Journal of Sciences, Islamic Republic of Iran 35(3): 243 - 256 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Evaluating Feature Selection Methods for Macro-Economic Forecasting, Applied for Iran's Macro-Economic Variables

M. Goldani*

Department of Political Science and Economics, Faculty of Literature and Humanities, Hakim Sabzevari University, Sabzevar, Islamic Republic of Iran

Received: 15 October 2024 / Revised: 3 January 2025 / Accepted: 2 February 2025

Abstract

This research examines different feature selection methods to enhance the predictive accuracy of macroeconomic forecasting models, focusing on Iran's economic indicators derived from World Bank data. Fourteen feature selection techniques were thoroughly compared, classified into Filter, Wrapper, Embedded, and Similarity-based categories. The evaluation utilized Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics under a 10-fold cross-validation scheme. The findings highlight that Stepwise Selection, Tree-based approaches, and Similarity-based methods, especially those employing Hausdorff and Euclidean distances, consistently outperformed others with average MAE values of 32.03 for Stepwise Selection and 62.69 for Hausdorff Distance. Conversely, Recursive Feature Elimination and Variance Thresholding exhibited weaker results, yielding significantly higher average MAE scores. Similarity-based approaches achieved an average rank of 9.125 across datasets, demonstrating their robustness in managing high-dimensional macroeconomic data. These outcomes underscore the value of integrating similarity measures with traditional feature selection techniques to improve the efficiency and reliability of predictive models, offering meaningful insights for researchers and policymakers in economic forecasting.

Keywords: Feature Selection; Predictive Accuracy; World Bank Indicators; Macroeconomic Analysis; Similarity Methods.

Introduction

The primary challenge of working with high-dimensional data lies in the exponential growth in complexity and sparsity that such data introduces. Additionally, the costs associated with storage and transmission increase, visualization becomes more challenging, and redundant or irrelevant features often complicate analysis (1). To address these challenges, dimensionality reduction techniques, feature selection, regularization methods, and meticulous data preprocessing are essential. These approaches help to extract valuable insights while mitigating the negative impacts of high dimensionality on data analysis and machine learning tasks. Feature selection is a key technique in dimensionality reduction, focusing on carefully identifying a relevant subset of features (variables or predictors) for model development. It plays a critical role in the data preprocessing workflow. Among various dimensionality reduction strategies, feature

* Corresponding Author: Tel: 05144410104; Email: m.goldani@hsu.ac.ir

Table 1. Summary of the specific objectives of the research

Objective	Details
Exploring Similarity-Based Feature Selection	Investigate using distance measures (e.g., Hausdorff, Euclidean, Dynamic Time Warping) as feature selection tools.
Benchmarking Against Conventional Methods	Compare similarity-based methods with Filter, Wrapper, and Embedded approaches using RMSE and MAE.
Assessing Practical Implications	Evaluate computational simplicity, robustness, and real-world applicability in economic forecasting.
Demonstrating Relevance with Case Studies	Use Iran's macroeconomic indicators (1990–2022) to validate findings and provide actionable insights.

selection stands out as a significant approach that retains only relevant features and eliminates redundant or irrelevant ones (2). Feature selection is vital in machine learning and data analysis, particularly when handling high-dimensional datasets. Identifying and selecting the most important features enhances model performance by improving predictive accuracy, reducing overfitting, and lowering computational costs. In the context of target variables, the feature selection process significantly contributes to achieving more accurate predictions.

Through systematically identifying and preserving relevant features while removing irrelevant or redundant ones, feature selection boosts a model's ability to capture underlying patterns and relationships within the data. This results in improved predictive accuracy, typically reflected in lower RMSE and MAE values. For instance, accurately forecasting GDP growth or inflation rates requires isolating key economic indicators such as broad money supply, government expenditure, and foreign direct investment. Similarly, understanding the drivers of unemployment or manufacturing growth necessitates focusing on the most impactful predictors. Effective feature selection not only enhances accuracy but also improves model interpretability and computational efficiency, aiding in better economic analysis and policymaking. Reducing model complexity also helps prevent overfitting, ensuring forecasts remain robust and reliable when applied to new data.

Feature selection (FS) is the process of identifying the most relevant and effective subsets of features to enhance the robustness of predictive models. This step is performed during the preprocessing phase of machine learning workflows. Before any training or testing, choosingthe most pertinent features based on the target variable is essential. While many FS techniques have been proposed in the literature, some methods, such as time series similarity methods, can also identify the most relevant features. A review of existing literature reveals that no studies have yet applied time series similarity methods specifically for feature selection. However, there are similarities between these two approaches that make time series similarity a promising alternative. Time series similarity measures the distance between two time series, which forms the foundation for clustering and classification tasks. A smaller distance between a feature and the target variable indicates that the feature is more relevant and should be included in the model. Thus, the goal of this research is to explore whether time series similarity methods can be as effective as traditional feature selection methods for identifying relevant feature subsets. The significance of this inquiry lies in the simplicity of the preprocessing step is just as important as the effectiveness of the methods employed, potentially saving both time and resources.

The overarching goal of this study is to evaluate the effectiveness of similarity-based methods as feature selection tools for high-dimensional macroeconomic forecasting. Table 1 summarizes the specific objectives of the research:

By addressing these objectives, the study contributes to advancing feature selection methodologies and provides practical recommendations for integrating similarity-based approaches in macroeconomic forecasting and other domains.

In the following sections, the methodology for integrating time series similarity measures into the feature selection framework is discussed (Section 2), the empirical results of the study are presented (Section 3), and the implications of the findings are analyzed in the discussion and conclusion (Section 4).

Literature review

Feature selection is essential for improving machine learning models accuracy, interpretability, and computational performance. By isolating the most significant features and eliminating those that are redundant or irrelevant, it addresses many of the challenges associated with high-dimensional datasets.

While feature selection has been extensively studied, the direct use of similarity measures as an independent method has not received much attention. Nevertheless, various studies have leveraged similarity measures indirectly to enhance feature selection techniques, as outlined below.

Similarity-based methods have shown potential, particularly in unsupervised feature selection. For example, Zhu et al. (3) proposed the Feature Selection-based Feature Clustering (FSFC) algorithm, which employs clustering driven by similarity measures to group and select features effectively. Similarly, Mitra (4) introduced an algorithm for unsupervised feature selection in large, high-dimensional datasets. This method evaluates features redundancy using similarity metrics, achieving greater efficiency and scalability.

Building on these ideas, Shi et al. (5) developed the Adaptive-Similarity-based Multi-modality Feature Selection (ASMFS) approach. This technique constructs a similarity matrix to capture inherent relationships across different modalities in high-dimensional data. The method demonstrated superior performance in tasks such as Alzheimer's disease classification, showcasing the value of similarity-based strategies in feature selection.

Recent research has refined similarity-based approaches to make them more robust and adaptable. Mehri et al. (6) employed similarity measures to identify and eliminate redundant features by examining their resemblance to others. Shen, Chen, and Garibaldi (7) proposed a meta-learning framework that integrates fuzzy similarity measures for recommending optimal feature selection techniques tailored to diverse datasets. Their approach automates feature selection, enhancing adaptability across dataset characteristics.

Goldani and Asadi (8) explored the application of similarity measures in financial forecasting, utilizing methods such as Haus Dorff distance and variance thresholds. These measures effectively selected predictive features, particularly in scenarios involving fluctuating data volumes. Similarly, Mathisen et al. (9) enhanced automated similarity measures for clustering, case-based reasoning, and one-shot learning, demonstrating their adaptability and utility in diverse applications.

Matrix factorization techniques have also leveraged similarity measures for feature selection. QI et al. (10) introduced the Regularized Matrix Factorization Feature Selection (RMFFS) method, which employs matrix factorization to capture feature correlations and applies a combination of l1 and l2 norms to ensure sparsity in the feature weight matrix. Du et al. (11) proposed the Robust Unsupervised Feature Selection via Matrix Factorization (RUFSM) method, which decomposes the data matrix into latent cluster centers and sparse representations. This approach achieves high-accuracy feature selection by identifying orthogonal cluster centers.

Hu et al. (12) extended this line of research with the Graph Self-Representation Sparse Feature Selection (GSR-SFS) method. Integrating a subspace-learning model into a sparse feature-level self-representation approach, improves both the interpretability and stability of the selected features.

Feature selection methods have found significant applications in medical and dynamic datasets. Remeseiro and Bolon-Canedo (2) reviewed feature selection techniques in medical imaging, biomedical signal processing, and DNA microarray data, highlighting their utility in solving domain-specific challenges. Venkatesh and Anuradha (13) addressed the limitations of traditional feature selection methods for dynamic, noisy datasets generated in IoT and web-based applications. Their work emphasized the need for scalable and robust methods to handle the evolving nature of such data.

The consensus among researchers, as highlighted by Guyon and Elisseeff (14), is that feature selection is crucial for improving the performance and interpretability of machine learning models. The choice of the feature selection method should be tailored to the specific problem and dataset, as there is no universal solution. Proper evaluation and validation are necessary to ensure the effectiveness of any feature selection technique. Jović et al. (15) investigated the calculation methods of standard filter, wrapper, and embedded methods. The result revealed that filters based on information theory and wrappers based on greedy stepwise approaches offer the best results.

The existing body of work highlights the potential of similarity-based methods to address challenges such as feature redundancy and relevance in high-dimensional data. While traditional feature selection methods such as Filter, Wrapper, and Embedded approaches have succeeded, integrating similarity measures directly into feature selection frameworks offers a promising alternative. However, their application remains underexplored in macroeconomic forecasting, which has motivated the current study to evaluate their feasibility and effectiveness in this context. This study bridges this gap by investigating the feasibility and effectiveness of using time series similarity methods as feature selection techniques. By systematically comparing these methods with established feature selection techniques, the research aims to evaluate their performance in identifying relevant subsets of features while ensuring computational simplicity and robustness. The findings have implications not only for improving the preprocessing of high-dimensional datasets but also for advancing methodologies in domains such as economic forecasting, healthcare, and beyond.

Materials and Methods

This section outlines the methodology employed in this research, consisting of four key steps as depicted in Figure 1.

Dataset

This paper aims to compare the predictive performance of datasets selected using feature selection techniques and time series similarity methods. The data set employed for this purpose is derived from the World Bank Development Indicators. To validate and assess the effectiveness of the dataset chosen through these methods, various target variables were selected, as summarized in the Table 2. These variables represent "Macroeconomic Indicators" for Iran, with data sourced from the World Bank website for 1990–2022.

Preprocessing data

As an initial step in the data preprocessing process, variables with a high proportion of missing data—specifically, those with more than 80% of their values absent—are systematically removed from the dataset to ensure the reliability and integrity of subsequent analyses. This step helps eliminate variables that otherwise provide insufficient information for meaningful insights. For the remaining variables, which have a missing data rate of less than 80%, the gaps in the dataset are addressed through the application of the K-Nearest Neighbors (KNN) imputation method. This technique leverages the patterns and relationships between existing data points to estimate and fill in

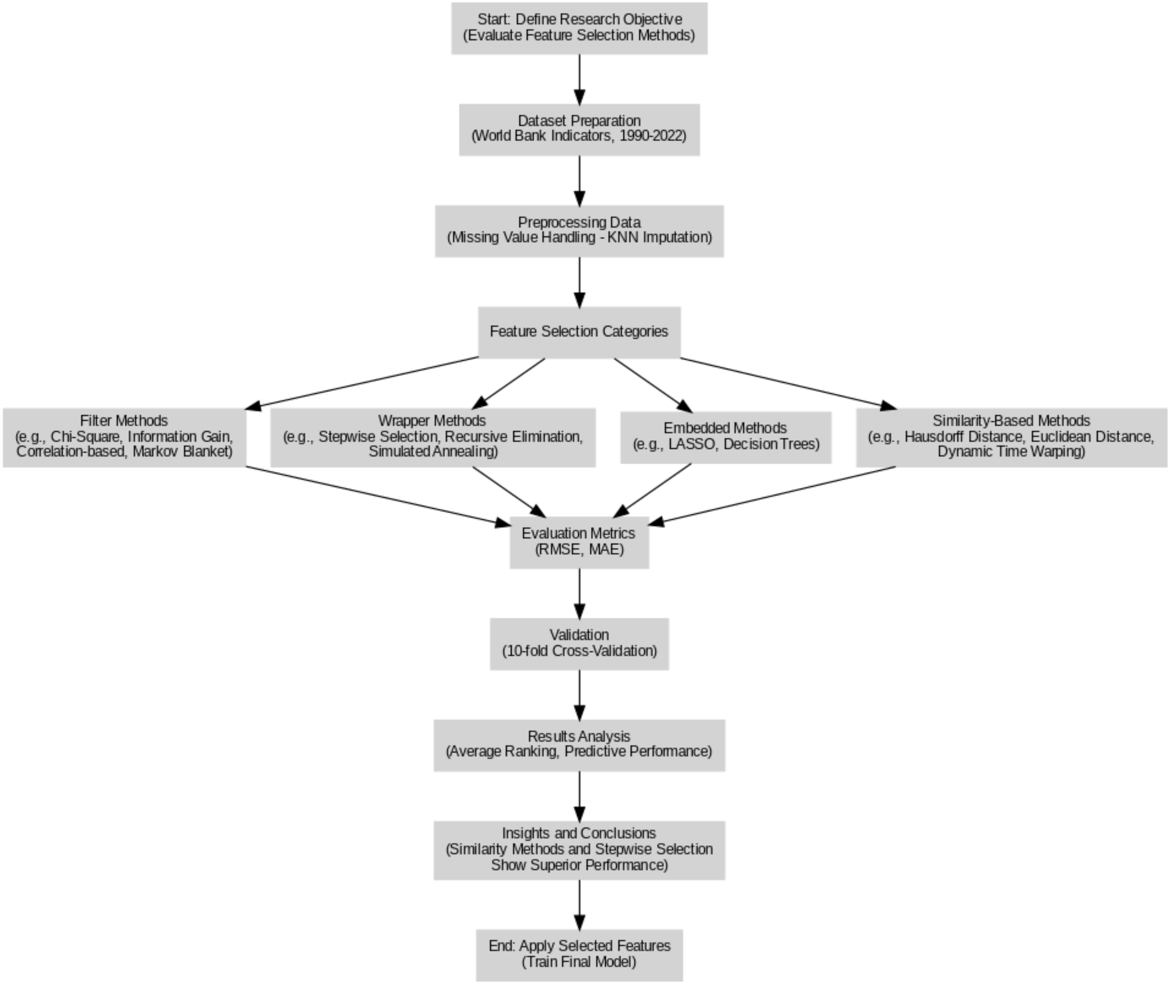

Start: Define Research Objective
(Evaluate Feature Selection Methods)
Dataset Preparation
(World Bank Indicators, 1990-2022)
Preprocessing Data
(Missing Value Handling - KNN Imputation)
Feature Selection Categories
Filter Methods
(e.g., Chi-Square, Information Gain, Correlation-based, Markov Blanket)
Wrapper Methods
(e.g., Stepwise Selection, Recursive Elimination, Simulated Annealing)
Embedded Methods
(e.g., LASSO, Decision Trees)
Similarity-Based Methods
(e.g., Hausdorff Distance, Euclidean Distance, Dynamic Time Warping)
Evaluation Metrics
(RMSE, MAE)
Validation
(10-fold Cross-Validation)
Results Analysis
(Average Ranking, Predictive Performance)
Insights and Conclusions
(Similarity Methods and Stepwise Selection Show Superior Performance)
End: Apply Selected Features
(Train Final Model)

Figure 1. The complete methodology

Table 2. The list of target variables

Variable	Description
Adjusted Savings: Consumption of Fixed Capital	Annual adjusted savings considering fixed capital usage.
Broad Money	Total money supply in the economy.
Food Production Index (2014–2016 = 100)	Measure of food production, base year 2014–2016.
Foreign Direct Investment (Net Inflows as % of GDP)	Net inflows of FDI as a percentage of GDP.
GDP Growth	Annual growth rate of GDP.
General Government Final Consumption Expenditure (% of GDP)	Government consumption as a percentage of GDP.
GNI	Gross National Income.
Gross Domestic Income	Total income generated domestically.
Gross Domestic Saving	National saving as a percentage of GDP.
Gross National Expenditure (% of GDP)	Total expenditure as a percentage of GDP.
Gross Value Added at Basic Prices	Value addition by all sectors at basic prices.
Households and NPISHs Final Consumption Expenditure Per Capita (Constant 2015 US$)	Per capita household expenditure in constant dollars.
Imports of Goods and Services (Constant 2015 US$)	Value of imports adjusted to constant 2015 US$.
Manufacturing Value Added (Annual % Growth)	Annual growth in manufacturing output.
Official Exchange Rate (LCU per US$, Period Average)	Average local currency exchange rate per US dollar.
Stocks Traded (Total Value as % of GDP)	Value of traded stocks as a percentage of GDP.
Total Debt Service (% of Exports of Goods, Services, and Primary Income)	Debt repayment as a percentage of exports.
Unemployment (Total % of the Labor Force, Modeled ILO Estimate)	Total unemployment rate as estimated by ILO.
Wholesale Price Index (2010 = 100)	Index measuring wholesale price levels (base 2010).
Consumer Price Inflation	Annual inflation based on consumer prices.

missing values, thereby preserving the completeness of the dataset while maintaining its statistical validity (16). This approach ensures that the data set is robust and suitable for further analysis.

Conventional feature selection methods

Feature selection (FS) techniques are employed to determine and preserve the most significant and insightful features of the data, ensuring the construction of precise predictive models. The dataset includes many features, leading to the presence of noise, irrelevant details, and redundant information. Hence, this increases the computational time and error rate of the learning algorithm (17). Three main categories of feature selection methods exist: filter, wrapper, and embedded. A brief description of each selection method is given in Table 3.

They become particularly valuable in complex scenarios where neither filter, wrapper, nor can embedded methods alone achieve the desired outcomes.

The proposed approach

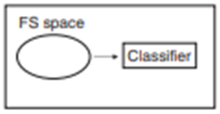
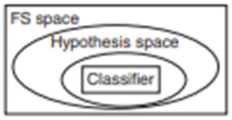
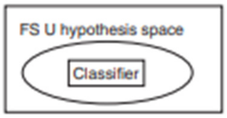
The proposed method falls under Filter techniques, which evaluate feature importance based on their correlation with the target variable. Figure 2 illustrates the framework of the suggested methodology, emphasizing its four main stages.

At the heart of this approach lies the application of similarity measures. This study examines feature selection (FS) by utilizing various distance metrics, including Euclidean Distance, Dynamic Time Warping (DTW), Edit Distance on Real Sequences (EDR), Longest Common Subsequence (LCSS), and Edit Distance with Real Penalty (ERP). These metrics are crucial for assessing the similarity between time series, a fundamental task in the clustering and classifying of temporal data. The primary goal is to determine the distance between two time series, which is vital for analyzing temporal patterns and trends.

In earlier applications, time series similarity was a direct statistical inference tool to uncover relationships between time series originating from different datasets (19). However, with the exponential growth of data collection in recent years, time series data has become increasingly prevalent, leading to a surge in analytical tasks such as regression, classification, clustering, and segmentation. These tasks often hinge on selecting a suitable distance metric to effectively quantify the degree of similarity between time series.

Given the importance of similarity measures, this study explores various methods to determine the distance between time series. These methods are broadly classified into three main categories: stepwise measures,

Table 3. Conventional feature selection methods

	Univariate		
 Filter	- Fast - Scalable - Independent of the classifier	- Ignores feature dependencies - Ignores interaction with the classifier	χ^2 (Chi-square test) Euclidean distance i-test Information gain Gain ratio
	Multivariate		
	- Models feature dependencies - Independent of the classifier - Better computational complexity than wrapper methods	- Slower than univariate techniques - Less scalable than univariate techniques - Ignores interaction with the classifier	Correlation-based feature selection (CFS) Markov blanket filter (MBF) Fast correlation-based feature selection (FCBF)
	Deterministic		
 Wrapper	- Simple - Interacts with the classifier - Models feature dependencies - Less computationally intensive than randomized methods	- Risk of overfitting - More prone than randomized algorithms to getting stuck in a local optimum (greedy search) - Classifier dependent selection	Sequential forward selection (SFS) Sequential backward elimination (SBE) Recursive Feature Elimination
	Randomized		
	- Less prone to local optima - Interacts with the classifier - Models feature dependencies	- Computationally intensive - Classifier dependent selection - Higher risk of overfitting than deterministic methods	Simulated annealing Randomized hill climbing Genetic algorithms Estimation of distribution algorithms
 Embedded	- Interacts with the classifier - Better computational complexity than wrapper methods - Models feature dependencies	- Classifier dependent selection	Decision trees LASSO Feature selection using the weight vector of SVM

which align time series elements sequentially; distribution-based measures, which focus on statistical properties; and geometric methods, which emphasize spatial relationships and patterns. Understanding and leveraging these approaches is essential for advancing time series analysis and enhancing its applications across diverse fields.

Stepwise Metrics

These metrics compare time-series samples one by one based on their time indices (20). A significant limitation of these methods is the requirement for identical sample sizes in the time series. The most notable stepwise metrics are Euclidean Distance and Correlation Coefficient, which are detailed below.

o The Euclidean Distance is the simplest measure for comparing time series. It calculates the shortest distance between two points in Euclidean space using the Pythagorean theorem. The Euclidean Distance between two time series x and y of length n is defined as:

$$Deuc = (\sum_{i=1}^{n}(x_i - y_i))^{1/2} \quad (1)$$

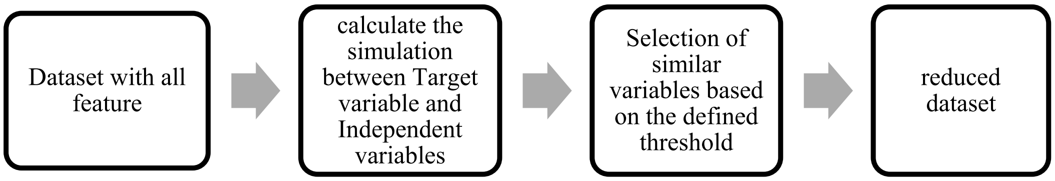

Dataset with all feature
calculate the simulation between Target variable and Independent variables
Selection of similar variables based on the defined threshold
reduced dataset

Figure 2. The framework of the proposed feature selection

This distance is widely used due to its simplicity and ease of understanding. However, a key limitation of Euclidean Distance is its sensitivity to time-axis transformations, such as scaling and shifting (21). Moreover, it cannot compare time series with different sample sizes. As it relies on point-to-point mapping, it is highly sensitive to noise and temporal misalignments, thus making it unsuitable for handling local shifts in time.

A straightforward extension of Euclidean Distance is to calculate the similarity using extracted features rather than raw time-series data.

o Pearson Correlation Coefficient is a widely used metric for assessing the linear relationship between two time series. It is defined as:

$$corr(x,y) = \frac{E(XY)-E(X)E(Y)}{std(X)std(Y)} \qquad (2)$$

The Pearson Correlation Coefficient ranges between -1 and 1, where 1 indicates a perfect positive correlation, and -1 reveals a perfect negative correlation. However, it cannot distinguish between dependent and independent variables or capture non-linear relationships.

Elastic metrics

These metrics adjust the time axis by stretching or compressing it to minimize the effect of local variations. These methods are particularly effective for handling non-linear distortions on time. The most notable elastic methods include Dynamic Time Warping (DTW), Longest Common Subsequence (LCSS), and others.

o Dynamic Time Warping (DTW) is an algorithm for measuring similarity between time series that may vary in speed or timing. Unlike Euclidean Distance, DTW aligns sequences non-linearly by stretching or compressing the time axis to find the optimal alignment. The cumulative distance is calculated as:

$$DISTMATRIX = \begin{bmatrix} d(x_{1,}y_1) & d(x_{1,}y_2) & \dots & d(x_{1,}y_m) \\ d(x_{2,}y_1) & d(x_{2,}y_2) & \dots & d(x_{2,}y_m) \\ d(x_{n,}y_1) & d(x_{n,}y_2) & \dots & d(x_{n,}y_m) \end{bmatrix} \qquad (3)$$

$$\begin{cases} r(i,j) = d(i,j) + \min\{r(i-1,j), r(i,j-1), r(i-1,j-1)\} \\ DTW(x,y) = \min\{r(n,m)\} \end{cases} \qquad (4)$$

DTW allows comparisons between time series of different lengths and identifies similar shapes, even if they are out of phase. However, it is computationally intensive, making it less practical for large datasets.

o Longest Common Subsequence (LCSS) focuses on the longest matching subsequences between two time series while ignoring noise and distortions. For two sequences S_x and S_y of lengths n and m, the similarity is defined as:

$$M(i,j) = \begin{cases} 0 \quad ; \ i = 0 \ or \ j = 0 \\ 1 + M(i-1,j-1) \ ; \ x_i = y_j \ , \ i \geq 1 \ or \ j \geq 1 \\ Max\begin{cases} M(i-1,j) \\ M(i,j-1) \end{cases} ; x_i \neq y_j \ , \ i \geq 1 \ or \ j \geq 1 \end{cases} \qquad (5)$$

Where M (n,m) is calculated recursively:

$$M(i,j) = \begin{cases} 0 \quad ; \ i = 0 \ or \ j = 0 \\ 1 + M(i-1,j-1) \ ; \ (x_i - y_j) \leq \varepsilon \ , \ i \geq 1 \ or \ j \geq 1 \\ Max\begin{cases} M(i-1,j) \\ M(i,j-1) \end{cases} ; \ (x_i - y_j) > \varepsilon \ , \ i \geq 1 \ or \ j \geq 1 \end{cases} \qquad (6)$$

LCSS is robust to noise and suitable for comparing time series with different lengths. However, it heavily depends on the similarity threshold, which impacts its accuracy.

o The edit distance algorithm counts the number of insertion, deletion, and substitution operations required to transform one string into another. It can be applied to time series, where points X and Y match if their absolute distance is less than ε (22). Given two sequences Y, and X, of lengths n and mmm, respectively, the Edit Distance on Real sequence (EDR) between X and Y refers to the number of insertions, deletions, or substitutions required to transform X into Y. It is defined as follows:

$$EDR(X\,,Y) = \begin{cases} n \quad if \ m = 0 \\ m \quad if \ n = 0 \\ min\begin{Bmatrix} EDR\big(Rest(X), Rest(Y)\big) + subcost, \\ EDR(Rest(X), Y) + 1\,, EDR(X, Rest(y)) + 1 \end{Bmatrix} \end{cases} \qquad (7)$$

o ERP, as with the EDR method, is based on Edit Distance (ED) for measuring the similarity of time-series data (23). ERP, accompanied by the L1-norm and Edit Distance, supports local time shifts and is a metric, meaning it satisfies the triangular inequality. Non-metric distance functions complicate problems as violating the triangular inequality renders most indexing structures infeasible. The primary reason why EDR does not satisfy the triangular inequality is that when a gap needs to be added, it repeats the previous element. In contrast, ERP does not face this issue since it uses the L1-norm between two non-gap elements and is designed in such a way that it applies an actual penalty between two non-gap elements. However, it employs a fixed value for calculating the distance for gaps (23). When calculating ERP for two time series S_x and S_y with lengths n and mmm, they are aligned to the same length by adding certain symbols (referred to as gaps). Then, each element in one time series is matched with a gap or an element in another. Finally, the ERP distance between the two-time series S_x and S_y is defined recursively.

$$d_{erp} = \begin{cases} \sum_{i=1}^{m} |x_i - g| & if\ n = 0 \\ \sum_{j=1}^{n} |y_j - g| & if\ m = 0 \\ min \begin{Bmatrix} d_{erp}(Rest(x), Rest(y) + |x_1 - y_1|), \\ d_{erp}(Rest(x), y + |x_1 - g|) \\ d_{erp}(x, Rest(y) + |y_1 - g|) \end{Bmatrix} & \end{cases} \quad (8)$$

o Time Warped Edit Distance (TWED) combines the strengths of DTW and edit distance algorithms by allowing elastic matching with additional constraints. The similarity is measured as the minimum sequence of edit operations required to align two time series.

Geometric distances

Geometric distances focus on the spatial characteristics of trajectories, particularly their shapes. Examples include Hausdorff Distance, Discrete Frechet Distance, and SSPD (Symmetric Segment Path Distance).

o The Hausdorff Distance measures the maximum mismatch between two trajectories, defined as:

$$Haus(X, Y) = Max\{\sup_{x \in X} \inf_{y \in Y} \|xy\|_2, \sup_{x \in X} \inf_{y \in Y} \|xy\|_2\} \quad (9)$$

o Frechet Distance measures the similarity between curves by calculating the minimal "leash length" required to connect a dog and its owner walking along two separate paths. It is mathematically defined as:

$$D_{Frechet}(T^1, T^2) = \min_{w}\{\max_{k \in (0 \ldots |w|)} \|w_k\|_2\} \quad (10)$$

o SSPD shape-based distances such as Hausdorff and Frechet can align with corresponding paths but can not be compared as a unified entity. SSPD is a shape-based distance metric that does not consider the time index of the path. This metric calculates the point-to-segment distance for all samples of the reference path and all segments of the other path then report the average of the obtained distances for the path sample as the SSPD distance (24).

SSPD is defined as follows:

$$D_{SPD}(T^1, T^2) = \frac{1}{n_1} \sum_{i_1=1}^{n_1} D_{pt}(p_i^1, T^2)$$
$$(p_i^1, T^2) = min_{i_2 \in (0 \ldots . n_2 - 1)} D_{ps}(p_{i_1}^1, s_{i_2}^2) \quad (11)$$
$$D_{PT}(P_1^2, T^1) = \min_{i_1 \in (0 \ldots . n_1 - 1)} D_{PS}(P_1^2, S_{i_1}^1)$$

This distance is not symmetric. By considering the average of these distances, SSPD is defined as follows:

$$D_{SSPD}(T^1, T^2) = \frac{D_{SPD}(T^1, T^2) + D_{SPD}(T^2, T^1)}{2} \quad (12)$$

Validation methods

Stationarity is a key principle in time series analysis, defined as the condition where the statistical attributes of a time series, such as its mean, variance, and autocorrelation, remain unchanged over time (25). A stationary time series is essential for reliable analysis and modeling. In the subsequent phase of our methodology, statistical tests were conducted to evaluate significant variations among the reduced datasets.

To perform predictive analysis, a Linear Regression model was selected due to its straightforward nature and ease of interpretation. Nevertheless, alternative regression models may be applied based on the specific requirements of the study. To enhance the reliability of the model evaluation and mitigate the risk of overfitting, a 10-fold cross-validation technique was employed. This method involves splitting the dataset into ten roughly equal parts, with each subset alternately used for training and testing during the evaluation.

Model performance was measured using two key metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). RMSE captures the deviation between predicted and observed values, whereas MAE quantifies the average error magnitude in predictions. The evaluation was carried out across 10 iterations, generating unique RMSE and MAE scores for each run. This iterative approach ensured the robustness and consistency of the results, providing a comprehensive validation of the methodology.

Results and Discussion

In this study, we present the results of predicting performance across 14 datasets, each selected using a different feature selection (FS) technique. These techniques include seven filter methods, five wrapper methods, three embedded methods, and four similarity-based methods. The similarity methods as FS techniques are also evaluated within this framework. The chosen methods were selected for their widespread recognition in literature, allowing for a clear comparison. To assess predictive accuracy, we use two evaluation metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), applied to the performance of a Linear Regression model. To evaluate the efficiency of each dataset selected by the FS methods, we implemented the techniques on the World Bank dataset, which includes

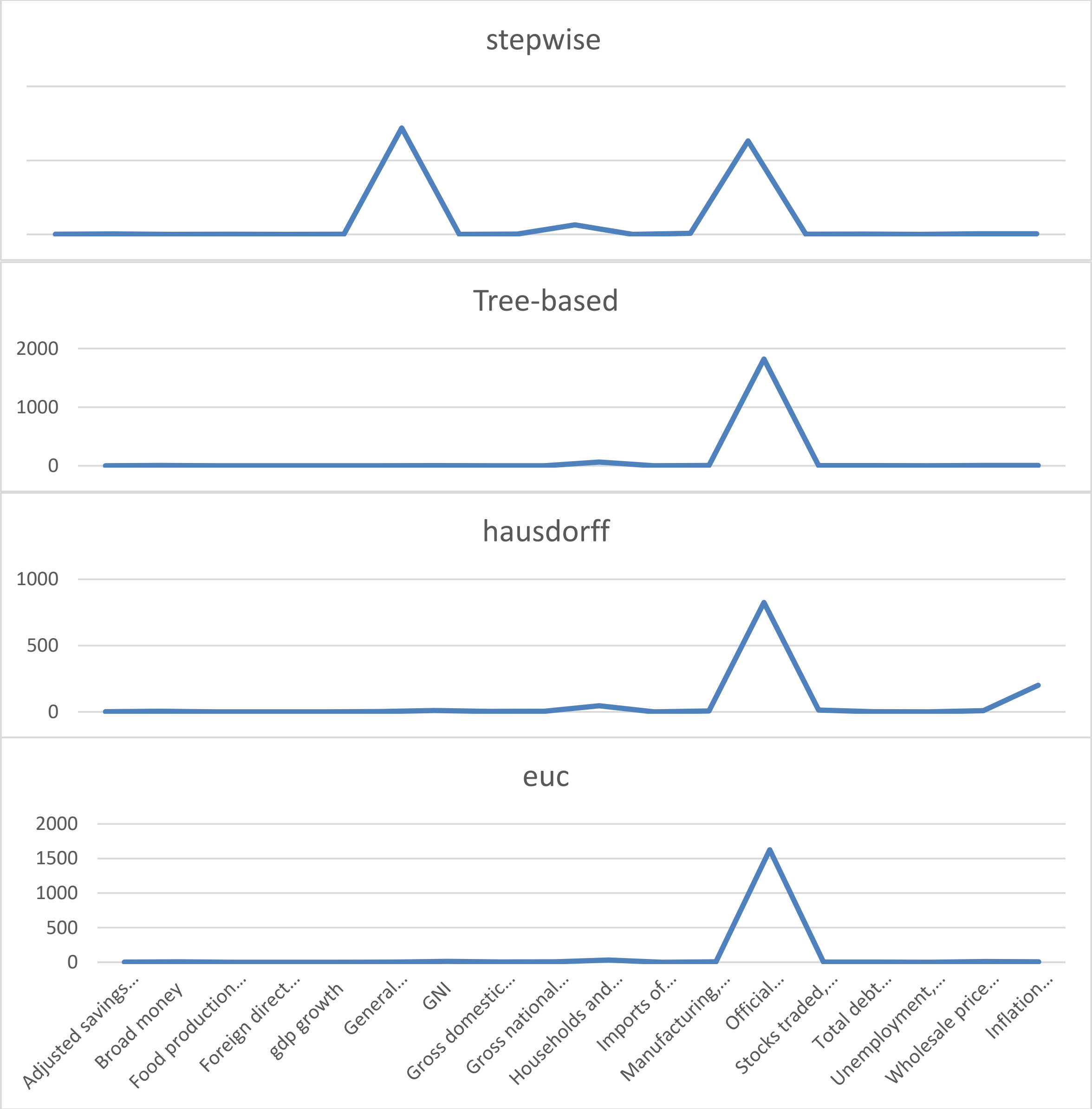


Figure 3. The top four feature selection models based on 14 datasets chosen by Feature selection techniques

Table 4. Average Mean Absolute Error (MAE) of datasets

Category	Methods	Average
Wrappers	stepwise	32/0299
similarity	frechet	51/6163
similarity	hausdorff	62/68829
similarity	sspd	91/70364
similarity	epr	91/88632
similarity	dtw	91/93176
similarity	euc	95/02939
Embedded	Tree-based	106/3909
Wrappers	recursive	270/5572
similarity	lcsso	292/8808
similarity	edr	298/4402
Filters	MI_Score	963/5397
Filters	inf	1683/06
similarity	Sparse	3/98E+08
Wrappers	forward	6/4E+08
Wrappers	simulated_annealing	8/13E+08
Filters	fisher	1/83E+09
Embedded	lasso	3/06E+12
Filters	chi	4/83E+13
Filters	corrolation	4/83E+13
Filters	data_dispersion	8/16E+13
Filters	var	6/41E+14
Wrappers	backward	6/47E+14

various target variables. In total, 20 different datasets were used, and FS methods were employed to identify the best feature subsets from each.

Figure 3 illustrates the results of a 10-fold cross-validation evaluation for each FS method. The datasets selected by these four methods consistently exhibited the lowest RMSE and MAE, indicating superior predictive accuracy.

Table 4 presents the average MAE values for datasets processed using various feature selection (FS) methods. The Mean Absolute Error (MAE) averaged across 20 datasets for each target variable.Those derived using the stepwise feature selection approach demonstrated superior predictive accuracy among the subsets generated. These subsets consistently exhibited the smallest MAE values compared to others. Following closely were the subsets identified through similarity-based techniques, which also achieved notably low average MAE scores, underscoring their effectiveness in prediction tasks.

Figure 4 indicates the average ranking of MAE selected based on FS methods. The ranking of each feature selection (FS) method was determined based on its ability to select the best subset of datasets with the lowest Mean Absolute Error (MAE). To provide a comprehensive analysis, the rank of each of the 20 datasets across all FS methods was averaged. According to the results, the best predictive accuracy methods were Stepwise Selection, Tree-based methods, Hausdorff, Euclidean (Euc), and MI_Score. In contrast, Recursive Feature Elimination with Cross-Validation (RFECV) and Variance Thresholding exhibited the poorest performance.

The average ranking across the feature selection categories (Figure 5) indicates that, on average, similarity-based methods outperformed the other approaches. Specifically, similarity methods achieved an average rank of 9.125, highlighting their superior performance in selecting the most relevant feature subsets compared to other methods.

The results underscore the potential of similarity-based methods as viable alternatives to traditional feature selection techniques, with implications for a wide range of applications, particularly macroeconomic forecasting.

Effectiveness of Similarity-Based Approaches

The strong performance of similarity-based methods, particularly Frechet and Hausdorff distances, demonstrates their ability to identify features that exhibit high relevance to target variables. These methods leverage the inherent structure of time series data, effectively capturing relationships that might be overlooked by traditional approaches. For instance, the Frechet Distance, which accounts for the shape and continuity of data trajectories, excels in handling time series with local distortions, while the Hausdorff Distance, which measures the greatest distance between points of two datasets, is robust against outliers and

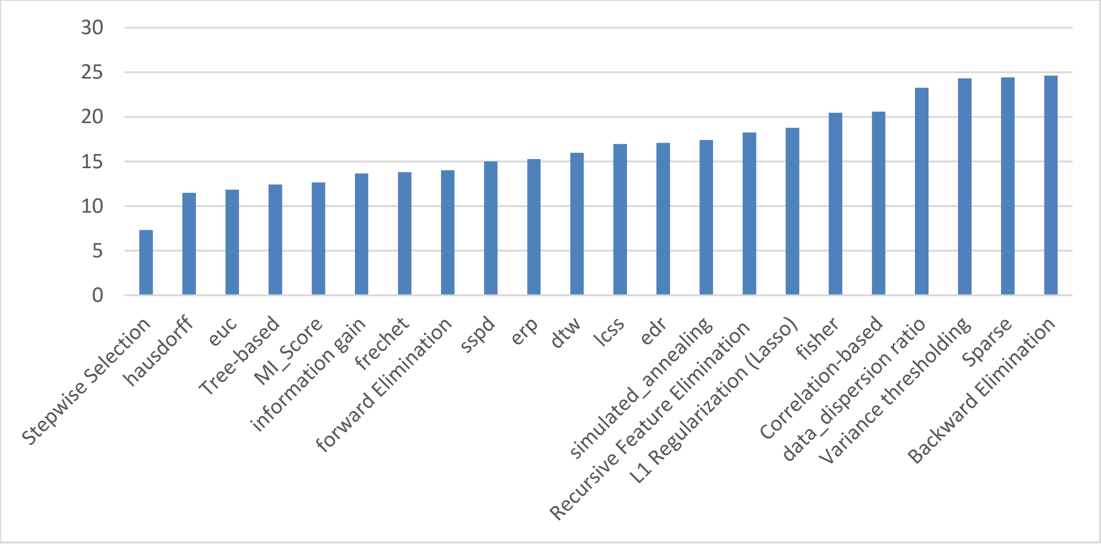


Figure 4. The average ranking of MAE selected based on FS methods

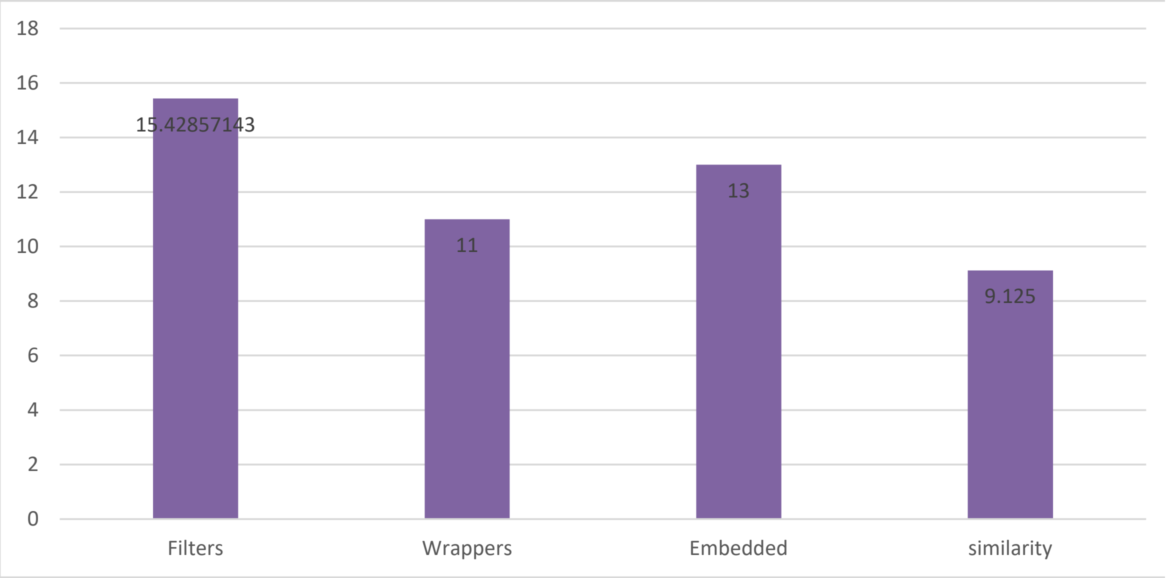


Figure 5. The ranking of the category of feature selection methods

noise.

This capability aligns with clustering and classification literature findings, where similarity measures are frequently employed to quantify relationships between data points. By applying these measures to feature selection, this study extends their utility into a new domain, validating their effectiveness in identifying subsets of features that enhance model performance. Further, their simplicity and computational efficiency make similarity-based methods suitable for real-world scenarios where quick and accurate analysis is critical.

Comparison with Traditional Methods

Traditional feature selection methods, such as Stepwise Selection and Tree-based approaches, remain benchmarks in the field due to their consistent performance and well-established methodologies.

Stepwise Selection, in particular, excels in identifying key features through iterative inclusion or exclusion, making it a preferred choice for many predictive modeling tasks. Similarly, Tree-based methods, such as Random Forests, offer an embedded mechanism for ranking features by their importance, balancing accuracy and interpretability.

However, similarity-based methods emerge as strong contenders, offering a computationally efficient alternative especially advantageous in high-dimensional scenarios. Unlike traditional methods that often rely on iterative testing or classifier-specific criteria, similarity-based approaches operate independently of classifiers, enabling faster preprocessing and reducing the risk of overfitting. This makes them particularly appealing for datasets with numerous variables, where computational resources and time constraints are significant considerations.

Implications for Macroeconomic Forecasting

Macroeconomic forecasting heavily relies on accurate predictions of key indicators, such as GDP growth, inflation rates, and unemployment levels. The use of similarity-based methods in this context provides several advantages:

✓ Simplification of Preprocessing: By directly measuring the relationship between features and target variables, similarity-based methods eliminate redundant preprocessing steps. This simplifies the pipeline and lowers the risk of introducing errors during data preparation.

✓ Enhanced Interpretability: The straightforward nature of similarity measures, such as distances or correlations, allows for easier interpretation of results. Policymakers and economists can gain clearer insights into which features drive predictions, thus facilitating more informed decision-making.

✓ Robust Forecasting Tools: By focusing on the most relevant features and minimizing noise, these methods contribute to developing robust and reliable forecasting models. This is particularly critical for policymaking, where accurate predictions can guide interventions and resource allocation.

Conclusion

In this study, we investigated which feature selection (FS) and similarity methods most effectively enhance the predictive performance of models for various macroeconomic variables. The analyzed indicators included a diverse range of metrics, such as adjusted savings (consumption of fixed capital), broad money, the food production index, imports of goods and services (constant 2015 US$), manufacturing value-added (annual % growth), official exchange rate (LCU per US$), stocks traded (total value as a % of GDP), total debt service (% of exports), unemployment (% of the total labor force, ILO estimate), the wholesale price index (2010 = 100), and consumer price inflation. To achieve this, we evaluated 23 different FS and similarity methods to identify the most effective techniques for selecting features that provide accurate predictions of these macroeconomic indicators.

Time series similarity algorithms, though rarely utilized as standalone feature selection methods, were a key focus of this research. By comparing these algorithms against traditional FS approaches, we aimed to assess their potential in identifying relevant features. Each FS and similarity method was applied to the datasets, and their performance was evaluated using both MAE and RMSE metrics. The current findings are hence in agreement with the studies of Zhu et al. and Mitra, who applied the methods of similarity measures for feature grouping and selection to increase clustering performance. Additionally, robustness from similarity metrics obtained herein further supports conclusions from Mehri et al. and Goldani and Asadi, demonstrating their viability in high-dimensional and financial forecasting setups. This work extends these approaches toward macroeconomic forecasting, hence addressing an important lacuna in the related literature. Besides, the traditional feature selection methods, such as stepwise selection and tree-based methods, were confirmed to be reliable benchmarks, which agrees with the results obtained by Jović et al. However, the similarity-based methods were their strong competitors, providing equal or higher predictive accuracy with computational simplicity. Unlike other methods, such as Recursive Feature Elimination and Variance Thresholding, which did not perform well in our analysis, results consistent with the critiques of Guyon similarity-based approaches provided a more robust alternative for high-dimensional datasets. Findings revealed that methods such as Stepwise Selection paired with Tree-based techniques, Hausdorff distance, Euclidean distance, and Mutual Information Score consistently outperformed other approaches, demonstrating higher predictive accuracy. Conversely, methods like Recursive Feature Elimination with Cross-Validation and Variance Thresholding showed comparatively weaker results, suggesting limited utility in this context. These results highlight the potential of similarity-based algorithms as effective tools for feature selection in macroeconomic forecasting.

By systematically comparing these methods with established feature selection techniques across 20 datasets of macroeconomic indicators, the key findings were obtained as follows:

Performance of Similarity-Based Methods:

- Similarity-based methods, particularly Frechet and Hausdorff distances, demonstrated strong performance in identifying relevant features, with competitive Mean Absolute Error (MAE) values compared to traditional techniques.
- The computational efficiency and robustness of similarity-based methods make them suitable for high-dimensional datasets, offering an alternative to Filter, Wrapper, and Embedded methods.

Advancing Feature Selection:

- Traditional approaches such as Stepwise Selection and Tree-based methods remain benchmarks thanks to their high accuracy and established methodologies. However, similarity-based methods provide a complementary approach, particularly in applications requiring computational simplicity and adaptability.

Macroeconomic Implications:

- The adoption of similarity-based feature selection can improve forecasting accuracy for critical economic indicators such as GDP growth, inflation, and unemployment. These tools enhance interpretability and simplify preprocessing, making them valuable for policymakers and economic analysts.

Studies could explore hybrid models that integrate similarity-based techniques with traditional feature selection frameworks to leverage the strengths of both approaches. For example, combining similarity measures with Wrapper methods could further boost accuracy while maintaining computational efficiency. Since the performance of similarity-based methods depends on the choice of distance metrics, research should focus on developing adaptive or data-driven methods for selecting optimal metrics based on dataset characteristics.

Adopting similarity-based feature selection methods significantly advances macroeconomic forecasting and policy analysis. These methods would improve the accuracy and efficiency of models while maintaining transparency and interpretability. By prioritizing adopting and developing these techniques, policymakers can make more informed decisions, better allocate resources, and enhance their ability to respond to economic challenges. Future efforts should focus on refining these methods, scaling their use across various domains, and integrating them into comprehensive, real-time forecasting systems to support dynamic and effective policymaking.

References

1. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. ACM Comput Surv. 2017; 50(6):1–45.
2. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019; 112:103375.
3. Zhu Z, Ong YS, Dash M. Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans Syst Man Cybern. 2007; 37(1):70–6.
4. Mitra P, Murthy CA, Pal SK. Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell. 2002; 24(3):301–12. doi:10.1109/34.990133
5. Shi J, Wang B, Shi Q, et al. Adaptive-similarity-based multi-modality feature selection for multimodal classification in Alzheimer's disease. Med Image Anal. 2020; 60:101618.
6. Mehri M, Chaieb R, Kalti K, Héroux P, Mullot R, Essoukri Ben Amara N. A comparative study of two state-of-the-art feature selection algorithms for texture-based pixel-labeling task of ancient documents. J Imaging. 2018; 4(8):97.
7. Shen Z, Chen X, Garibaldi JM. A novel meta-learning framework for feature selection using data synthesis and fuzzy similarity. In: 2020 IEEE Int Conf Fuzzy Syst (FUZZ-IEEE). 2020. p. 1–8.
8. Goldani M, Tirvan SA. Sensitivity assessing to data volume for forecasting: introducing similarity methods as a suitable one in feature selection methods. arXiv preprint arXiv:2406.04390. 2024.
9. Mathisen BM, Aamodt A, Bach K, Langseth H. Learning similarity measures from data. Prog Artif Intell. 2020; 9(2):129–43.
10. Qi M, Wang T, Liu F, Zhang B, Wang J, Yi Y. Unsupervised feature selection by regularized matrix factorization. Neurocomputing. 2017; 273:593–610.
11. Du S, Ma Y, Li S, Ma Y. Robust unsupervised feature selection via matrix factorization. Neurocomputing. 2017; 241:115–27.
12. Hu R, et al. Graph self-representation method for unsupervised feature selection. Neurocomputing. 2015; 220:130–7.
13. Venkatesh B, Anuradha J. A review of feature selection and its methods. Cybern Inf Technol. 2019; 19(1):3–26.
14. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–82.
15. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2015. p. 1200–5.
16. Goldani M. Comparative analysis of missing values imputation methods: a case study in financial series (S&P500 and Bitcoin value data sets). Iran J Finance. 2024; 8(1):47–70.
17. Ali M, Mazhar T, Shahzad T, Ghadi YY, Mohsin SM, Akber SMA, et al. Analysis of feature selection methods in software defect prediction models. IEEE Access. 2023.
18. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19):2507–17.
19. Chen H, GAO X. A new time series similarity measurement method based on fluctuation features. Tehnički Vjesnik.

2020; 27:1134–41.

20. Salarpour A, Khatunloo H. A segmental distance-based similarity criterion using time deviation. J Electr Eng Univ Tabriz. 2019; (2):645–56.

21. Keogh E, Pazzani M. Derivative dynamic time warping. In: Proceedings of the 2001 SIAM International Conference on Data Mining. 2001. p. 1–11. doi:10.1137/1.9781611972719.1

22. Besse PC, Guillouet B, Loubes JM, Royer F. Review and perspective for distance-based clustering of vehicle trajectories. IEEE Trans Intell Transp Syst. 2016; 17(11):3306–17.

23. Chen L, Ng RT. On the marriage of Lp-norms and edit distance. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB). 2004. p. 792–803.

24. Besse PC, Guillouet B, Loubes JM, Royer F. Review and perspective for distance-based clustering of vehicle trajectories. IEEE Trans Intell Transp Syst. 2016; 17(11):3306–17.

25. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci. 2012; 191:192–213.