Bozorov, O. (2023). Using TF-IDF in text classification. AIP Conference Proceedings. https://doi.org/10.1063/5.0145520

Kaastra, I., & Boyd, M. (2016). Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10(3), 215-236. https://doi.org/10.1016/0925-2312(95)00039-9v

Khan, A., Rasheed, M. T., & Khan, H. (2023). An empirical study of deep learning-based feature extractor models for imbalanced image classification. Advances in Computational Intelligence, 3(6). https://doi.org/10.1007/s43674-023-00067-x

Kim, J., Pearl, J., & Bareinboim, E. (2022). Combining predictive modeling and causal inference for improved decision-making. Machine Learning, 111(7), 1789-1812. https://doi.org/10.1007/s10994-021-06123-x

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

KhosrowHassibi. (2016, October 28). *Machine Learning vs. Traditional Statistics: Different philosophies, Different Approaches*. Data Science Central. https://www.datasciencecentral.com/machine-learning-vs-traditional-statistics-different-philosophi-1/

Λιαγκούρας, K., & Metaxiotis, K. (2020). Stock market forecasting by using support vector machines. In Learning and analytics in intelligent systems (pp. 259–271). https://doi.org/10.1007/978-3-030-49724-8_11

Li, X., Liang, C., & Ma, F. (2022). Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model. Annals of Operation Research/Annals of Operations Research.

Li, Y., Mansmann, U., Du, S., & Hornung, R. (2022). Benchmark study of feature selection strategies for multi-omics data. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-04962-x

Liu, Z., Chen, S., & Wang, Y. (2022). Deep feature selection for time series forecasting. Pattern Recognition, 122, 108318. https://doi.org/10.1016/j.patcog.2021.108318

López, O. a. M., López, A. M., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Springer eBooks* (pp. 109–139). https://doi.org/10.1007/978-3-030-89010-0_4

Long, W., Gao, J., Bai, K., & Lu, Z. (2024). A hybrid model for stock price prediction based on multi-view heterogeneous data. *Financial Innovation*, *10*(1). https://doi.org/10.1186/s40854-023-00519-w

Mengash, H. A., Aljunid, N., Kouki, F., Singla, C., Elhameed, E. A., & Mahmud, A. (2023). Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries. Biomimetics, 9(1), 1. https://doi.org/10.3390/biomimetics9010001

Mirzaei, G., Soltani, A., Soltani, M., & Darabi, M. (2018). An integrated data-mining and multi-criteria decision-making approach for hazard-based object ranking with a focus on landslides and floods. Environmental Earth Sciences, 77(16). https://doi.org/10.1007/s12665-018-7762-2

Meng, Q., & Li, K. (2019). A rank correlation coefficient based particle filter to estimate parameters in non-linear models. International Journal of Distributed Sensor Networks, 15(4), 155014771984127. https://doi.org/10.1177/1550147719841273

Naik, N., Vikranth, B., & Yogesh, N. (2022b). Recursive feature elimination technique for technical indicators selection. In Communications in computer and information science (pp. 139–145). https://doi.org/10.1007/978-3-031-08277-1_12

Napier.tex.paper (n.d.). Researchgate.Net. https://www.researchgate.net/profile/Andrew-Napier-8/publication/320762638_Napier_A_et_al_-_SciPy_a_Pythonbased_ecosystem_for_scientific_and_numerial_computing_BIS_2018/links/5c6ee2e1299bf1bc768c8c19/Napier-A-et-al-SciPy-a-Pythonbased-ecosystem-for-scientific-and-numerial-computing-BIS-2018.pdf?origin=publication_detail

Nasir, I. M., Khan, M. A., Yasmin, M., Shah, J. H., Gabryel, M., Scherer, R., & Damaševičius, R. (2020). Pearson Correlation-Based feature selection for document classification using balanced training. Sensors, 20(23), 6793. https://doi.org/10.3390/s20236793

Navon, D., & Bronstein, A. M. (2022, August 17). Random Search Hyper-Parameter tuning: expected improvement estimation and the corresponding lower bound. arXiv.org. https://arxiv.org/abs/2208.08170

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00299-5

Özyurt, F. (2019). Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures. ˜the œJournal of Supercomputing/Journal of Supercomputing, 76(11), 8413–8431. https://doi.org/10.1007/s10994-013-5337-8.pdf

Omolara, A. E., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., & Alkhawaldeh, R. S. (2021). A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Computing & Applications, 33(22), 15091–15118. https://doi.org/10.1007/s00521-021-06406-8

Orra, A., Sahoo, K., & Choudhary, H. (2023). Machine Learning-Based Hybrid Models for trend forecasting in financial instruments. In Lecture notes in networks and systems (pp. 337–353). https://doi.org/10.1007/978-981-19-6525-8_26

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172. https://doi.org/10.1016/j.eswa.2014.10.031

Patil, R., & Kolhe, S. R. (2022). Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets. Social Network Analysis and Mining, 12(1). https://doi.org/10.1007/s13278-022-00877-w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. PloS One, 18(12), e0289724. https://doi.org/10.1371/journal.pone.0289724

Pleil, J. D., Wallace, M. a. G., Stiegel, M. A., & Funk, W. E. (2018). Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. Journal of Toxicology and Environmental Health. Part B, Critical Reviews, 21(3), 161–180. https://doi.org/10.1080/10937404.2018.1490128

Qiu, J., Wang, B., & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. PloS one, 15(1), e0227222. https://doi.org/10.1371/journal.pone.0227222

R. Aishwarya, K. Pavitra, P. V. Miranda, K. Keerthana and L. Kamatchi Priya, "Parkinson's Disease Prediction using Fisher Score based Recursive Feature Elimination," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 1-8, https://doi: 10.1109/InCACCT57535.2023.10141768 .

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., … & Zhang, M. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 1-10. 68.pdf (stanford.edu)

R. Balabhadrapathruni and S. De, "A Study on Analysing the impact of Feature Selection on Predictive Machine Learning Algorithms," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 10-15, doi: 10.1109/PDGC50313.2020.9315801.

Rickert, C. A., Henkel, M., & Lieleg, O. (2023). An efficiency-driven, correlation-based feature elimination strategy for small datasets. APL Machine Learning, 1(1). https://doi.org/10.1063/5.0118207

Rubyabdullah. (2023, October 29). Overfitting vs. Underfitting: A Practical Guide to Model Generalization. Medium. https://medium.com/@rubyabdullah14/overfitting-vs-underfitting-a-practical-guide-to-model-generalization-b560baaf4124

Saidi, R., Bouaguel, W., & Essoussi, N. (2018). Hybrid feature selection method based on the genetic algorithm and Pearson correlation coefficient. In Studies in computational intelligence (pp. 3–24). https://doi.org/10.1007/978-3-030-02357-7_1

Salmerón-Gómez, R., García-García, C., & García-Pérez, J. (2020). A guide to using the R package "MultiColl" for detecting multicollinearity. *Computational Economics*, *57*(2), 529–536. https://doi.org/10.1007/s10614-019-09967-y

Sampaio, C. (2023, July 2). *Understanding SVM hyperparameters*. Stack Abuse. https://stackabuse.com/understanding-svm-hyperparameters/

Sancar, S., & Uzun-Per, M. (2023). Testing the performance of feature selection methods for customer churn analysis: case study in B2B business. In Lecture notes in networks and systems (Online) (pp. 509–519). https://doi.org/10.1007/978-3-031-27099-4_39

Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics, 19(1). https://doi.org/10.1186/s12859-018-2451-4

Sarker, I. H. (2021). Machine learning: algorithms, Real-World applications and research directions. SN Computer Science/SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Scott, W. (2021, December 7). TF-IDF from scratch in python on a real-world dataset. Medium. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. International Journal of Financial Studies, 7(2), 26. https://doi.org/10.3390/ijfs7020026

Shah, R. (2022, August 8). Performance comparison of tuned and untuned classification models. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/performance-comparision-of-tuned-and-untuned-classification-models/

Shetty, A. M., Aljunid, M. F., Manjaiah, D. H., & Afzal, A. L. (2024). Hyperparameter optimization of machine learning models using grid search for Amazon Review sentiment analysis. In Lecture notes in networks and systems (pp. 451–474). https://doi.org/10.1007/978-981-99-7814-4_36

Shetty, A. M., Aljunid, M. F., Manjaiah, D. H., & Afzal, A. L. (2024b). Hyperparameter optimization of machine learning models using grid search for Amazon Review sentiment analysis. In Lecture notes in networks and systems (pp. 451–474). https://doi.org/10.1007/978-981-99-7814-4_36

Sidhom, O., Ghazouani, H., & Barhoumi, W. (2023). Three-phases hybrid feature selection for facial expression recognition. ˜the œJournal of Supercomputing/Journal of Supercomputing. https://doi.org/10.1007/s11227-023-05758-3

Silipo, R. (2021, December 13). Machine learning algorithms and the art of hyperparameter selection. Medium. https://towardsdatascience.com/machine-learning-algorithms-and-the-art-of-hyperparameter-selection-279d3b04c281

Singh, H. (2024, February 14). Forward Feature selection in Machine Learning: A Comprehensive guide. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/forward-feature-selection-and-its-implementation/

Suresha, M., Kuppa, S., & Raghukumar, D. S. (2020). A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. International Journal of Multimedia Information Retrieval (Internet), 9(2), 81–101. https://doi.org/10.1007/s13735-019-00190-x

Turney, S. (2024, February 10). Pearson Correlation Coefficient (r) | Guide & Examples. Scribbr. https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model using Random Forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134–60149. https://doi.org/10.1109/access.2019.2914999

Victoria, A. H., & Maragatham, G. (2020). Automatic tuning of hyperparameters using Bayesian optimization. Evolving Systems, 12(1), 217–223. https://doi.org/10.1007/s12530-020-09345-2

Virmani, D., & Pandey, H. (2022). Comparative analysis on effect of different SVM kernel functions for classification. In Lecture notes in networks and systems (Online) (pp. 657–670). https://doi.org/10.1007/978-981-19-3679-1_56

Vishnoi, V. K., Kumar, K., & Kumar, B. V. R. (2021). A comprehensive study of feature extraction techniques for plant leaf disease detection. Multimedia Tools and Applications, 81(1), 367–419. https://doi.org/10.1007/s11042-021-11375-0

Wang, L., Zhu, J., & Zou, H. (2021). High-dimensional predictive modeling with feature selection. Journal of the American Statistical Association, 116(536), 1810-1828. https://doi.org/10.1080/01621459.2021.1938632

Xiang, L. (2022). Application of an improved TF-IDF method in literary text classification. Advances in Multimedia, 2022, 1–10. https://doi.org/10.1155/2022/9285324

Xu, J., Wang, X., & Li, Y. (2023). Deep convolutional neural networks for image classification and object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 456-472. https://doi.org/10.1109/TPAMI.2022.3212345

Yadav, G. S., Guha, A., & Chakrabarti, A. (2020). Measuring complexity in financial data. Frontiers in Physics, 8. https://doi.org/10.3389/fphy.2020.00339

Yadav, G. S., Guha, A., & Chakrabarti, A. (2020). Measuring complexity in financial data. Frontiers in Physics, 8. https://doi.org/10.3389/fphy.2020.00339

Zach. (2020, November 2). Introduction to Linear Discriminant Analysis. Statology. https://www.statology.org/linear-discriminant-analysis/

Zhang, Y., Gong, D., & Yang, J. (2021). Ensemble sparse coding for robust feature selection. IEEE Transactions on Neural Networks and Learning Systems, 32(9), 3909-3923. https://doi.org/10.1109/TNNLS.2020.3024877

Zhao, K. (2022, March 30). Feature Extraction using Principal Component Analysis — A Simplified Visual Demo. Medium. https://towardsdatascience.com/feature-extraction-using-principal-component-analysis-a-simplified-visual-demo-e5592ced100a

Zhou, H., Wang, X., & Zhu, R. (2021). Feature selection based on mutual information with correlation coefficient. Applied Intelligence, 52(5), 5457–5474. https://doi.org/10.1007/s10489-021-02524-x

Zhu, Y., Li, T., & Li, W. (2022). An efficient hybrid feature selection method using the Artificial Immune Algorithm for High-Dimensional data. Computational Intelligence and Neuroscience, 2022, 1–21. https://doi.org/10.1155/2022/1452301