Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Extraction New Features

Nidhal Hasan Hasaan[a], Lamia Abed Noor Muhammed[b]

AL-Qadisiyah University, College of Computer Science and Information Technology, Computer Science Department, Diwaniyah, Iraq. Email: cm.post23.13@qu.edu.iq

AL-Qadisiyah University, College of Computer Science and Information Technology, Computer Science Department, Diwaniyah, Iraq.Email: lamia.abed@qu.edu.iq

ARTICLE INFO

Article history:
Received: 4 /1/2025
Rrevised form: 10 /1/2025
Accepted : 13 /1/2035
Available online: 30 /3/2025

Keywords:

Extraction Features

ABSTRACT

Feature extraction is a key part of machine learning, aiming to transform raw facts into more representative and effective features for models. This process involves selecting features that facilitate classification or prediction tasks. Feature extraction helps reduce the dimensions of the data while keeping basic information, which contributes to reducing computational complexity, improving the accuracy of predictions, and increasing the efficiency and speed of models. Common feature extraction methods include several methodologies, including statistical techniques such as calculating means, standard deviation, and variance, principal component analysis (PCA) to reduce dimensions while preserving as much variance in the data as possible, and independent factor analysis (ICA) is used to separate mixed signals and extract statistically independent features, and using advanced techniques such as linear discriminant analyses (LDA) and autoencoders to extract important features. In addition, clustering techniques such as K-Means also play an important role in identifying hidden patterns in data by grouping them into clusters and then using the properties of these clusters as features.

Feature extraction is an essential process for improving the effectiveness of models in terms of performance. It is indispensable in various analytical applications, such as medical diagnosis, image analysis, fraud detection, voice recognition, and text analysis.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.11960

1. Introduction

With an increase in the volume of data available across various domains [1], feature extraction is transforming raw data into more useful features that enhance the performance and accuracy of models and algorithms in machine learning, feature extraction has become one of the most critical steps that directly affect the effectiveness of models [2], Since Primary data often contains unnecessary, noisy, or irrelevant information, feature extraction reduces data

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

complexity while retaining only the essential information. This process is not only key to reducing data size but also enhances classification and prediction accuracy, and reduces processing time[3].
The necessity of feature extraction lies in the increase in accuracy and efficiency of the Templates used in machine learning [4], it also helps with Dimensionality Reduction: where Feature extraction helps prune the size of data by selecting important characters, which decreases computational complexity and improves the speed of training and testing models[5]. also, Reduction of Redundancy and Overlap: where Feature extraction contributes to eliminating redundant or overlapping features that may negatively impact model performance[6], In addition, it Contributes to Increased Model Accuracy: where Feature extraction helps identify relevant characteristics, enabling the model to handle data more efficiently, and Overfitting: By reducing features to the essential features, It will decrease the likelihood of the model learning unhelpful details or noise in the training data decreases. This reduces the chance of the model becoming overfitted, thereby improving its performance on new test data [7].
Not all features are of high quality or beneficial for classification or prediction[8]. Some features may have lower quality, such as Redundant or Overlapping Features, or Irrelevant Features, there are types of features Continuous Features These are features that can take any numeric values such as (age, weight, temperature, and income). and Categorical Features These features take a specific set of discrete values or categories. These values are often textual or symbolic and express categories or groups. Examples: gender, income class. Therefore, I must extract new features to represent the original data better[9].

They used several feature extraction methods, the most prominent of which is, Factor Analysis (FS)[10], Principal Component Analysis (PCA)[10] [11] [12], Linear Discriminant Analysis (LDA)[10][13][14][15], Locally Linear Embedding (LLE)[11], and t-SNE [11], Independent Component Analysis (ICA)[16][17] Kernel Principal Component Analysis(KPCA)[18], [19], Probabilistic Principal Component Analysis(PPCA)[20], Fisher discriminant analysis[21], Statistical methods (such as calculating the mean, standard deviation, and variance)[22] [23], convolutional autoencoders[24], Autoencoders [25],(RBM)Restricted Boltzmann Machine[26] , Hybrid methods that combine the advantages of different technologies to deliver better results [27][28][29][30][31][32]. they used clustering as a means of extracting features [33][34][35] [36] [37]

Each algorithm or method has advantages and disadvantages that make it suitable for some applications and not others. This balance between limitations and benefits encourages researchers to constantly find new methods or develop existing algorithms to improve accuracy and performance and obtain better results.

This paper has been formatted, in part 2, types of feature extraction methods. In Part 3, the related work contains previous works on different methods and techniques for feature extraction. In part 4, the discussions are presented, and in part 5, the conclusion.

2. Methods

Feature extraction is the extraction of potentially hidden Patterns from abundant, incomplete, noisy, ambiguous, and random data[38], It is a conversion process of raw data into a representation containing useful and necessary information about the problem the model is trying to solve in machine learning[39], The original data is often complex or contains redundant information and feature extraction comes into play to identify the most important features, thus helping the model learn from them[40] . The relationship between machine learning and feature extraction is solid, as the value of the model's performance depends on the quality of its given features: This relationship has several basic aspects. It improves the accuracy and efficiency of the model, reducing dimensionality, accelerating the learning process, resisting noise, and reducing the overfitting problem[41].

One of the most common methods is:

2.1.(PCA) Principal Component Analysis

It is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional form, the steps to implement it:

1-Standardization Principal Component Analysis (PCA) is sensitive to the data size so that each feature has a mean of 0 and a standard deviation of 1.

2- The variance/covariance matrix is calculated, which describes the variation between different features in the data. This matrix is calculated to determine the relationship between these features.

3- we calculate the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors represent the new directions explaining the greatest variation in the data, while the eigenvalues represent the variation along each principal component.

4- Order the eigenvectors by their corresponding eigenvalues from largest to smallest. The eigenvectors associated with the largest eigenvalues represent the principal components that account for the most variance.

5-Transform Data Project the original data onto the selected principal components to obtain the reduced-dimensional representation [10][11][13][14][12].

2.2. Linear Discriminant Analysis (LDA):

It is a technique used for dimension reduction and improved classification. LDA aims to find a line that separates multiple classes in a dataset [10][14] [15]

2.3. Factor analysis (FA):

statistical method used to understand the structure of relationships among a set of variables through dimensionality reduction. FA aims to identify underlying factors that explain variance in data [10].

2.4. Locally Linear Embedding (LLE):

It is a nonlinear dimensionality reduction technique. aims to project high-dimensional data onto a lower-dimensional space while preserving the local structure of the data [11].

2.5.T-SNE (t-Distributed Stochastic Neighbor Embedding):

is a dimensionality reduction method commonly used in high-dimensional data analysis. t-SNE aims to represent data in low dimensions (often 2 or 3) while preserving local structure and neighboring data well. t-SNE [11].

2.6. Component Analysis (ICA):

It is a technique that separates independent signals from a set of mixed signals. ICA is commonly used in signal processing [13][16][17][42]

2.7. Kernel Principal Component Analysis (KPCA):

It is a dimensionality reduction technique based on the classical principal component analysis (PCA) method with the introduction of a kernel function to handle high-dimensional data and nonlinear structure. KPCA provides an efficient way to extract low-dimensional features from nonlinear data [18].[19]

2.8.Probabilistic Principal Component Analysis (PPCA):

It is used for dimensionality reduction which is useful in dealing with high-dimensional data. PPCA combines the concept of principal components (PCA) with a probabilistic model, allowing for the estimation of data uncertainty and reconstruction of missing data [20].[43]

2.1. Fisher's Discriminant Analysis (FDA):

It is a statistical technique used in classification and dimensionality reduction and it seeks to identify a linear combination of properties that can be used to discriminate between two or more groups [21].

2.10. MRE (Mean, Relative Amplitude, and Entropy):

It is a statistical approach used in feature extraction, and its components: Mean: Represents the average value of a dataset. Relative Amplitude: Measures the proportion of the signal's amplitude relative to some reference or baseline [22].

2.11. Convolutional Autoencoder (CAE):

It is a special type of autoencoder that uses convolutional neural networks. It is commonly used in image processing, where it takes advantage of the spatial properties of the data [24].

2.12. Autoencoder (AE):

It is a type of neural network used in unsupervised learning and aims to learn a compact representation of data. An autoencoder consists of two main parts: an encoder and a decoder. Here are some key points about an autoencoder [25] [44]

2.13. (RBM)Restricted Boltzmann Machine:

It is used in feature extraction tasks and is a building block of deep neural networks such as deep generative networks and convolutional neural networks. Originally developed by Geoffrey Hinton, RBM consists of two layers: a visible layer and a hidden layer, with connections between these layers only (i.e., there are no connections between units in the same layer) [26][45].

2.14. Clustering
A process that aims to discover hidden patterns in data sets, and it consists of grouping data elements into separate groups so that the data in each group is similar and different from the data in the other groups [35][46][47].

3. literature survey

Feature extraction is the process of transforming raw data into efficient representations that help improve the performance of models and reduce computational complexity. Many feature extraction techniques have been developed over time, varying according to the type of data and the nature of the applications. Among the most popular traditional methods are:

Ruhul Amin et al. [2023][10], proposed an integrated projection-based statistical feature extraction method that combines (PCA), (LDA), and (FA). This method improves classification accuracy by addressing feature redundancy, missing values, and outliers. The UCI dataset for patients with Indian Liver Disease (ILPD) was used. The model was evaluated using different machine learning classifiers (e.g. Random Forest, K-nearest Neighbors Support machine vector, multi-layer perceptron). The proposed method performed better than other approaches, the Random Forest classifier achieved the highest performance, with 88.10% accuracy, 85.33% precision, 92.30% recall, F1 score of 88.68 %, and AUC of 88.20%. The limitation of this paper run time of the model is longer than some existing methods. The study was also limited to statistical techniques based on projection.

Ashir Javeed et al. [2024][11], studied integrated systems combining (LLE, PCA, ICA, and TSNE) with ML classifiers (LR, DT, KNN,) for depression reduction. The dataset used in this study comes from the (SNAC)Swedish National Study on Aging and Care. The combination of logistic regression with PCA achieved the highest accuracy of 89.04% for depression classification, outperforming other feature extraction methods such as ICA, LLE, and TSNE. Also, TSNE showed a good performance in combination with the naive Bayes model. Md Shamim Reza and Jinwen Ma [2016] [13], proposed two feature extraction methods that integrate PCA and ICA to create a set of features. In the first method, PCA is applied to the data, followed by ICA. In the second method, both PCA and ICA feature. the following

datasets (simulated, wine, breast cancer, crab) were used. The IC-PC method achieved the highest classification accuracy, outperforming traditional methods like PCA and ACI, particularly in the classification of breast cancer and wine. However, the limitation of this study is that the performance depends on the type of data used. In addition, ICA requires ordering independent components, which can be a complex process.

Shifei Ding et al. [2012][14], proposed various feature extraction methods, from classical approaches like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to more advanced non-linear techniques such as Independent Component Analysis (ICA) and kernel-based methods that transform non-linear problems into linear ones. It also explores improvements using information theory and artificial neural networks. artificial datasets with varying complexity and dimensionality (e.g. Four-gauss, Easy doughnut, Difficult doughnut), and Three real datasets (Wine, Iris, and Breast Cancer Wisconsin Diagnostic (WDBC) datasets were used. Non-linear methods such as Independent Component Analysis (ICA) and Kernel Principal Component Analysis (KPCA) demonstrated superior performance when handling non-linear data, preserving important features while reducing dimensionality. Many advanced feature extraction techniques assume statistical independence or linearity, which may not hold for all datasets. Additionally, some methods are computationally expensive.

Nojun Kwak and Chong-Ho Choi [2003][16], Proposed The authors a modified ICA algorithm for feature extraction in binary classification problems. Their method appends class label information to the ICA process, allowing features that contain relevant information to be extracted for classification. The article uses various synthetic and real datasets, including the IBM and UCI datasets, datasets such as Wisconsin Breast Cancer, and the Sonar Target dataset. The ICA-based method outperformed PCA and traditional ICA in terms of classification accuracy. The method is currently limited to binary classification problems. Furthermore, the algorithm's reliance on ICA may be limited by the statistical properties of the data.

Nojun Kwak et al. [2001][17], proposed a new feature extraction algorithm using independent component analysis (ICA) for supervised classification problems, it incorporates the result class information together with the input features to extract new independent and relevant features for the classification task, this method lessens the dimensionality of the feature while maintaining the classification accuracy. This study was applied to the data set (Chess End game, breast cancer). The extracted features are compared with the basic features and the features extracted by the MIFS-U algorithm, The classification performance is tested using the C4.5 and multi-layer perceptron (MLP) algorithms. The proposed ICA-based method achieved a classification accuracy of 95.42% with only one extracted feature (breast cancer data), achieving an accuracy of 98.80% with only three extracted features (chess data). The limits of this method the algorithm is designed for two-class classification problems, the success of the method depends on the selection number of parameters, such as the threshold for shrinking small weights.

Yong Xu et al. [2007] [18], proposed an improved KPCA (IKPCA) method that speeds up feature extraction by approximating feature extractors as linear combinations of a smaller subset of training samples (called nodes). This reduces the number of kernel functions that must be calculated and thus improves efficiency. The method was tested on four datasets (Splice, Diabetes, Banana, Cancer). IKPCA outperformed KPCA in feature extraction efficiency, for example, in the "Splice" dataset, IKPCA had a Significantly lower classification error rate (17.8%) compared to KPCA (21.8%) with fewer feature extractors and improved speed. The method relies on the empirical determination of the number of nodes, which may affect the overall efficiency, and the node selection process in IKPCA is computationally expensive.

Padathala Visweswara Rao et al. [2024][20], proposed a method that uses Probabilistic Principal Component Analysis (PPCA)for selection and feature extraction to reduce dimensionality and enhance the accuracy of cardiovascular disease classification models. The method was tested on multiple datasets (CHDS), (SHDS), and (SAHDS). The PPCA-based approach outperformed traditional methods like PCA and ReliefF in terms of classification metrics (Accuracy 98.6% (PPCA), 85.4% (PCA), and 80.3% (ReliefF). Precision: 97.5% (PPCA), 83.8% (PCA), 79.1% (ReliefF). Specificity: 97.2% (PPCA), 82.9% (PCA), 81.6% (ReliefF). One limitation of the method is its reliance on the assumption that features contribute to the classification.

Xiaoming Wang and Hong Peng [2012][21], proposed a technique that depends on the MCVSVM algorithm, which combines the strengths of SVM and FDA to provide better class separation and feature extraction by considering both boundary and distribution samples. The proposed method is tested on the Wine dataset. The proposed method outperformed MMDA in projecting the dataset into a lower-dimensional space while maintaining a smaller scatter for same-class samples. The limits of this method are the paper evaluates the technique using a small dataset (Wine dataset) with two classes, which may limit the generalizability of the results to larger and more complex datasets.

Olayinka Ogundile, et al. [2024][22], use an Ensemble Hidden Markov Model with two methods (PCA) and MRE (Mean, Relative Amplitude, and Entropy)), These techniques are used to improve fraud detection accuracy and reduce computational complexity. A publicly available credit card transaction dataset from Kaggle was used, The PCA-EHMM showed significant performance improvement, and the MRE-EHMM achieved similar performance with

much lower computational complexity. One of the limitations of this method is that it focuses on feature extraction techniques without exploring other machine learning models that might further optimize computational efficiency or prediction accuracy.

Wei Li, et al. [2015] [23], proposed a novel statistical feature extraction and evaluation method. The extraction procedure is based on the central limit theorem, which improves classification accuracy. It evaluates the statistical features using a decoupling technique, providing an analytical guarantee for classification accuracy in terms of (FCR). Data was collected from a machinery fault simulator under five different health conditions (normal, rotor imbalance, and three bearing faults). Using the proposed method classification accuracy was significantly improved over traditional ANN and SVM methods, achieving a classification accuracy of 100%, while ensuring a false classification rate (FCR) of less than 0.3%. While the method improves accuracy and provides analytical guarantees, its computational load is higher due to the calculation of statistical features.

Zahra Salek Shah Rezaee et al. [2023][24], proposed this study investigates the impact of feature extraction and data sampling techniques to handle class imbalance and high dimensionality. It evaluates two feature extraction methods Convolutional Autoencoder (CAE) and (PCA) and three data sampling techniques (RUS, SMOTE, and SMOTE Tomek) using four classifiers: Random Forest, XGBoost, LightGBM, and CatBoost. The Credit Card Fraud Detection Dataset from Kaggle was used. the combination of Random sampling (RUS) and Convolutional Autoencoder (CAE) yielded the highest F1 score of 0.909 and an AUC of 0.988 when using LightGBM and XGBoost, respectively. The main limitation was the increased computational cost associated with the use of CAE.

Simon Cramer et al. [2022][25], presented a comparison between two feature extraction methods (PCA) and autoencoders. PCA reduces linear correlations, while autoencoders handle both non-linear and linear correlations. with two regression models: Support Vector Regression (SVR) and a feed-forward neural network (FFNN). The proposed method has been applied to samples from industrial production processes. The combination of PCA and SVR yielded the best performance, with the lowest MAPE (2.91%). This combination proved efficient in removing feature correlations and producing accurate predictions with relatively low computational complexity. The limitation of this method's autoencoder's performance in combination with SVR was unstable, It was expected that the combination of autoencoder and FFNN would yield better results, but was outperformed by PCA-SVR due to the complexity and computational cost.

Jianxin Zhu, et al. [2024][26], The authors propose an improved multilayer Restricted Boltzmann Machine (RBM) for feature extraction. This method uses the reconstruction error to determine the optimal number of RBM layers and employs a weighted pruning approach to remove redundant nodes and improve computational efficiency while maintaining accuracy. I applied the method to the Kaggle Credit Scoring Dataset, the China Union Pay (CUP)dataset. The improved RBM model significantly reduced data dimensionality and achieved high classification performance. After pruning, the model maintained high AUC, accuracy, and recall, particularly with 10% and 20% pruning ratios. the limit of the method is that over-pruning can degrade performance.

Changsheng Zhu et al. [2019][27], proposed a model that integrates PCA for dimensionality reduction with K-means clustering and logistic regression for improved prediction accuracy. PCA is used to enhance clustering performance by determining better initial centroids, reducing redundancy, and improving the logistic regression classification. Was used The Pima Indian Diabetes dataset. The proposed model achieved an accuracy of 97.40% using PCA and K-means, outperforming other models such as K-means and logistic regression alone. The evaluation included k-fold cross-validation, confusion matrix, ROC curve analysis, and comparisons with other algorithms like SVM and KNN. The ROC value was 0.967, and the Kappa statistic value was 0.942. the limitation of this study was the model's performance is dependent on the quality of the initial PCA-based transformation.

Ade Jamal, et al. [2018][28], The proposed methodology involves applying PCA and K-Means for dimensionality reduction and then using SVM and XGBoost for classification. WBCD Dataset from UCI was used. PCA combined with SVM yielded up to 97.07% accuracy, and K-Means with three clusters combined with SVM achieved 97.8% accuracy, 98.7% specificity, and 96.1% sensitivity. One of the limitations of this method is that Sensitivity was low with fewer clusters.

Neha Sharma, and Deeksha Kumari [2022][29], The authors propose a hybrid approach combining Principal Component Analysis (PCA) for feature extraction, K-Means clustering for data grouping, and a voting classifier (integrating GNB, BNB, RF, and Support Vector Machine algorithms) for final prediction. The dataset used in this study was COVID-19 data from Mexico. The voting classifier outperformed other models, achieving 94% accuracy, 94% precision, and 94% recall. Among the limits of this method is that the model's performance could be further enhanced by incorporating deep learning techniques and comparing it with other existing prediction models.

Bichen Zheng, et al. [2014][30], proposed an approach combining Support Vector Machines (K-SVM) and K-means clustering. K-means is used to extract and cluster abstract features of tumors, which are then input into an SVM for classification. The dataset used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from UCI. The K-SVM

achieved 97.38% accuracy in diagnosing breast cancer. While the approach reduces the feature space, it does not decrease the size of the training sample set.

NirmalaDevi, et al. [2013] [31], proposed A combined model is developed that integrates k-means clustering with KNN, along with multiple pre-processing steps. These steps include removing noisy data and replacing missing values with means or medians. data is then classified using KNN to achieve better results. The performance was tested using (PIDD) dataset from UCI. The combined model achieved a classification accuracy of 97.4% with K=5. The Limitation of this method is the computational complexity increases significantly with higher K values.

Nitin Arora, et al. [2022][32], the proposed method used a hybrid approach that combines K-Means clustering for feature extraction and Support Vector Machine (SVM) for classification. The approach clusters data and uses those clusters as features to train the SVM classifier. The Pima Indians Diabetes Database was used. The proposed architecture outperformed other machine learning models, achieving 98.7% accuracy, 98.6% precision, 96.8% recall, and 97.5% F1 score. The study is limited to the Pima Indians Diabetes Database, which may not be representative of all populations.

Srinivasa K. G et al. [2006][33], proposed a generic feature extraction method based on Fuzzy C-Means (FCM) clustering. The system preprocesses data (including normalization and transformation), performs clustering using FCM, and then extracts relevant features from the clusters, which are used for classification with Artificial Neural Networks (ANN) and Support Vector Machines (SVM). the following dataset (Physics, sonar, Dermatology, Waveform Generator) was used. The results showed that classification accuracy was close to problem-specific feature extraction methods (physics dataset: 70% (generic) vs 73% (problem-specific with SVM), Sonar dataset: 88% (generic) vs 90.5% (problem-specific with ANN), Dermatology dataset: 83.3% (generic) vs 85% (problem-specific with ANN). Waveform dataset: 85.6% (generic) vs 86% (problem-specific with Bayesian Classification)). However, the limitation of this study was While the system performs well across multiple datasets, the performance might slightly lag behind dataset-specific feature extraction methods.

Chih-Fong Tsai, et al. [2011][34], proposed method a distance-based feature extraction method. Two types of distances are considered (the distance between a data point and its intra-cluster center, and the sum of distances between the data point and extra-cluster centers). It used ten datasets from UCI, including datasets like Balance Scale, Abalone, Tic-Tac-Toe Endgame, and Iris(xx). The proposed method improved classification accuracy when combining (the original with distance-based) features, particularly for datasets with lower dimensionality and fewer samples. The proposed method struggled with high-dimensional datasets.

Maciej Piernik, Tadeusz Morzy [2021] [35], The proposed framework incorporates clustering as a feature extraction step for classification. It clusters the dataset, encodes clusters as new features, and evaluates various classifiers on these enhanced datasets. The framework is agnostic to the type of clustering algorithm, similarity measures, classifiers, and datasets. The paper investigates 10 critical questions regarding this approach. Datasets Used: 16 publicly available datasets from the UCI. including Breast-cancer-Wisconsin, Ecoli, Iris, Glass, opt digits, Pen digits, and Pima-Indians-diabetes. The results were that Clustering-generated features can significantly increase the effectiveness of linear classifiers like (SVM), penalized multinomial regression, and discriminant analysis. However, for nonlinear classifiers like K-nearest neighbors (KNN) and random forests, clustering-generated features may degrade performance. One of the limitations of this framework Clustering-generated features may not always improve classification and can negatively impact nonlinear models like KNN and random forests. The method's success is highly contingent on the dataset characteristics and the number of clusters.

Chulhee Lee and David A. Landgrebe [1993][48], proposed a method that depends on decision boundaries, this approach identifies discriminately informative and discriminately redundant features by evaluating their relation to the decision boundary, The method predicts the minimum number of features needed to achieve the same classification accuracy. Synthetic data (Gaussian distributed datasets with known statistics) and real data (Multispectral remote sensing data) were used. The result was with synthetic data, one feature was sufficient to match the performance of the original 2D space. For the real-world remote sensing data, the method achieved a classification accuracy of 87.4% with just ten features, The method depends on the decision boundary matrix, which may become computationally intensive for very large datasets or high-dimensional data.

4. **Discussions**

Feature extraction is a fundamental technique in machine learning and a key factor for the success of many applications. Discussions focus on how to apply and compare different techniques to extract features effectively[49]such as descriptive statistics to understand the underlying distribution of data, Fisher discriminant analysis that can distinguish between different classes, and autoencoder that relies on neural networks to discover representations that preserve important information, as well as principal component analysis to reduce dimensions, and clustering to identify hidden patterns, which opens up new horizons for various applications such as classification and prediction[50].

Ref id	Author	Year	Dataset	Proposed method	Result	Limitation
[10]	Ruhul Amin et al.	2023	(ILPD) from UCI	A method combining (PCA), (FA), and (LDA)an integrated projection-based statistical feature extraction.	the Random Forest better result, 88.10% accuracy, 85.33% precision, 92.30% recall, 88.20% AUC, 88.68% F1 score.	model's runtime is longer, the study was also limited to statistical projection-based techniques.
[11]	Ashir Javeed, et al.	2024	(SNAC)	Built integrated systems combining (PCA, ICA, LLE, and TSNE) with (LR, KNN, and DT).	combination of PCA with (LR) achieves an accuracy of 89.04% and outperforms ICA, LLE, and TSNE. Also, TSNE showed good performance in combination with the naive Bayes model.	
[13]	Md Shamim Reza, Jinwen Ma	2016	simulated, breast cancer, wine, crab).	two feature extraction methods that integrate PCA and ICA. In the first method, PCA is applied to the data, followed by ICA. In the second method, both PCA and ICA feature.	The IC-PC method achieved the highest classification accuracy, outperforming PCA and ACI, particularly in the classification of breast cancer and wine.	the performance depends on the type of data used, ICA requires independent components, which can be a complex process.
[14]	Shifei Ding et al.	2012	artificial datasets (e.g. Four-gauss, Easy doughnut, Difficult doughnut), and real datasets (Wine, Iris, and (WDBC)	use (PCA) and (LDA), (ICA), and kernel-based methods that transform non-linear problems into linear ones	Non-linear methods such as (ICA) and (KPCA) demonstrated superior performance when handling non-linear data, preserving important features while reducing dimensionality	feature extraction techniques are based on assumptions of statistical independence, which may not hold for all datasets, and some methods are computationally expensive
[16]	Nojun Kwak, Chong-Ho Choi	2003	IBM and UCI datasets, the Wisconsin Breast Cancer and Sonar Target datasets	Proposed a modified ICA algorithm for feature extraction in binary classification problems	The ICA-based method outperformed PCA and traditional ICA in terms of classification accuracy	The method is currently limited to binary classification problems. The algorithm's reliance on ICA may be limited by the statistical properties of the data
[17]	Nojun Kwak et al.	2001	Chess End game, breast cancer	using independent component analysis (ICA) for supervised classification problems	using the C4.5 and (MLP). accuracy of 95.42% (breast cancer data) and accuracy of 98.80% (chess data)	the algorithm is designed for two-class classification problems, and the success of the algorithm depends on the choice of some parameters
[18]	Yong Xu et al.	2007	Splice, Diabetes, Banana, and Cancer.	an improved KPCA (IKPCA) method that speeds up feature	IKPCA outperformed KPCA in feature extraction efficiency, in the "Splice" dataset, IKPCA had a Significantly lower classification error rate (17.8%) compared to KPCA(%21.8)	The method relies on the empirical determination of the number of nodes, and the node selection in IKPCA is computationally expensive

[20]	Padathala Visweswara Rao et al.	2024	Cleveland Heart Dataset (CHDS), Statlog Heart Dataset (SHDS), and South African Heart Dataset (SAHDS).	uses Probabilistic Principal Component Analysis (PPCA) for feature extraction and selection	The PPCA-based approach outperformed PCA and ReliefF, and it had an accuracy of 98.6%, Precision: of 97.5%, and Specificity of 97.2%)	the method relies on the assumption that features contribute independently to theclassification
[21]	Xiaoming Wang, Hong Peng	2012	Wine dataset	using a technique based on the MCVSVM algorithm, which combines the strengths of SVM and FDA to provide better class separation	The proposed method outperformed MMDA in projecting the dataset into a lower-dimensional space while maintaining a smaller scatter for same-class samples	which may limit the generalizability of the results to larger and more complex datasets
[22]	Olayinka Ogundile, et al.	2024	credit card transaction dataset from Kaggle	use of an Ensemble Hidden Markov Model combined with two techniques (PCA) and MRE	The PCA-EHMM showed significant performance improvement, the MRE-EHMM achieved similar performance with much lower computational complexity	focuses on feature extraction techniques without exploring other machine learning models that might further optimize computational efficiency or prediction accuracy
[23]	Wei Li, et al.	2015	Data was collected from a machinery fault simulator (normal, rotor imbalance, and three bearing faults)	The extraction procedure is based on the central limit theorem, it evaluates the statistical features using a decoupling technique	accuracy of 100%, while ensuring a false classification rate (FCR) of less than 0.3%	computational load is higher due to the calculation of statistical features
[24]	Zahra Salek et al.	2023	The Credit Card Fraud Detection Dataset from Kaggle	use two feature extraction methods (PCA) and (CAE)) and three data sampling techniques RUS, SMOTE Tomek, SMOTE, and	the combination of (RUS) and (CAE), yielded the highest F1 score of 0.909 and an AUC of 0.988 when using LightGBM and XGBoost, respectively.	the increased computational cost associated with the use of CAE
[25]	Simon Cramer et al.	2022	samples from industrial production processes	comparison between (PCA) and autoencoders. PCA reduces linear correlations, while autoencoders handle both linear and non-linear correlations	The combination of PCA and SVR yielded the best performance, with the lowest MAPE(%2.91)	autoencoder's performance in combination with SVR was unstable, and the combination of autoencoder and FFNN, expected to give better results, was outperformed by PCA-SVR due
[26]	Jianxin Zhu, et al.	2024	the Credit Scoring Dataset from Kaggle, the China Union Pay (CUP)dataset	using an improved multilayer Restricted Boltzmann Machine (RBM) for feature extraction	The improved RBM model significantly reduced data dimensionality and achieved high classification performance	that over-pruning can degrade performance
[27]	Changsheng Zhu et al.	2019	PIDD from UCI	A method that combines PCA with K-means clustering and logistic regression.	achieved an accuracy of 97.40% using PCA and K-means	the model's performance is dependent on the quality of the initial PCA-based transformation
[28]	Ade Jamal, et al.	2018	The Wisconsin Breast Cancer Dataset from UCI.	applying PCA and K-Means for dimensionality reduction and then using SVM and XGBoost for classification	PCA with SVM achieved 97.07% accuracy and K-Means with SVM achieved 97.8% accuracy, 98.7% specificity, and 96.1% sensitivity	Sensitivity was low with fewer clusters
[29]	Neha Sharma, Deeksha Kumari	2022	COVID-19 data from Mexico	A method combining (PCA) and K-Means clustering, and a voting classifier (GNB, BNB, RF, and SVM)	94% accuracy, 94% precision, and 94% recall	that the model's performance could be further enhanced by incorporating deep learning techniques.

[30]	Bichen Zheng, et al.	2014	WDBC dataset from UCI	combining K-means clustering and(K-SVM)	Accuracy of 97.38%	the approach does not decrease the size of the training sample set
[31]	Nirmala Devi, et al.	2013	(PIDD) from UCI	A method combined k-means clustering with KNN	accuracy of 97.4% with K=5	the computational complexity increases significantly with higher K values
[32]	Nitin Arora, et al	2022	The Pima Indians Diabetes Database	Combines (SVM) and K-Means clustering	98.7% accuracy, 98.6% precision, 96.8% recall, and 97.5% F1 score	The study is limited to P IDD
[33]	Srinivasa K. G et al.	2006	dataset (Physics, sonar, Dermatology, Waveform Generator)	A method based on Fuzzy C-Means (FCM) clustering	accuracy (physics 70%, Sonar 88%, Dermatology 83.3%, Waveform 85.6%)	While the system performs well across multiple datasets, the performance might slightly lag behind dataset-specific feature extraction methods
[34]	Chih-Fong Tsai, et al.	2011	Balance Scale, Tic-Tac-Toe Endgame, Abalone, and Iris(xx) from UCI	a novel distance-based feature extraction method	improved classification accuracy when combined distance-based with the original features.	struggled with high-dimensional datasets
[35]	Maciej Piernik, Tadeusz Morzy	2021	16publicly available datasets from UCI	using clustering as a feature extraction step for classification	that Clustering-generated features can improve the performance of linear classifiers like (SVM) penalized multinomial regression, and discriminant analysis	Clustering-generated features may not always improve classification and The method's success is contingent on the dataset characteristics and the number of clusters
[48]	Chulhee Lee, David A. Landgrebe	1993	Synthetic data (Gaussian distributed datasets with known statistics) and real data (Multispectral remote sensing data	propose an extraction method based on feature decision boundaries	with synthetic data, one feature was sufficient to match the performance of the original 2D space. As for the real data, an accuracy of 87.4% with just ten features	The method relies on calculating the decision boundary feature matrix, which may become computationally intensive for very large datasets

5. Conclusion

Feature extraction is a key part of machine learning because it Contributes to reducing dimensionality, improving model performance, and understanding hidden patterns in data using techniques such as PCA, LDA, convolutional analysis, clustering methods, etc. In addition, the advancement in feature extraction techniques shows signs of new possibilities in data analysis and machine learning, and the choice of the most appropriate method depends on the nature of the data, and the main goal is to achieve a balance between the simplicity of the model and its accuracy, which contributes to obtaining reliable and fast results.

References

[1] R. Imamguluyev, "The Rise of GPT-3: Implications for Natural Language Processing and Beyond," *International Journal of Research Publication and Reviews*, vol. 4, pp. 4893-4903, 03/03 2023, doi: 10.55248/gengpi.2023.4.33987.

[2] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover," *PeerJ Computer Science*, vol. 7, p. e443, 2021.

[3] E. Sadikoğlu, M. Gök, M. Mijwil, and I. Kosesoy, "The Evolution and Impact of Large Language Model Chatbots in Social Media: A Comprehensive Review of Past, Present, and Future Applications," vol. 6, pp. 67-76, 12/21 2023.

[4] D. Valiaiev, "Detection of Machine-Generated Text: Literature Survey," *arXiv preprint arXiv:2402.01642*, 2024.

[5] L. Dugan, D. Ippolito, A. Kirubarajan, and C. Callison-Burch, "RoFT: A tool for evaluating human detection of machine-generated text," *arXiv preprint arXiv:2010.03070*, 2020.

[6] A. Das and R. M. Verma, "Can machines tell stories? a comparative study of deep neural language models and metrics," *IEEE Access*, vol. 8, pp. 181258-181292, 2020.

[7] I. Dergaa, K. Chamari, P. Zmijewski, and H. Ben Saad, "From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and potential threats of ChatGPT in Academic Writing," *Biology of Sport*, vol. 40, pp. 615-622, 03/07 2023, doi: 10.5114/biolsport.2023.125623.

[8] A. Rauchfleisch, M. Sele, and C. Caspar, "Digital astroturfing in politics: Definition, typology, and countermeasures," *Studies in Communication Sciences*, vol. 18, pp. 69-85, 11/14 2018, doi: 10.24434/j.scoms.2018.01.005.

[9] J. Peng, R. K. K. Choo, and H. Ashman, "Astroturfing Detection in Social Media: Using Binary n-Gram Analysis for Authorship Attribution," in *2016 IEEE Trustcom/BigDataSE/ISPA*, 23-26 Aug. 2016 2016, pp. 121-128, doi: 10.1109/TrustCom.2016.0054.

[10] S. S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, and A. Bedi, "A Survey on the Possibilities & Impossibilities of AI-generated Text Detection," *Transactions on Machine Learning Research*, 2023.

[11] S. Katzenbeisser and F. Petitcolas, *Information hiding*. Artech house, 2016.

[12] U. Topkara, M. Topkara, and M. J. Atallah, "The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions," in *Proceedings of the 8th workshop on Multimedia and security*, 2006, pp. 164-174.

[13] M. J. Atallah *et al.*, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation," in *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, 2001: Springer, pp. 185-200.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] H. Ueoka, Y. Murawaki, and S. Kurohashi, "Frustratingly easy edit-based linguistic steganography with a masked language model," *arXiv preprint arXiv:2104.09833*, 2021.

[16] Z. M. Ziegler, Y. Deng, and A. M. Rush, "Neural linguistic steganography," *arXiv preprint arXiv:1909.01496*, 2019.

[17] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang, "Provable robust watermarking for ai-generated text," *arXiv preprint arXiv:2306.17439*, 2023.

[18] S. S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, and A. S. Bedi, "Towards possibilities & impossibilities of ai-generated text detection: A survey," *arXiv preprint arXiv:2310.15264*, 2023.

[19] J. Qiang, S. Zhu, Y. Li, Y. Zhu, Y. Yuan, and X. Wu, "Natural language watermarking via paraphraser-based lexical substitution," *Artificial Intelligence*, vol. 317, p. 103859, 2023/04/01/ 2023, doi: https://doi.org/10.1016/j.artint.2023.103859.

[20] X. Zhao, Y.-X. Wang, and L. Li, "Protecting language generation models via invisible watermarking," *arXiv preprint arXiv:2302.03162*, 2023.

[21] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang, "Robust distortion-free watermarks for language models," *arXiv preprint arXiv:2307.15593*, 2023.

[22] M. Christ, S. Gunn, and O. Zamir, "Undetectable Watermarks for Language Models," *arXiv preprint arXiv:2306.09194*, 2023.

[23] B. Y. Idrissi, M. Millunzi, A. Sorrenti, L. Baraldi, and D. Dementieva, "Temperature Matters: Enhancing Watermark Robustness Against Paraphrasing Attacks," 2023.

[24] A. B. Hou *et al.*, "Semstamp: A semantic watermark with paraphrastic robustness for text generation," *arXiv preprint arXiv:2310.03991*, 2023.

[25] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," *arXiv preprint arXiv:2301.10226*, 2023.

[26] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," *arXiv preprint arXiv:2301.11305*, 2023.

[27] N. Nashid, M. Sintaha, and A. Mesbah, "Retrieval-based prompt selection for code-related few-shot learning," in *Proceedings of the 45th International Conference on Software Engineering (ICSE'23)*, 2023.

[28] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv," *arXiv preprint arXiv:2303.13408*, 2023.

[29] M. H. I. Abdalla, S. Malberg, D. Dementieva, E. Mosca, and G. Groh, "A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers," *Information*, vol. 14, no. 10, p. 522, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/10/522.

[30] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, *Automatic Detection of Generated Text is Easiest when Humans are Fooled*. 2020, pp. 1808-1822.

[31] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature," *arXiv preprint arXiv:2310.05130*, 2023.

[32] R. Zellers *et al.*, "Defending against neural fake news," *Advances in neural information processing systems*, vol. 32, 2019.

[33] K. Mohammadi, "Human vs machine generated text detection in Persian," ed, 2023.

[34] F. Xiong, T. Markchom, Z. Zheng, S. Jung, V. Ojha, and H. Liang, "Fine-tuning Large Language Models for Multigenerator, Multidomain, and Multilingual Machine-Generated Text Detection," *arXiv preprint arXiv:2401.12326*, 2024.

[35] F. Rangel and P. Rosso, "Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter," *Working notes papers of the CLEF 2019 evaluation labs*, vol. 2380, pp. 1-7, 2019.

[36] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing methods for twitter sentiment analysis," *arXiv preprint arXiv:1505.02973*, 2015.

[37] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, *A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis*. 2017.

[38] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Information Systems*, vol. 121, p. 102342, 2024/03/01/ 2024, doi: https://doi.org/10.1016/j.is.2023.102342.

[39] S. Vasudevan, "Enhancing the Sentiment Classification Accuracy of Twitter Data using Machine Learning Algorithms," 2021, pp. 189-199.

[40] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199-22213, 2022.

[41] H. Cheng, S. Liu, W. Sun, and Q. Sun, "A Neural Topic Modeling Study Integrating SBERT and Data Augmentation," *Applied Sciences*, vol. 13, no. 7, p. 4595, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4595.

[42] H. Face, "all-MiniLM ". [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[43] H. Face, "mpnet " 2022. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[44] N. a. G. Reimers, Iryna, "paraphrase-multilingual-mpnet-base-v2," 2019. [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

[45] "sentence-transformers/gtr-t5-large," 2021. [Online]. Available: https://huggingface.co/sentence-transformers/gtr-t5-large.

[46] L. Tunstall *et al.*, "Efficient few-shot learning without prompts," *arXiv preprint arXiv:2209.11055*, 2022.

[47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[48] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," *Applied Sciences*, vol. 12, no. 19, p. 9691, 2022.

[49] B. Jiang, L. Wang, J. Tang, and B. Luo, "Context-Aware Graph Attention Networks," *arXiv preprint arXiv:1910.01736*, 2019.

[50] J. Chen and H. Chen, "Edge-featured graph attention network," *arXiv preprint arXiv:2101.07671*, 2021.

[51] E. Alothali, M. Salih, K. Hayawi, and H. Alashwal, "Bot-mgat: A transfer learning model based on a multi-view graph attention network to detect social bots," *Applied Sciences*, vol. 12, no. 16, p. 8117, 2022.

[52] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, pp. 1-23, 2019.

[53] S. Zhang, C. Zhou, Y. Li, X. Zhang, L. Ye, and Y. Wei, "Irregular Scene Text Detection Based on a Graph Convolutional Network," *Sensors*, vol. 23, no. 3, p. 1070, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1070.

[54] N. R. Aljohani, A. Fayoumi, and S.-U. Hassan, "Bot prediction on social networks of Twitter in altmetrics using deep graph convolutional networks," *Soft Computing*, vol. 24, no. 15, pp. 11109-11120, 2020.

[55] L. Zhang, H. Song, N. Aletras, and H. Lu, "Node-Feature Convolution for Graph Convolutional Networks," *Pattern Recognition*, vol. 128, p. 108661, 2022/08/01/ 2022, doi: https://doi.org/10.1016/j.patcog.2022.108661.

[56] X. Zhu, L. Zhu, J. Guo, S. Liang, and S. Dietze, "GL-GCN: Global and local dependency guided graph convolutional networks for aspect-based sentiment classification," *Expert Systems with Applications*, vol. 186, p. 115712, 2021.

[57] L. Zhang and H. Lu, "A feature-importance-aware and robust aggregator for GCN," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1813-1822.

[58] H. B. Giglou, M. Rahgouy, A. Rahmati, T. Rahgooy, and C. D. Seals, "Profiling Irony and Stereotype Spreaders with Encoding Dependency Information using Graph Convolutional Network," in *CLEF*, 2022, pp. 1613-0073.

[59] V. Jimenez-Villar, J. Sánchez-Junquera, M. Montes-y-Gómez, L. Villaseñor-Pineda, and S. P. Ponzetto, "Bots and gender profiling using masking techniques: Notebook for PAN at CLEF 2019," in *CEUR Workshop Proceedings*, 2019, vol. 2380: RWTH Aachen, pp. 1-8.

[60] A. Mahmood and P. Srinivasan, "Twitter Bots and Gender Detection using Tf-idf," in *CLEF (Working Notes)*, 2019.

[61] I. Vogel and P. Jiang, "Bot and Gender Identification in Twitter using Word and Character N-Grams," in *CLEF (Working Notes)*, 2019.

[62] R. Goubin, D. Lefeuvre, A. Alhamzeh, J. Mitrovic, E. Egyed-Zsigmond, and L. G. Fossi, "Bots and Gender Profiling using a Multi-layer Architecture," in *CLEF (Working Notes)*, 2019.

[63] M. Polignano, M. G. de Pinto, P. Lops, and G. Semeraro, "Identification Of Bot Accounts In Twitter Using 2D CNNs On User-generated Contents," in *Clef (working notes)*, 2019.

[64] E. Puertas, L. G. Moreno-Sandoval, F. M. Plaza-del Arco, J. A. Alvarado-Valencia, A. Pomares-Quimbaya, and L. Alfonso, "Bots and gender profiling on twitter using sociolinguistic features," *CLEF (Working Notes)*, pp. 1-8, 2019.

[65] D. Kosmajac and V. Keselj, "Twitter User Profiling: Bot and Gender Identification: Notebook for PAN at CLEF 2019," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, 2020: Springer, pp. 141-153.

[66] T. Fagni and M. Tesconi, "Profiling Twitter Users Using Autogenerated Features Invariant to Data

[67] W. Qu *et al.*, "Provably Robust Multi-bit Watermarking for AI-generated Text via Error Correction Code," *arXiv preprint arXiv:2401.16820*, 2024.