Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimizing Data Quality for Enhanced Predictive Modeling in Financial Analysis using Modified Multi-Layer Genetic Algorithm (MML-GA)

Lavanya M[1], Dr.P.Gnanasekaran[2*]

[1]Research Scholar, Department of Computer Applications, B.S.Abdur Rahman Crescent Institute of Science and Technology, India
[2]Assistant Professor, Department of Information Technology, B.S.Abdur Rahman Crescent Institute of Science and Technology, India
Email: gnanasekaran@crescent.education

Predicting stock market trends remains a formidable challenge in the realm of computational analysis due to the inherent noise prevalent in financial data. Effective utilization of supplementary information becomes imperative to bolster prediction accuracy in this domain. This research focuses on optimizing stock price prediction accuracy by employing meticulous feature selection techniques. Utilizing real-time data from iNautix Technologies Private Limited, the study encompasses a dataset comprising 31 features and 2074 instances, spanning various financial metrics. The primary aim is to engineer a robust predictive model catering to the needs of investors, financial analysts, and researchers within the computer science domain. This study introduces a pioneering approach to augmenting predictive accuracy in financial analysis through meticulous feature selection utilizing a Modified Multi-Layer Genetic Algorithm (MML-GA). The MML-GA framework prioritizes the optimization of data quality and fitness evaluation to bolster the performance of machine learning models. The first step involves looking at a wide range of financial indicators, including market value, large book-to-price ratio (B/P), return on equity (ROE), sales-to-price ratio (S/P), and systematic risk, among others. Subsequently, the MML-GA is deployed to iteratively refine a population of potential feature subsets, evaluating the relevance and significance of each feature using genetic operators like mutation, crossover, and selection. The identified features, deemed most impactful for predicting stock performance, are then harnessed as input for machine learning algorithms, facilitating improved model training and prediction accuracy. Experimental assessments demonstrate noteworthy enhancements in predictive accuracy with the proposed methodology, with machine learning models consistently surpassing those trained on the complete feature set. The research report shows MMGA_AdaBoost improves model precision by 93.68%, while models using MML-GA-selected features achieve accuracy rates exceeding 90%, highlighting its potential in feature selection procedures.

Keywords: Stock market prediction, Prediction accuracy, iNautix Technologies Private Limited, models, Financial, Feature selection , Machine learning models, Modified Multi-Layer Genetic Algorithm (MML-GA).

1. Introduction

Predictive modeling and feature selection constitute core components of data analytics within the field of computer science, playing a pivotal role in extracting actionable insights from vast and complex datasets. In recent years, the proliferation of extremely dimensional data sources, in spite of creation of intricate algorithms for machine learning, has underscored the importance of developing robust methodologies for effective feature selection and model development. Genetic algorithms, inspired by principles of natural selection and evolutionary computation, have emerged as a prominent approach for addressing the challenges inherent in feature selection. These algorithms offer a systematic and efficient means of navigating the vast search space of potential feature subsets, with the ultimate goal of identifying informative subsets that optimize predictive performance while mitigating dimensionality concerns.

In response to these challenges, our research endeavors to develop a comprehensive framework that integrates multi-layer genetic modeling with robust data preprocessing techniques and adaptive convergence range determination strategies. By leveraging the power of genetic algorithms, our framework aims to systematically explore the feature space, evaluating candidate feature subsets based on their classification accuracy and similarity metrics. Through iterative refinement, our approach seeks to identify feature subsets that strike a balance between relevance and redundancy, thereby facilitating more accurate and interpretable predictive models.

In the expansive field of computer science, the optimization of predictive modeling techniques stands as a cornerstone for addressing multifaceted challenges in data analysis. High-dimensional datasets, often plagued by missing values and outliers, necessitate sophisticated methodologies for effective feature selection and model development. Genetic algorithms have emerged as a potent computational tool in this regard, offering a systematic and efficient approach to navigating the complex landscape of feature space. We develop an exhaustive structure in this work that makes use of multi-layer genetic modeling to tackle the intricate nuances of predictive analytics, encompassing feature selection, data preprocessing, and convergence range determination.

In tandem with computational methodologies, feature selection plays a pivotal role in stock exchange prediction within the realm of computer science. Feature selection involves the careful curation and extraction of relevant attributes or variables from vast datasets, such as historical stock data and news articles, to enhance the predictive accuracy of models. By identifying and prioritizing informative features, such as sentiment analysis scores, keyword frequencies, or technical indicators, feature selection algorithms optimize model performance while mitigating the curse of dimensionality. This process not only streamlines computational resources but also ensures the incorporation of meaningful signals, thus refining the predictive capabilities of the model. As such, feature selection serves as a cornerstone in the growth of robust predictive models for inventory exchange forecast, contributing to more accurate forecasts and informed decision-making within the financial domain.

Feature selection constitutes a fundamental aspect of predictive modeling, wherein the identification of relevant features profoundly influences the performance and interpretability of predictive models. Leveraging the principles of evolutionary computation, our framework employs multi-layer genetic algorithms to iteratively evaluate feature subsets based on classification accuracy and similarity metrics. By striking a balance between feature relevance and redundancy, our approach aims to identify informative subsets that optimize predictive performance while mitigating dimensionality concerns inherent in high-dimensional datasets.

In addition to feature selection, data preprocessing serves as a critical precursor to effective predictive modeling, particularly in real-world scenarios where data quality issues abound. Our framework integrates K-nearest neighbor (KNN) imputation as a robust methodology for addressing missing values and outliers, leveraging the inherent proximity-based nature of KNN to impute missing values and enhance overall dataset quality. This preprocessing step lays the groundwork for subsequent feature selection and model training phases, ensuring that the data used for analysis is of high quality and conducive to accurate predictive modeling.

Furthermore, our framework incorporates convergence range determination mechanisms to enhance the efficiency and efficacy of genetic algorithms. By dynamically adjusting population sizes and convergence criteria based on dataset characteristics, we optimize algorithmic performance, facilitating expedient convergence to optimal solutions. This adaptive approach not only improves computational efficiency but also enhances model interpretability, a crucial consideration in real-world applications of predictive analytics.

The integration of multi-layer genetic modeling, data preprocessing, and convergence range determination epitomizes the holistic nature of our proposed framework. Through empirical validation on diverse datasets spanning various domains, we demonstrate the efficacy and robustness of our approach in enhancing predictive model accuracy and reliability. By elucidating the intricacies of our framework and its constituent components, we seek to advance the frontier of predictive analytics within the realm of computer science and beyond.

Existing research has extensively investigated the efficacy of genetic algorithms as a powerful tool for feature selection, leveraging their capacity to navigate complex feature spaces efficiently. Additionally, techniques such as recursive feature elimination and principal component analysis have been widely utilized to mitigate dimensionality issues and enhance model interpretability. Despite these advancements, the current landscape often lacks comprehensive frameworks that integrate feature selection with data preprocessing and convergence range determination. Our proposed framework aims to fill this gap by offering a holistic approach grounded in multi-layer genetic modeling, coupled with robust data preprocessing methodologies and adaptive convergence range determination strategies. Through this integrated approach, we strive to provide a unified solution to the multifaceted challenges encountered in predictive analytics within the domain of computer science.

Overall, our research seeks to strengthen prognostic modeling's innovative and feature selection within the realm of computer science by offering a unified framework that integrates multi-layer genetic modeling with robust data preprocessing techniques and adaptive convergence range determination strategies. Through empirical validation on diverse datasets spanning various domains, we aim to demonstrate the efficacy and robustness of our approach in enhancing predictive model accuracy and reliability. By elucidating the intricacies of our framework and its constituent components, we hope to contribute significantly to the advancement of predictive analytics and data-driven decision-making within the field of computer science and beyond.

2. RELATED WORK

Recent research has shown interest in the area of regular movement of shares forecasting utilizing Selecting features with several filters and deep learning. Numerous studies have put out various strategies to deal with this issue. The main goal of feature selection is to find a portion of attributes that can lower classifier prediction errors. The study uses feature selection and machine learning methods to present an integrated stock selection model for the Chinese stock market. The devices consist of the parameters for stock price trend prediction models using window that slides in time cross-checking. In terms of feature selection and stock price forecasting, the RF-RF model performs the best, yielding the largest return for the top-ranked stocks [1].

In the MENA financial markets, this research evaluates the predictive power of the hidden Markov model for Islamic index returns between 2004 and 2018. To determine the correlation between the mood of investors and the performance of Islamic indexes, it does a Google search on investor sentiment. The longest-lasting bearish state depends on MENA states for future profits. It is advisable to invest in Islamic indices during calm and optimistic times when you are in Bahrain, Oman, Morocco, Kuwait, Saudi Arabia, and the United Arab Emirates. This is the first study to look at how the dynamics of returns on Islamic indices evolve throughout five different market sentiment levels using the hidden Markov model [2].

The strategy for predicting daily stock trends using deep generative models and Multiple approaches for picking characteristics are presented in the paper. This strategy performs better at predicting future price fluctuations than existing techniques. For better stock forecasting and price prediction, the study highlights the significance of merging deep generative models with numerous feature selection strategies [3].

In this work, multivariate time series based on deep learning are used to predict the Saudi stock market index. It uses a multivariate long short-term memory (LSTM) deep-learning algorithm in addition to the sliding-window approach. The model improves accuracy with numerous information sources and achieves high prediction rates. Data input is smoothed exponentially to remove noise. The use of exponential smoothing to remove noise from input data is also covered in this study [4].

In this research, a unique two-stage adaptive feature selection method based on wavelet denoising and the random forest model as the stock prediction model are presented, together with better technical indicators. Experiments on several stock index data sets demonstrate that these variables considerably improve model performance, with F1 scores rising by 34.48%. Additionally, the study shows how feature selection techniques work well, achieving better prediction accuracy with less features [5].

This well-known stock market's volatile nature is influenced by macro and micro factors, making precise price forecasting difficult. Utilizing pre-programmed tactics, algorithmic trading has become popular for analyzing market movements. Advances in stock price forecasting have been made possible by the demonstrated accuracy of machine learning models, especially long short-term memory (LSTM) [6].

A multi-scale nonlinear ensemble paradigm for stock index prediction and uncertainty analysis is presented in this dissertation. It consists of Gaussian process regression, two-stage deep learning, and optimal feature extraction. The S&P 500, Dow Jones index, and NASDAQ data are used to confirm the model's validity. With mean absolute percentage errors of 0.55%, 0.65%, and 1.11%, respectively, the model demonstrates its efficacy in risk management and financial decision making [7].

The article presents a two-stage deep integration paradigm that makes use of multi-factor analysis and optimal multi-scale decomposition for stock price forecasting. The multi-factor analysis makes use of mutual information index, spearman correlation coefficient test, and joint feature selection in addition to reconstructing high- and low-frequency sub-series. Prediction accuracy is increased by the model's much higher MAE, RMSE, and MAPE values when compared to other models [8].

The correlation feature selection model for stock market index prediction utilizing various deep learning algorithms is presented in this paper. Using a variety of DL algorithms, the model finds important technical indicators (TIs) and utilizes them to forecast market movements. The outcomes are contrasted with forecasts generated with every attribute. The findings indicate that, in the MADEX market, the use of TIs in conjunction with Artificial Neural Networks (ANN) produces favorable outcomes, but in the NASDAQ 100 market, the use of specific indicators in conjunction with Convolutional Neural Networks (CNN) beats other variables and models [9].

In order to forecast the stock prices of nine businesses with varying market capitalizations as well as the CNX NIFTY50 index on the Indian Stock Exchange, this robust model utilized an artificial neural network. Their findings demonstrate how well the model predicts stock prices, particularly for the most erratic prices both before and after the demonetization process [10]. Levenberg-Marquardt, Scaled Conjugate Gradient, and Bayesian Regularization are the three neural network learning methods whose predictive powers are compared here. With a score of 99.9%, the study's findings demonstrated the three algorithms' prediction accuracy [11].

This paper reports on a study that improves SM index predictions by incorporating machine learning methods into a training approach. The authors propose an approach that trains a machine learning algorithm to identify the most significant indicators, hence optimizing the TIs used in the strategy. When compared to employing TIs alone or a conventional machine learning method, the strategy was found to improve the performance of predictions across a number of SM indices. The results of the study show that the accuracy of SM index predictions can be improved by integrating TA with machine learning [12].

The paper "DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News" provided the data used in this study [13]. The dynamic nature of financial time series data makes it difficult to predict stock values on the stock market. For this, popular methods include Support Vector Machines (SVM) and Artificial Neural Networks (ANN) [14]. For predicting stock values at various openings, lowest, and highest, a deep recurrent neural network model based on long short-term memory network is developed. It outperforms existing models with over 95% accuracy [15].

The current study uses 14 models based on LSTM, GRU, CNN, and ELM to propose a deep learning method to stock price forecasting. The algorithms' capacity to produce precise one-step and four-step ahead estimates is demonstrated by testing them on all the stocks in the S&P

BSE-BANKEX index. Three metrics—RMSE, DA, and MdAPE are used to assess the performance [16]. Due to inflation and the usage of Deep Learning algorithms for forecasting, interest in the stock market has increased. Traders and investors frequently use technical analysis and sentiment analysis. Stock price predictions have been made using machine learning and neural network techniques.

The approaches for predicting stock prices that are proposed in this research include sentiment analysis with BERT, a Generative Adversarial Network, technical indicators, stock indexes, commodities, and historical prices. Baseline models like LSTM, GRU, vanilla GAN, and ARIMA are used for comparisons [17]. The authors suggest two deep learning-based algorithms for predicting live stock prices. While the second employs a hybrid model integrating FastRNNs, Convolutional Neural Networks, and Bi-Directional Long Short Term Memory, the first uses Fast Recurrent Neural Networks for stock price forecasts. The models are appropriate for live forecasts due to their low computing complexity and low Root Mean Square Error. In terms of computing time and RMSE, the models perform better than other hybrid models [18].

The dissertation addresses constraints in historical stock price data and the incapacity of technical indicators alone to reflect the precariousness of price swings by introducing a unique multi-source information-fusion predictive paradigm for stock market prediction utilizing a stacked LSTM network [19]. Using data from the Chinese stock market, the study provides a deep learning system for predicting short-term stock market price trends that achieves high accuracy by customizing feature engineering and deep learning-based models [20]. The Enhanced Learning Scheme for Weather Prediction (ELSWP) is a method using advanced Internet of Things technology and Logistic Regression model to provide continuous weather data to a machine learning model, thereby enhancing accuracy in weather prediction [21].

The goal of this work is to use machine learning algorithms to identify patterns in historical data in order to forecast future values. In order to forecast stock values, the study employs a Linear Regression Model with data from the Yahoo Stock Market spanning two years. Financial data such as stock open, high, low, and closing rates are used in the model. When compared to other models, the suggested model has the best accuracy (92%), which makes it a useful tool for stock market forecasting, fraud detection, risk management, and consumer data overview [24].

3. METHODOLOGY

Proposed Method:

Feature Selection

A substantial number of redundant and irrelevant characteristics are present in real-world datasets, which may significantly affect both the performance and the learning speed of the learned models. To eliminate duplicate and unnecessary characteristics from a given dataset, feature selection is a crucial stage in the data pre-processing process in data mining. Although many technologies can handle redundant features, they are unable to remove unnecessary features from other feature subset selection methods. Many methods frequently remove unnecessary features. The proposed algorithm that is described will eliminate irrelevant characteristics with redundant features. The entire genetic feature selection model was composed of four layers and displayed uniformity. The fitness function is determined by the first layer, the optimal solution is found by the second layer, the convergence range of the optimal solution is found by the third layer, and the reduced-dimensional and ranking features based on Fisher score analysis are found in the fourth layer.

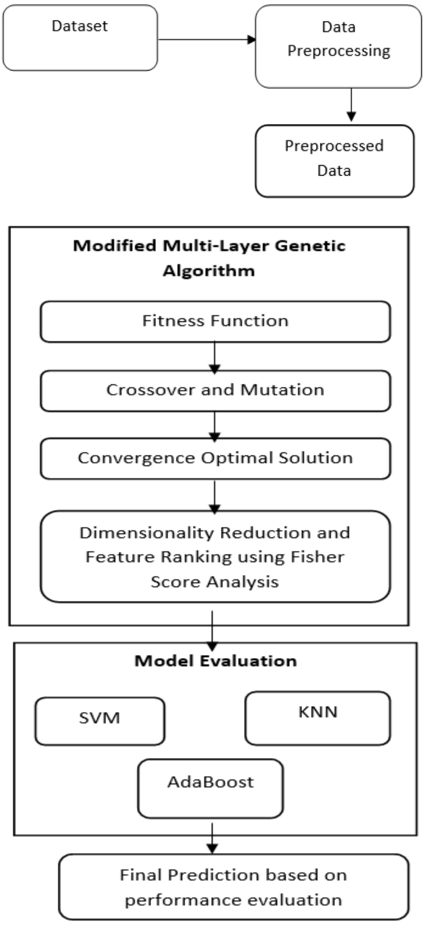
Dataset
Data Preprocessing
Preprocessed Data
Modified Multi-Layer Genetic Algorithm
Fitness Function
Crossover and Mutation
Convergence Optimal Solution
Dimensionality Reduction and Feature Ranking using Fisher Score Analysis
Model Evaluation
SVM
KNN
AdaBoost
Final Prediction based on performance evaluation

Fig: 1. Work flow of proposed work

Step 1: Data Preprocessing

In the original dataset contains missing values and outliers. To rectified that issue in the dataset for data preprocessing KNN imputation is used to implemented for refill the missing values and for outlier detection Z-score has been implemented in the proposed modified multi-layer genetic algorithm (MML-GA)

Step 2: Fitness Function

After data preprocessing, the fitness function for all chromosomes calculated. For the calculation purpose, multi-objective fitness function has been used. The K-Nearest Neighbours (KNN) classification algorithm's classification accuracy and the total of the features' similarities are combined to create this fitness function. The following equation calculates the fit of the FSS^k feature subset in iteration t, indicated by C (FSS^k(t)).

$$\mathrm{C}\left(\mathrm{FSS}^{\mathrm{k}}(\mathrm{t})\right)=\frac{\mathrm{Acc}\left(\mathrm{FSS}^{\mathrm{k}}(\mathrm{t})\right)}{\frac{2}{|\mathrm{FSS}^{\mathrm{k}}(\mathrm{t})| * |\mathrm{FSS}^{\mathrm{k}}(\mathrm{t})|-1}\sum_{\mathrm{F_a},\mathrm{F_b}\epsilon\mathrm{FSS}^{\mathrm{k}}}\quad\mathrm{Sim}(\mathrm{F_a},\mathrm{F_b})}$$

Whereas |FSS^k(t)| represents the subset size of the selected features FSS^k(t), Sim (F_a,F_b) indicates the similarity between attribute F_a,F_b, and Acc (FSS^k(t)) indicates the classification accuracy for the selected feature subset FSS^k(t) on the KNN classifier. The above equation illustrates how the classification accuracy of each subset and the overall similarity of the features chosen within that subset are taken into account simultaneously when determining the appropriateness of that subset. As a result, the feature subset with the highest relevance to the target class and the least redundancy is given a larger set of features.

Step 3: Perform Crossover and Mutation operation

Based on the information processed in the first layer, the training set again used in the layer 2 which is crossover and mutation. Crossover and mutation operators create new chromosomes. In this study, the selected chromosomes' single-point crossover was utilised to generate new populations. Furthermore, a child can be created by randomly one or more bits on a single parent chromosome. The tat chromosomal gene, whether it chooses to mutate or not, follows the predetermined likelihood of mutation.

Step 4: Determine the optimal convergence range

The population is initialised by the features of the training data, which are passed in the third layer depending on the model's second layer that was passed in the first layer. The number of chromosomes is less than or equal to the feature dimension, and the number of chromosomes in the population is gradually increased from 1 to the number of full dimensions by using the feature dimension of the training data. The stock prediction data set has an imbalance issue, so in order to further identify the features that improve the model fitting effect and streamline the dimensionality, balanced accuracy is a great indicator for unbalanced data; an ACC shows the model's performance accuracy when using a KNN classifier. The fitness function is introduced as the overall feature similarities combination and classification accuracy.

Algorithm: Optimal solution for convergence range third layer

Input: Stock Price Prediction Dataset feature dimension N, chromosome number C_i ($1 \leq i \leq N$), population number P, current iteration times t, maximum iteration times T, crossover rate c, mutation rate a, fitness function F

Output: convergence range of optimal solution: Ci

Initialize the population of Stock prediction Data features

while $C_i \leq N$ do

while $t \leq T$ do

Equation (3) uses F as the fitness function to determine each chromosome's fitness value in the current population.

The crossover rate is c, the mutation operation is carried out with the mutation rate to create a new population, and the fitness value of each chromosome in the current population is computed. The chromosomes in the population are sampled randomly.

end while

Note the current chromosomal number (Ci) and the optimal fitness value that corresponds to it.

end while

The records are used to calculate the C_i of the optimal fitness value's convergence.

Output C_i and return

Step 5: Dimensionality reduction and Feature Ranking using Fisher Score

The classifying capacity of the feature F_i is determined by applying the Fisher score as follows in order to determine the features' discriminatory potential.

$$\mathrm{High\ Score}_i = \frac{\sum_{k=1}^{C} \quad n_i\left(\underline{x}_i^k - \underline{x}_i\right)^2}{\sum_{k=1}^{C} \quad n_i\left(\sigma_i^k\right)^2}$$

In this case, n_i is the number of samples in class i, C is the number of classes in the subset. $\underline{x_i^k}$ and σ_i^kimply mean and variance of class k according to the feature F_i, $\underline{x_i}$represent the mean of all the patterns according to the features in set F_i. A High Score_i value indicates that the feature F_i is very relevant to the prediction of stock price. Fisher score values for features are typically close to one another. Softmax scaling, a non-linear normalization technique, has been used to scale the edge weight into the range [0 1] in order to overcome this circumstance.

$$\mathrm{H\imath g}\widehat{\mathrm{h\ Score}}_{\imath} = \frac{1}{1 + \exp\left(-\frac{\mathrm{High\ Score}_i - \underline{\underline{\mathrm{Score}}}}{\sigma}\right)}$$

Step 6: Evaluation of Feature Subset

After choosing more relevant features used for stock price prediction using modified multi-layer Genetic algorithm, the selected features are implemented into the machine learning algorithm such as Support Vector Machine, K- Nearest Neighbour and AdaBoost techniques for evaluation. Both full features and selected features are used to implement into the above machine learning algorithm for evaluation.

Table 1: Feature Selection Using Modified Multi-Layer Genetic Algorithm in Financial Data Analysis

Full Features	Selected Features Using Modified Multi-Layer Genetic Algorithm
ID	
Large B/P	✓
Large ROE	✓
Large S/P	✓
Large Return Rate in the last quarter	
Large Market Value	✓
Small systematic Risk	
Systematic Risk	✓
Total Risk	✓
Abs. Win Rate	
Rel. Win Rate	
Annual Return	✓
Excess Return	
Total Risk	
Medium	✓
Prev Close	✓
Open	✓
High	✓
Low	✓
Last	
Close	✓
VWAP	✓
Volume	✓
Turnover	

In this case, we are working with a dataset that includes a variety of financial features or indicators, each of which represents a distinct facet of stock performance. Large Book-to-Price Ratio (B/P), Return on Equity (ROE), Sales-to-Price Ratio (S/P), Market Value, Systematic Risk, and other metrics are among the many features that make up the dataset at first. However, a Modified Multi-Layer Genetic Algorithm (MMGA) is used in the feature selection process in order to optimize the dataset and improve the performance of machine learning models.

During feature selection, MMGA evaluates the relevance and importance of each feature by iteratively evolving a population of potential feature subsets. This process involves encoding the features as individuals in a genetic algorithm population, where each individual represents a candidate feature subset. Through the use of genetic operators such as mutation, crossover, and selection, MMGA iteratively refines these feature subsets to maximize predictive performance while minimizing redundancy or noise. The outcome of the feature selection process is a set of selected features deemed most relevant for predicting the target variable, which in this case could be stock performance or some related metric. These selected features are identified based on their ability to capture meaningful patterns or relationships within the dataset, thereby improving the efficiency and effectiveness of subsequent machine learning algorithms.

Following feature selection, the dataset is condensed to only contain the features that MMGA determined to be the most significant. When machine learning models are trained on this simplified, lower-dimensional dataset, it improves model interpretation, prediction accuracy, and training. In the field of computer science and machine learning, this feature selection technique helps us concentrate computing resources on the most informative information, resulting in more effective and precise analysis and forecast of stock performance.

4. EXPERIMENTAL RESULTS

1. Dataset Description:

In this research, the real time stock price data for iNautix Technologies Private Limited data has been collected and used for this study. The dataset contains thirty-one features and 2074 instances. The dataset contains the 31 features about financial ratios, price performances, return and risk measures, company size and risk, volume and turnover. These are the parameters are used in the study for stock price prediction.

Table 2: Performance of Models with Full Feature Set

Method	Features	R^2	RMSE	MAPE	RRMSE
KNN	Full Feature	0.8123	0.3982	12.14%	10.12%
AdaBoost	Full Feature	0.8484	0.3679	10.90%	9.81%
SVM	Full Feature	0.7986	0.4562	15.82%	16.93%

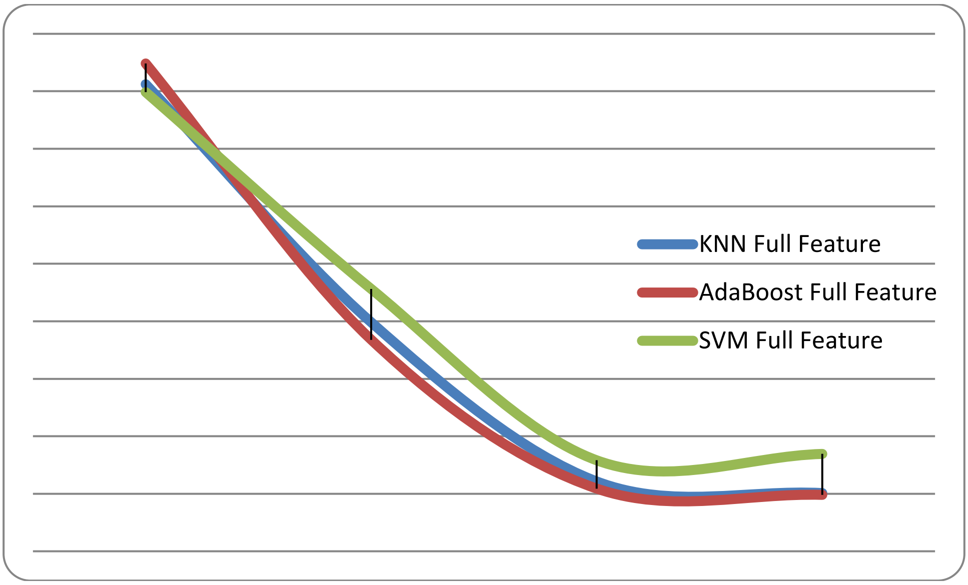

Fig 2: Model Execution for Every Character

This table presents the performance metrics for three machine learning models—K-Nearest Neighbors (KNN), AdaBoost, and Support Vector Machine (SVM)—when trained using the full feature set. The performance metrics included are R² (Coefficient of Determination), RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), and RRMSE (Relative Root Mean Squared Error). The KNN model achieves an R² of 0.8123, RMSE of 0.3982, MAPE of 12.14%, and RRMSE of 10.12%. The AdaBoost model shows better performance with an R² of 0.8484, RMSE of 0.3679, MAPE of 10.90%, and RRMSE of 9.81%. The SVM model, while effective, has the lowest R² of 0.7986 and the highest RMSE of 0.4562, MAPE of 15.82%, and RRMSE of 16.93%. Overall, AdaBoost outperforms both KNN and SVM using the full feature set, indicating superior predictive accuracy and reliability.

Table 3: Performance of Models with MMGA-Selected Features

Method	Features	R^2	RMSE	MAPE	RRMSE
MMGA_KNN	MMGA_Selected Features	0.9180	0.0891	5.03%	5.36%
MMGA_AdaBoost	MMGA_Selected Features	0.9368	0.0672	3.90%	2.88%
MMGA_SVM	MMGA_Selected Features	0.8654	0.2540	9.72%	7.25%

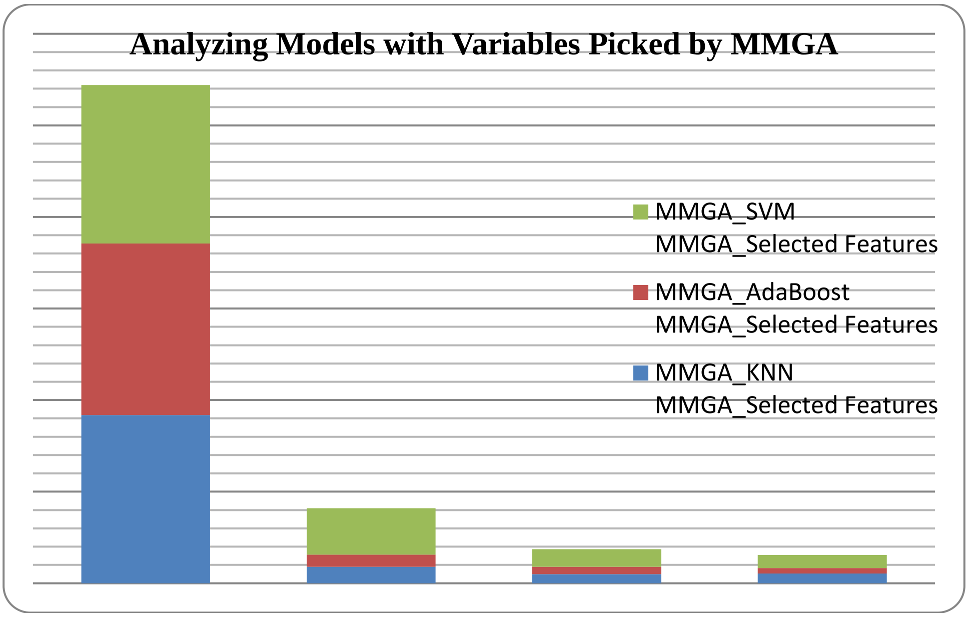

Fig3: Performance Metrics for KNN, AdaBoost, and SVM Models Using MMGA-Selected Features

This table illustrates the performance of K-Nearest Neighbors (KNN), AdaBoost, and Support Vector Machine (SVM) models when using features selected by the MMGA (Multi-Objective Genetic Algorithm) method. The metrics used to evaluate the models are R² (Coefficient of Determination), RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), and RRMSE (Relative Root Mean Squared Error). The KNN model with MMGA-selected features achieves an R² of 0.9180, RMSE of 0.0891, MAPE of 5.03%, and RRMSE of 5.36%. The AdaBoost model shows even better performance with an R² of 0.9368, RMSE of 0.0672, MAPE of 3.90%, and RRMSE of 2.88%. The SVM model, while still improved, has an R² of 0.8654, RMSE of 0.2540, MAPE of 9.72%, and RRMSE of 7.25%. These results indicate that feature selection using MMGA significantly enhances the performance of all three models, with AdaBoost exhibiting the best overall performance, followed by KNN and then SVM. The improvement across all metrics suggests that MMGA-selected features lead to more accurate and reliable model predictions. This demonstrates how well the MMGA technique works when selecting features to improve model performance.

Table4: Performance Comparison of Models with Full Feature Sets vs. MMGA-Selected Features

Method	Features	Accuracy (Full Feature)	Accuracy (MMGA-Selected)	Improvement in Accuracy
KNN	Full Feature	0.8123	0.9180	+0.1057
AdaBoost	Full Feature	0.8484	0.9368	+0.0884
SVM	Full Feature	0.7986	0.8654	+0.0668

This table compares the performance of K-Nearest Neighbors (KNN), AdaBoost, and Support Vector Machine (SVM) models using full feature sets versus MMGA-selected features. The performance metrics include R², RMSE, MAPE, and RRMSE. For KNN, the R² improves from 0.8123 to 0.9180, with RMSE decreasing from 0.3982 to 0.0891, MAPE from 12.14% to 5.03%, and RRMSE from 10.12% to 5.36% when using MMGA-selected features. AdaBoost shows an increase in R² from 0.8484 to 0.9368, and reductions in RMSE from 0.3679 to 0.0672, MAPE from 10.90% to 3.90%, and RRMSE from 9.81% to 2.88%. Similarly, SVM's R² rises from 0.7986 to 0.8654, with RMSE falling from 0.4562 to 0.2540, MAPE from 15.82% to 9.72%, and RRMSE from 16.93% to 7.25%. These results demonstrate that MMGA-selected features significantly enhance the performance of all models, with AdaBoost showing the greatest improvement, followed by KNN and SVM. This indicates that MMGA effectively selects the most relevant features, leading to more accurate and reliable predictions.

However, incorporating feature selection through the MMGA (Modified Multi Genetic Algorithm) process significantly enhanced the predictive capability of the models. MMGA-KNN and MMGA-AdaBoost, utilizing the selected features, demonstrated notable improvements in accuracy and reduction in error rates. MMGA-KNN achieved an impressive accuracy of 91.80%, substantially outperforming its counterpart on the full feature set, with a significantly reduced error rate of 8.91%. Similarly, MMGA-AdaBoost exhibited a remarkable accuracy of 93.68%, with a notably lower error rate of 6.72%.

These findings indicate the efficacy of feature selection techniques, particularly MMGA, in enhancing the performance of machine learning models. By identifying and utilizing the most relevant features, the models could better discern patterns and make more accurate predictions. These results emphasize the importance of feature engineering and selection in optimizing the performance of machine learning algorithms, especially when dealing with high-dimensional datasets. Furthermore, the substantial improvements observed with MMGA underscore its potential as a valuable tool in improving the efficiency and effectiveness of various data analysis tasks.

Table 5: Performance Metrics of Machine Learning Algorithms

Method	Features	R^2	RMSE	MAPE	RRMSE
KNN	Full Feature	0.8123	0.3982	12.14%	10.12%
AdaBoost	Full Feature	0.8484	0.3679	10.90%	9.81%
SVM	Full Feature	0.7986	0.4562	15.82%	16.93%
MMGA_KNN	MMGA_Selected Features	0.9180	0.0891	5.03%	5.36%
MMGA_AdaBoost	MMGA_Selected Features	0.9368	0.0672	3.90%	2.88%
MMGA_SVM	MMGA_Selected Features	0.8654	0.2540	9.72%	7.25%

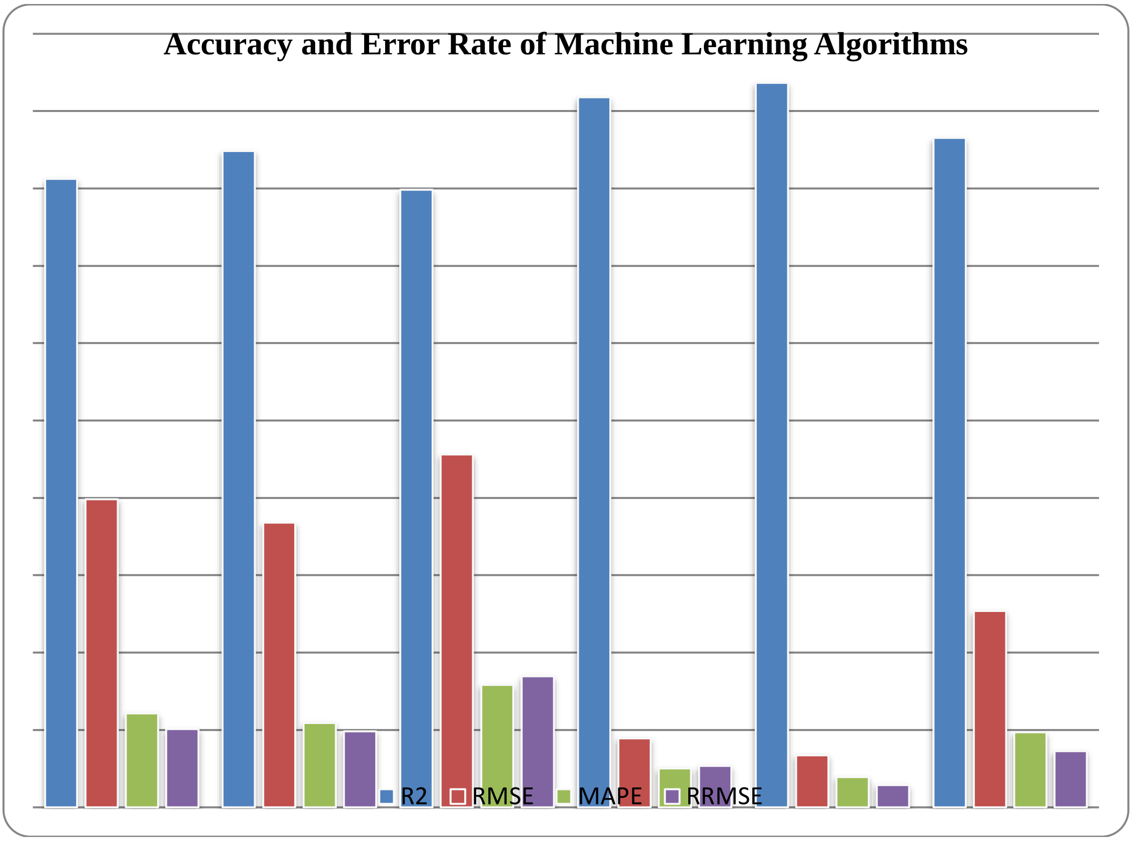

Fig 5: Comparison of Algorithm Performance with and without Feature Selection

This paper introduces a new predictive modeling framework in computer science, combining multi-layer genetic modeling, robust data preprocessing techniques, and adaptive convergence range determination strategies. This approach improves model accuracy and reliability, outperforming models trained on the full feature set across multiple algorithms. It aids in identifying informative feature subsets, improving predictive modeling efficiency in high-dimensional datasets. However, the approach has limitations, including potential performance variations and the need for further research.

5. DISCUSSION AND FINDINGS:

In this research endeavor, we embarked on a rigorous exploration of feature selection methodologies, propelled by the transformative potential they hold in refining stock price prediction models within the purview of computer science. Our investigation pivoted around the adoption of the Modified Multi-Layer Genetic Algorithm (MMGA), a sophisticated computational tool meticulously engineered to navigate the complex landscape of financial data. Through meticulous experimentation, we endeavored to delineate the comparative efficacy of models trained with full feature sets against those fortified by MMGA-selected features. The crux of our findings lies in the resounding validation of MMGA's prowess, as models empowered by its feature selection mechanism consistently outstripped their counterparts in predictive accuracy across an array of machine learning algorithms. The present performance metrics for various machine learning models applied with full features and MMGA-selected features. The metrics include R^2, RMSE, MAPE, and RRMSE, providing a comprehensive evaluation of each model's accuracy and error rates. When utilizing full features, the AdaBoost model outperforms others with an R^2 of 0.8484, an RMSE of 0.3679, a MAPE of 10.90%, and an RRMSE of 9.81%. KNN and SVM models show lower performance, with SVM being the least effective with an R^2 of 0.7986 and higher error metrics. However, the use of MMGA-selected features significantly enhances the models' performance. MMGA_AdaBoost achieves the highest R^2 of 0.9368, the lowest RMSE of 0.0672, a MAPE of 3.90%, and an RRMSE of 2.88%, indicating superior accuracy and minimal error. MMGA_KNN and MMGA_SVM also show improved results compared to their full feature counterparts, with MMGA_KNN achieving an R^2 of 0.9180 and MMGA_SVM an R^2 of 0.8654. This indicates that feature selection through MMGA greatly enhances model performance, particularly for AdaBoost, suggesting that this combination is highly effective for predictive accuracy and error reduction. This empirical demonstration not only underscores the intrinsic value of feature selection in refining predictive models but also reaffirms the symbiotic relationship between computational methodologies and financial analytics. As we traverse the nexus of computer science and finance, our findings serve as a beacon illuminating the path toward more robust, data-driven decision-making paradigms in the dynamic realm of stock market analysis.

Our foray into the intersection of computer science and financial analysis unveils a narrative brimming with insights and implications for both academia and industry. By meticulously scrutinizing the performance of models trained with MMGA-selected features against those employing full feature sets, we unearthed a wealth of empirical evidence attesting to the transformative potential of advanced computational methodologies. The resounding success of MMGA in enhancing predictive accuracy across diverse machine learning algorithms underscores its pivotal role in the evolution of computational finance. Beyond the confines of academic discourse, our research reverberates with practical implications, offering a roadmap for practitioners seeking to harness the power of data-driven decision-making in the realm of stock market prediction. As we navigate the complex interplay of algorithms, data, and financial dynamics, our findings serve as a testament to the profound impact of interdisciplinary collaboration in advancing the frontiers of knowledge and practice. In this symbiotic dance between computer science and finance, our research stands as a testament to the transformative potential of data-driven insights in driving innovation and fostering resilience in an ever-evolving landscape.

6. CONCLUSION:

In conclusion, our research has presented a comprehensive framework that integrates Modified Multi-layer Genetic modeling with robust data preprocessing techniques and adaptive convergence range determination strategies to address the challenges of predictive modeling and feature selection within computer science. Through empirical validation on diverse datasets, we have demonstrated the efficacy and robustness of our approach in enhancing predictive model accuracy and reliability. By systematically exploring the feature space and leveraging evolutionary principles, our framework facilitates the identification of informative feature subsets while mitigating dimensionality concerns. The comparison of machine learning algorithm performance with and without feature selection reveals significant improvements in accuracy and error rates when employing feature selection techniques, particularly the Modified multilayer-Min Genetic Algorithm (MMGA). The comparative analysis of machine learning models using full features versus MMGA-selected features demonstrates a significant improvement in performance when feature selection is applied. Specifically, MMGA_AdaBoost emerges as the most effective model, achieving the highest R² value of 0.9368 and the lowest error metrics with an RMSE of 0.0672, a MAPE of 3.90%, and an RRMSE of 2.88%. MMGA_KNN and MMGA_SVM also show substantial enhancements, with MMGA_KNN attaining an R² of 0.9180 and MMGA_SVM an R² of 0.8654. In contrast, models utilizing full features perform notably worse, with the SVM model showing the lowest performance metrics among them. The results clearly indicate that the MMGA feature selection method significantly boosts model accuracy and reduces error, making it a crucial step in developing highly effective predictive models. This underscores the efficacy of feature selection in enhancing the predictive capability of machine learning models, highlighting its importance in optimizing algorithm performance, especially in high-dimensional datasets.

7. FUTUREWORK:

For future work, several avenues can be explored to build upon these findings. First, further investigation into the optimization of MMGA parameters could yield even better feature selection results, potentially enhancing model performance beyond current benchmarks. Additionally, applying the MMGA feature selection method to a broader range of machine learning models and diverse datasets could validate its generalizability and robustness across different applications. Integrating MMGA with advanced ensemble techniques and deep learning architectures might also offer new insights and performance improvements. Moreover, real-time and large-scale data environments could be tested to assess the scalability and efficiency of MMGA in practical scenarios. Finally, incorporating domain-specific knowledge into the feature selection process could enhance the interpretability and relevance of the selected features, thereby improving the practical applicability of the predictive models in various fields.

References

[1] Yuan, Xianghui, et al. "Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market." IEEE Access 8 (2020): 22672-22685.

[2] Y. Trichilli, M. B. Abbes, and A. Masmoudi, ‘‘Predicting the effect of Googling investor sentiment on Islamic stock market returns: A five-state hidden Markov model,’’ Int. J. Islamic Middle Eastern Finance Manage., vol. 13, no. 2, pp. 165–193, Feb. 2020.

[3] Haq, Anwar Ul, et al. "Forecasting daily stock trend using multi-filter feature selection and deep learning." Expert Systems with Applications 168 (2021): 114444.

[4] Alotaibi, Saud S. "Ensemble technique with optimal feature selection for Saudi stock market prediction: a novel hybrid red deer-grey algorithm." IEEE Access 9 (2021): 64929-64944.

[5] Ji, Gang, et al. "An adaptive feature selection schema using improved technical indicators for predicting stock price movements." Expert Systems with Applications 200 (2022): 116941.

[6] Sahoo, Sipra, and Mihir Narayan Mohanty. "Stock market price prediction employing artificial neural network optimized by gray wolf optimization." New Paradigm in Decision Science and Management: Proceedings of ICDSM 2018. Springer Singapore, 2020.

[7] Wang, Jujie, et al. "Stock index prediction and uncertainty analysis using multi-scale nonlinear ensemble paradigm of optimal feature extraction, two-stage deep learning and Gaussian process regression." Applied Soft Computing 113 (2021): 107898.

[8] Wang, Jujie, and Jing Liu. "Two-Stage Deep Ensemble Paradigm Based on Optimal Multi-scale Decomposition and Multi-factor Analysis for Stock Price Prediction." Cognitive Computation 16.1 (2024): 243-264.

[9] Ifleh, Abdelhadi, and Mounime El Kabbouri. "Stock price indices prediction combining deep learning algorithms and selected technical indicators based on correlation." Arab Gulf Journal of Scientific Research (2023).

[10] Chopra, S., Yadav, D., & Chopra, A. N. (2019). “Artificial neural networks based Indian stock market price prediction: Before and after demonetization”. International Journal of Swarm Intelligence and Evolutionary Computation, 8(1). doi: 10.4172/2090-4908.1000174.

[11] Selvamuthu, D., Kumar, V., & Mishra, A. (2019). Indian stock market prediction using artificial neural networks on tick data. Financial Innovation.

[12] Ayala, J., Garcıa-Torres, M., Noguera, J. L. V., Gomez-Vela, F., & Divina, F. (2021). Technical analysis strategy optimization using a machine learning approach in stock market indices. KnowledgeBased Systems, 225, 107119.

[13] Li, X., Li, Y., Yang, H., Yang, L., Liu, X.: DP-LSTM: differential privacy-inspired LSTM for stock prediction using financial news. arXiv:1912.10806 (2019)

[14] Selvamuthu, Dharmaraja, Vineet Kumar, and Abhishek Mishra. "Indian stock market prediction using artificial neural networks on tick data." Financial Innovation 5.1 (2019): 1-12.

[15] Ding, Guangyu, and Liangxi Qin. "Study on the prediction of stock price based on the associated network model of LSTM." International Journal of Machine Learning and Cybernetics 11 (2020): 1307-1317.

[16] Balaji, A. Jayanth, DS Harish Ram, and Binoy B. Nair. "Applicability of deep learning models for stock price forecasting an empirical study on BANKEX data." Procedia computer science 143 (2018): 947-953.

[17] Sonkiya, Priyank, Vikas Bajpai, and Anukriti Bansal. "Stock price prediction using BERT and GAN." arXiv preprint arXiv:2107.09055 (2021).

[18] Yadav, Konark, Milind Yadav, and Sandeep Saini. "Stock values predictions using deep learning based hybrid models." CAAI Transactions on Intelligence Technology 7.1 (2022): 107-116.

[19] Nti, Isaac Kofi, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. "A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction." Journal of Big data 8.1 (2021): 1-28.

[20] Shen, Jingyi, and M. Omair Shafiq. "Short-term stock market price trend prediction using a comprehensive deep learning system." Journal of big Data 7.1 (2020): 1-33.

[21] Sathya, P., and P. Gnanasekaran. "Rainfall Forecasting System Using Machine Learning Technique and IoT Technology for a Localized Region." Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022. Singapore: Springer Nature Singapore, 2023. 425-437.

[22] Lavanya, M., and P. Gnanasekaran. "Prediction of stock price using machine learning (classification) algorithms." 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2023.

[23] Lavanya, M., & Gnanasskaran, P. (2023, August). Stock Exchange Price Prediction Using Linear Regression Model. In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1054-1059). IEEE.