The Scientific Temper (2025) Vol. 16 (2): 3736-3746

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.06

https://scientifictemper.com/

RESEARCH ARTICLE

A robust feature selection approach for high-dimensional medical data classification using enhanced correlation attribute evaluation

G. Vanitha[1*], M. Kasthuri[2]

Abstract

The challenge of high-dimensional feature spaces and redundant attributes significantly impacts classification performance in medical datasets. Addressing this, the proposed Enhanced Correlation Attribute Evaluation (E-CAE) method effectively integrates multiple correlation measures such as Pearson, Spearman, Kendall, Biweight Midcorrelation, and Distance Correlation to rank and select the most relevant features. This hybrid feature selection technique was rigorously tested on three datasets: the Darwin dataset, Parkinson's speech dataset, and the Dyslexia dataset. The E-CAE method demonstrated superior classification performance across various models, achieving a remarkable 95.64% accuracy on the Darwin dataset, 93.42% accuracy on the Parkinson's dataset, and 90.86% accuracy on the Dyslexia dataset. These results notably outperformed traditional feature selection techniques. The novelty of this approach lies in its composite scoring mechanism, which ensures robust feature evaluation and significantly enhances classification accuracy across diverse biomedical datasets.

Keywords: Attribute evaluation, Disease classification, Feature selection, High-dimensional data, Medical diagnosis

Introduction

The fields of healthcare along with other domains benefit significantly from machine learning (ML) and artificial intelligence (AI) advancements because they produce efficient data-driven methods to identify diseases (Kasthuri and Jency 2020; Reddy *et al*., 2023; Khalifa *et al*., 2024; Faiyazuddin *et al*., 2025). Early diagnosis of neurological disorders, including dyslexia, Parkinson's disease, and Alzheimer's disease, is essential for prompt medical intervention and improved treatment outcomes (Usman *et al*., 2021; Jha & Kumar, 2024). The standard diagnostic approach employs clinical examinations alongside expert opinions for testing which requires considerable time and demonstrates subjectivity as well as human operational mistakes. The increasing availability of high-dimensional medical datasets offers opportunities to apply machine learning techniques to automate and enhance disease detection processes (Hider *et al*., 2024). However, leveraging these datasets effectively poses significant challenges, including high dimensionality, class imbalance, and the need for interpretability in decision-making (Gholampour 2024; Wilson & Anwar 2024).

High dimensionality is a common issue in medical datasets, where many features or variables are collected for analysis (Zebari *et al*., 2020). While high-dimensional data can capture complex patterns, it often contains redundant or irrelevant features that may negatively impact model performance (Chin *et al*., 2024). The label "curse of dimensionality" describes a problem that causes models to experience higher complexity in computation alongside increased potential to incorrectly learn noise patterns instead of true relationships. Feature selection

[1]Research Scholar/Assistant Professor, Department of Information Technology, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620 024, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli- 620 024, Tamil Nadu, India.

***Corresponding Author:** G. Vanitha, Department of Information Technology, Bishop Heber College (Autonomous), Affiliated Bharathidasan University, Tiruchirappalli-620 024, Tamil Nadu, India., E-Mail: vanithaguna11@gmail.com

How to cite this article: Vanitha, G., Kasthuri, M. (2025). A robust feature selection approach for high-dimensional medical data classification using enhanced correlation attribute evaluation. The Scientific Temper, **16**(2):3736-3746.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.06

Source of support: Nil

Conflict of interest: None.

Received: 12/01/2025 **Accepted:** 24/02/2025 **Published:** 20/03/2025

techniques help solve the issue by pinpointing essential features so models become more dimensionally reduced while achieving better generalizability (Ali *et al*., 2024). The process of choosing superior features presents an intricate challenge due to the need to find the right balance between model ease and predictive performance achievement.

Class imbalance is another critical challenge in medical data analysis. In many healthcare applications, datasets contain disproportionately fewer samples of diseased individuals compared to healthy controls (Kitova *et al*., 2024). Machine learning models experience an unbalanced distribution that promotes the majority class and adversely affects their ability to identify significant but uncommon healthcare circumstances. Traditional classifiers show substandard performance in detecting minority class instances while achieving high overall accuracy because they ignore the class imbalance problem. Techniques such as oversampling, undersampling, and cost-sensitive learning have been explored to mitigate class imbalance, yet these approaches have limitations, including increased computational costs and the risk of overfitting (Kavitha & Kasthuri 2024).

Several studies have attempted to address these challenges through innovative machine-learning techniques. An online gamified test was developed for predicting dyslexia risk using random forest classifiers, achieving 78% accuracy (Luz Rello *et al*. 2020). The proposal offered an easy-to-use dyslexia screening solution yet faced limitations due to the sophisticated machine learning system complexity and requirement of substantial datasets. A bio-inspired method that uses a genetic algorithm (GA) and binary particle swarm optimization (BPSO) was combined with 11 ML classifiers for Parkinson's disease classification (Akram and Latha 2020). They achieved 89% accuracy but faced challenges related to high computational time due to iterative convergence processes.

Feature selection-based machine learning models were introduced for Parkinson's disease prediction using Boruta, RFE, and Random Forest algorithms, achieving 82.35% accuracy (Nazmun Nahar *et al*. 2021). While effective, these methods risk overfitting and can be computationally expensive. Faisal *et al*. (2022) enhanced Parkinson's disease prediction by integrating principal component analysis (PCA) achieving an accuracy of 88.33%. However, the wrapper-based feature selection methods used in their study resulted in high computational overhead.

For dyslexia detection, an ensemble learning technique was proposed by combining various ML models with feature selection methods like select k best and mutual information gain (Tabassum Jan *et al*. 2022). Their approach achieved 90% accuracy but lacked evaluation using F1-score metrics. Karim Gasmi *et al*. (2024) developed an adaptive genetic algorithm-based ensemble learning model for dyslexia prediction, attaining 90% accuracy. Despite its effectiveness, this method was computationally intensive. Shahriar Kaisar and Abdullahi Chowdhury (2022) explored the integration of oversampling and ensemble learning for imbalanced dyslexia datasets, achieving notable performance improvements. However, the full dataset training led to high computational costs (Vanitha and Kasthuri 2021).

Vectorial genetic programming (VEGP) is utilized for Alzheimer's disease prediction through handwriting analysis, achieving 71% accuracy (Irene *et al*. 2024). VEGP demonstrated robustness by avoiding genetic drift but required fine-tuning, limiting scalability.

Despite significant advancements in ML and feature selection techniques, challenges persist in handling high-dimensional medical datasets, especially for disease detection tasks. Existing feature selection techniques, such as recursive feature elimination (RFE), Boruta, principal component analysis (PCA), and genetic algorithms (GA), have demonstrated varying degrees of success. However, these methods often struggle to balance the trade-off between dimensionality reduction and classification accuracy. Many traditional approaches primarily focus on linear correlations and fail to capture complex, non-linear relationships between features and target variables. Additionally, techniques like wrapper-based methods, while effective in improving accuracy, are computationally expensive and unsuitable for large datasets. Class imbalance remains a critical issue, leading to biased models that favor majority classes. Furthermore, most advanced models lack transparency and interpretability, making it difficult for healthcare professionals to trust and adopt these solutions in clinical settings. There is a clear need for a feature selection method that can efficiently handle high-dimensional data, reduce overfitting, and improve interpretability while maintaining computational efficiency.

This research derives its motivation from healthcare's increasing need for machine learning models that achieve both accuracy and interpretability. Dyslexia and Parkinson's disease need early precise diagnoses to enable immediate treatment and appropriate interventions. However, existing machine learning models face limitations due to high-dimensional data, class imbalance, and lack of interpretability. These challenges hinder the practical implementation of automated diagnostic systems in clinical environments. The success of previous studies in applying machine learning techniques highlights the potential for data-driven solutions. Still, the consistent struggle with computational inefficiencies and the inability to fully exploit feature relationships necessitates a more robust approach. This research is driven by the need to develop a method that can overcome these limitations by effectively selecting relevant features, improving model performance,

and ensuring interpretability, thereby contributing to advancements in medical diagnostics.

The primary objective of this research is to develop and evaluate an enhanced correlation attribute evaluation (E-CAE) method for effective feature selection and classification in high-dimensional medical datasets.

Proposed Work

The proposed research introduces the E-CAE method to address the limitations of traditional feature selection techniques in handling high-dimensional medical datasets. Figure 1 represents the overall workflow of the proposed E-CAE method. This method is designed to improve feature relevance assessment, reduce dimensionality, and enhance classification performance.

2.1 Multi-Correlation Metric Integration

To assess feature relevance comprehensively, E-CAE employs multiple correlation measures. These metrics are designed to capture various relationships between feature variables and the target class.

- *2.1.1 Pearson Correlation Coefficient (PCC)*

PCC functions as a common statistical tool that evaluates both the strength and direction of a linear connection between two variables that exist on continuous value scales. Within feature selection applications the PCC establishes the linear relationship strength between feature

X_i and target variable Y. Linear correlation strength between feature and target information rises when the absolute Pearson coefficient value increases. To determine essential predictors for supervised learning tasks this evaluation method delivers valuable insights about feature contributions. The Pearson correlation coefficient connecting feature X_i to the target variable, Y exists in the form of equation 1

$$\rho_{X_i,Y} = \frac{Cov(X_i,Y)}{\sigma_{X_i}\sigma_Y} = \frac{\sum_{j=1}^{n}\left(X_{ij}-\overline{X_i}\right)\left(Y_j-\overline{Y}\right)}{\sqrt{\sum_{j=1}^{n}\left(X_{ij}-\overline{X_i}\right)^2}\sqrt{\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2}} \quad (1)$$

In this equation,

$\mathrm{Cov}\left(X_i,Y\right)$ represents the covariance between feature X_i and the target variable Y, while

σ_{X_i} and σ_Y denote the standard deviations of X_i and Y, respectively.

- *2.1.2 Spearman Rank Correlation (SRC)*

Non-parametric statistical analysis using SRC enables researchers to determine both the strength of monotonic associations and the direction of their relationship between two variables. The Spearman correlation evaluates rank

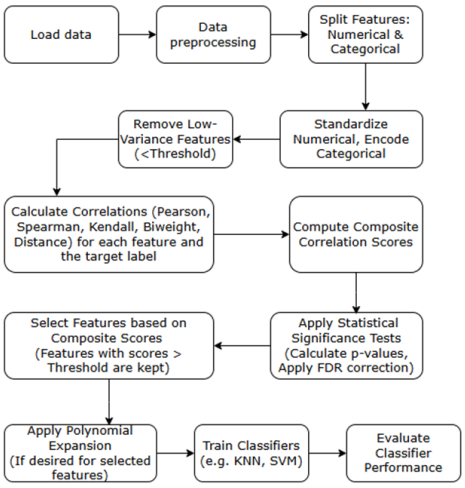

Figure 1: E-CAE flow diagram

order patterns while disregarding the evaluation of linear relationships that the Pearson correlation would perform. This monotonality capability of SRC makes it an ideal option for tracking linear and non-linear trends between target variables and features while evaluating high-dimensional medical datasets. The Spearman correlation coefficient between

X_i metrically measured feature and target variable Y is estimated through the formula above.,

$$\rho_s = 1 - \frac{6\sum_{j=1}^{n} d_j^2}{n\left(n^2-1\right)} \quad (2)$$

Here, d_j represents the difference between the ranks of the j-th observation in feature

X_i and the corresponding rank in the target variable Y, and n is the total number of observations.

- *2.1.3 Kendall Rank Correlation (KRC)*

The value τ or KRC indicates a non-parametric statistic that measures both the strength and direction of two variable relationships. The Kendall correlation method focuses on ordinal variable associations while it differs from both Pearson's linear association analysis and Spearman's ranked-based monotonic correlation tests. This makes it highly suitable for datasets where relationships between features and target variables may not be linear or even strictly monotonic, such as in complex medical datasets.

The Kendall correlation coefficient between a feature X_i and the target variable Y is defined as:

$$\tau = \frac{C-D}{\frac{1}{2}n\left(n-1\right)} \quad (3)$$

In this equation:

- C is the number of concordant pairs,
- D is the number of discordant pairs,
- n is the total number of data points.

- *2.1.4 Biweight Midcorrelation (Bicor)*

Bicor is a robust correlation measure that effectively reduces the influence of outliers and extreme values in the data. Unlike traditional correlation measures such as Pearson or Spearman, which can be sensitive to anomalies, Biweight Midcorrelation down-weights the impact of extreme observations, making it highly suitable for analyzing complex and noisy datasets. This property is particularly advantageous in high-dimensional medical datasets, where noisy or erroneous data can significantly distort traditional correlation metrics. The Biweight Midcorrelation between a feature X_i and the target variable Y is mathematically defined as:

$$\mathrm{Bicor}(X_i,Y)=\frac{\sum_{j=1}^{n}\left(X_{ij}-\widetilde{X_i}\right)\cdot\left(Y_j-\tilde{Y}\right)\cdot w_{Xij}\cdot w_{Yj}}{\sqrt{\sum_{j=1}^{n}\left(\left(X_{ij}-\widetilde{X_i}\right)^2\cdot w_{Xij}^2\right)}\cdot\sqrt{\sum_{j=1}^{n}\left(\left(Y_j-\tilde{Y}\right)^2\cdot w_{Yj}^2\right)}} \quad (3)$$

In this equation:

- $\widetilde{X_i}$ and $\tilde{Y}$ represent the medians of the feature X_i and the target variable Y, respectively.
- w_{Xij} and w_{Yj} are weight functions that down-weight the influence of data points that are far from the median.
- The weight function is defined as:

$$w_{Xij}=\left(1-\left(\frac{X_{ij}-\widetilde{X_i}}{9\cdot\mathrm{MAD}(X_i)}\right)^2\right)^2 \quad \text{for} \quad \left|\frac{X_{ij}-\widetilde{X_i}}{9\cdot\mathrm{MAD}(X_i)}\right|<1 \quad (4)$$

and similarly for w_{Yj}, where

$\mathrm{MAD}(X_i)$ is the Median Absolute Deviation (MAD) of feature X_i, a robust measure of statistical dispersion.

- *2.1.5 Distance Correlation (dCor)*

The statistical measure dCor expresses associations between datasets or random variables through linear relationships as well as non-linear associations. Distance correlation surpasses linear-only Pearson metrics by detecting wide-ranging dependent relationships so it functions well in high-dimensional heterogeneous datasets for feature selection. The distance correlation between a feature X_i and the target variable Y is defined as:

$$\mathrm{dCor}(X_i,Y)=\frac{\mathrm{dCov}(X_i,Y)}{\sqrt{\mathrm{dCov}(X_i,X_i)\cdot\mathrm{dCov}(Y,Y)}} \quad (5)$$

Here:

- $\mathrm{dCov}(X_i,Y)$ is the distance covariance between
- X_i and Y.
- $\mathrm{dCov}(X_i,X_i)$ and
- $\mathrm{dCov}(Y,Y)$ represent the self-distance covariances
- of X_i and Y, respectively.

The distance covariance is calculated as:

$$\mathrm{dCov}^2(X_i,Y)=\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}A_{jk}B_{jk} \quad (6)$$

Where:

- A_{jk} and
- B_{jk} are the centered distance matrices for
- X_i and Y, respectively.
- The distance matrices are computed by taking pairwise Euclidean distances between data points and then centering them using the following transformation:

$$A_{jk}=d\left(X_{ij},X_{ik}\right)-\overline{d_{j\cdot}}-\overline{d_{\cdot k}}+\overline{d_{\cdot\cdot}} \quad (7)$$

Similarly, for B_{jk}:

$$B_{jk}=d\left(Y_j,Y_k\right)-\overline{d_{j\cdot}}-\overline{d_{\cdot k}}+\overline{d_{\cdot\cdot}} \quad (8)$$

Where:

- $d\left(X_{ij},X_{ik}\right)$ is the Euclidean distance between observations j and k for feature X_i.
- $\overline{d_{j\cdot}}$ and $\overline{d_{\cdot k}}$ are the row and column means of the distance matrix, respectively.
- $\overline{d_{\cdot\cdot}}$ is the grand mean of all distances.

The Distance Correlation coefficient ranges from 0 to 1:

- A value of 0 indicates complete independence between the feature
- X_i and the target variable Y.
- A value of 1 suggests a perfect dependency, which could be linear or non-linear.

Composite Scoring and Feature Ranking

In the E-CAE method, an essential step after computing various correlation coefficients is the integration of these metrics into a unified score. This unified score, known as the composite correlation score, serves as a comprehensive measure to rank features based on their overall relevance to the target variable. The process of composite scoring and feature ranking ensures that both linear and non-linear relationships are considered, thereby enhancing the robustness of the feature selection process.

Composite Scoring

The composite correlation score is calculated by aggregating the values from multiple correlation metrics: PCC, SRC, KRC, Bicor, and dCor. Each of these measures captures different aspects of the relationship between a feature X_i and the target variable Y.

The composite score for each feature X_i is computed as a weighted sum of these correlation values:

$$\text{composite score}\left(X_i\right) = w_1 \cdot \left|\rho_{X_i,Y}\right| + w_2 \cdot \left|\rho_{s_{X_i,Y}}\right| + w_3 \cdot \left|\tau_{X_i,Y}\right| + w_4 \cdot \left|\rho_{bicor,X_i,Y}\right| + w_5 \cdot \left|dCor\left(X_i,Y\right)\right| \quad (9)$$

Where:

- $\rho_{X_i,Y}$ is the PCC.
- $\rho_{s_{X_i,Y}}$ is the SRC.
- $\tau_{X_i,Y}$ is the KRC.
- $\rho_{bicor,X_i,Y}$ is the Bicor.
- $\mathrm{dCor}\left(X_i,Y\right)$ is the dCor.
- w_1, w_2, w_3, w_4, w_5 are the weights assigned to each correlation metric.

In most cases, these weights are set equally to ensure that each metric contributes uniformly to the composite score:

$$w_1 = w_2 = w_3 = w_4 = w_5 = \frac{1}{5}$$

However, the weights can be adjusted to prioritize certain types of relationships depending on the dataset characteristics. For instance, in datasets with expected non-linear patterns, higher weights can be assigned to distance correlation and Biweight Midcorrelation.

Feature Ranking

Once the composite scores are computed, the features are ranked in descending order based on these scores. This ranking directly reflects the relevance of each feature to the target variable. Higher composite scores indicate stronger associations, making those features more significant for predictive modeling.

Let the set of composite scores for all features be:

$$\text{Composite Scores} = \left\{\mathrm{CS}\left(X_1\right), \mathrm{CS}\left(X_2\right), \ldots, \mathrm{CS}\left(X_p\right)\right\}$$

Where p is the total number of features. The features are sorted according to their composite scores:

$$\mathrm{Rank}\left(X_i\right) = \mathrm{argsort}\left(-\text{Composite Scores}\right)$$

Statistical Significance Testing

The statistical significance test allows researchers to confirm the importance of chosen features among variables. The analysis calculates *p-values* regarding feature correlation. Since multiple tests are conducted the false discovery rate (FDR) must be used to control Type I errors. The Benjamini-Hochberg (BH) procedure is applied for this purpose:

$$\text{Adjusted p-values} = \text{FDR Correction}\left(\text{p-values}\right)$$

Features with adjusted *p-value*s below a significance threshold (e.g., 0.05) are considered statistically significant and are retained for further analysis.

Feature Selection

Based on the ranking and statistical significance, the top-ranked features are selected. The selection criteria can be:

- A fixed number of top features (e.g., top 20 features).
- A threshold-based selection, where features with a composite score above a certain value are chosen.
- Significance-based selection, where features with adjusted *p-value*s below the threshold are selected.

Let T be the set of selected features:

$$T = \left\{X_i \middle| CS\left(X_i\right) \geq \theta \text{ and } P\text{-}adjusted\left(X_i\right) < 0.05\right\}$$

Where:

- θ is a user-defined threshold for the composite score.

This selected set T of features is then used for model training, ensuring that only the most relevant and statistically significant features contribute to the predictive model.

Adaptive Thresholding with FDR Correction

In high-dimensional datasets, where numerous features are evaluated for relevance to the target variable, the likelihood of selecting irrelevant or spurious features increases. This problem, known as the multiple comparisons problem, can lead to misleading conclusions due to the accumulation of Type I errors (false positives). To address this issue, the enhanced correlation attribute evaluation (E-CAE) framework incorporates adaptive thresholding combined with false discovery rate (FDR) Correction to ensure that feature selection is both statistically rigorous and robust.

Adaptive Thresholding

Adaptive thresholding dynamically adjusts the feature selection criteria based on the statistical significance of the computed correlation metrics. Unlike fixed thresholding, where a pre-defined cutoff is applied to correlation scores, adaptive thresholding evaluates the statistical reliability of each feature's association with the target variable.

After computing the composite correlation scores for each feature, a corresponding *p-value* is calculated to assess the likelihood that the observed correlation occurred by chance. Features with lower *p-value*s are more likely to have a genuine association with the target variable. The adaptive threshold for selecting features is defined as:

$$T = \{X_i \mid CS(X_i) \geq \theta \ and \ (X_i) < \alpha\}$$

were:

- T is the set of selected features.
- $\mathrm{CS}(X_i)$ is the composite score of the feature X_i.
- θ is the adaptive threshold for the composite score.
- $p(X_i)$ is the *p-value* associated with the feature X_i.
- α is the significance level, typically set to 0.05.

However, evaluating multiple features increases the risk of false discoveries. To mitigate this risk, FDR Correction is applied to control for multiple comparisons.

False Discovery Rate (FDR) Correction

The FDR represents the expected proportion of incorrectly rejected null hypotheses (false positives) among all rejected hypotheses. Controlling the FDR is crucial in feature selection because it balances the trade-off between discovering meaningful features and limiting false discoveries.

The BH procedure adjusts *p-value*s to account for the number of hypothesis tests, thereby reducing the likelihood of selecting features due to random chance. The Benjamini-Hochberg procedure operates as follows:

1. Compute p-values:

For each feature X_i, calculate the p-value p_i from its correlation with the target variable.

2. Rank the p-values:

Sort the p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

where m is the total number of features.

3. Calculate the BH critical value

For each sorted p-value $p_{(i)}$, compute the BH critical value:

$$\mathrm{BH}_i = \frac{i}{m} \cdot$$

where i is the rank of the p-value and á is the desired FDR level (commonly 0.05).

4. Identify significant features

Find the largest i such that: $p_{(i)} \leq \mathrm{BH}_i$

All features with p-values less than or equal to $p_{(i)}$ are considered statistically significant.

5. Adjust p-values

The adjusted *p-value*s control the FDR, ensuring that only features with strong evidence of relevance are selected. The adjusted *p-value*s are given by:

$$p_{\mathrm{adj}}(X_i) = \min\left(\frac{p_i \cdot m}{i}, 1\right) \cdot \qquad (10)$$

This correction reduces the chance of false discoveries while retaining meaningful features.

Integration of Adaptive Thresholding and FDR Correction

In the E-CAE method, adaptive thresholding and FDR correction are integrated to form a two-step filtering process:

1. Initial feature filtering

Features are first filtered based on their Composite Correlation Scores. Only features with scores above the adaptive threshold θ proceed to the next step.

2. Statistical Validation

The FDR-corrected *p-value*s are then used to validate the statistical significance of the remaining features. Features with $p_{\mathrm{adj}}(X_i) <$ are retained.

The final feature selection set S is defined as:

$$S = \{X_i \mid CS(X_i) \geq \theta \, and \, p_{adj}(X_i) < \alpha\}$$

This dual filtering process ensures that only features with strong and statistically significant relationships to the target variable are selected, improving both the reliability and interpretability of the model.

Results and Discussion

Dataset Description

The proposed E-CAE method was rigorously evaluated using three distinct medical datasets: the Darwin dataset (Alzheimer) (Cilia *et al*., 2018), the Parkinson's disease speech dataset (Dipayan 2019), and the dyslexia dataset (Rello 2020). These datasets were selected for their high dimensionality and varied complexity, which effectively test the scalability and performance of the proposed feature selection method. Each dataset presents unique challenges related to feature relevance, dimensionality reduction, and classification accuracy. Table 1 summarizes the datasets used in this study, providing detailed information on the number of features, total records, and data sources.

The Darwin dataset comprises 174 records with 451 features and focuses on Alzheimer's disease diagnosis. The Parkinson's disease speech dataset includes 756 records with 754 features, capturing diverse speech signal features essential for diagnosing Parkinson's disease. The dyslexia

Table 1: Datasets summary

S. No	*Dataset*	*# Features*	*# Records*
1	Alzheimer/Darwin dataset	451	174
2	Parkinson's disease speech dataset	754	756
3	Dyslexia dataset	197	3,644

dataset consists of 3,644 records and 197 features, designed for identifying individuals with dyslexia.

Analysis

Results for Alzheimer/Darwin Dataset

The performance evaluation of the proposed E-CAE framework on the Alzheimer/Darwin dataset demonstrates significant improvements across multiple classification models. The comparative analysis of classification results is presented in Table 2, highlighting the superiority of the E-CAE method over existing approaches (Figure 1).

The baseline model developed by Azzali *et al*. (2024) achieved an accuracy of 71% and a recall of 82%. However, it lacked comprehensive reporting on precision, F1-score, and ROC AUC, limiting a complete performance comparison. In contrast, the proposed E-CAE method consistently outperformed the baseline model across all evaluated classifiers.

The K-nearest neighbors (KNN) classifier, when integrated with the E-CAE method, achieved an accuracy of 77.36%. Notably, it attained a perfect precision score of 100%, indicating that all predicted positive cases were indeed correct. However, the recall was relatively lower at 52%, reflecting challenges in identifying all true positive cases. Despite this, the ROC AUC score reached 90.71%, showcasing the model's strong discriminative power.

The Naive Bayes classifier exhibited a more balanced performance, achieving the highest accuracy of 86.79% among all classifiers. It reported a precision of 87.5%, a recall of 84%, and an F1-score of 85.71%. The ROC AUC score was also robust at 86.64%, demonstrating the classifier's effectiveness in distinguishing between positive and negative cases. This balanced performance across evaluation metrics emphasizes Naive Bayes as a highly effective model for the Alzheimer/Darwin dataset when combined with E-CAE.

Logistic regression (LR) achieved an accuracy of 83.02%. It maintained a precision of 80.77%, a recall of 84%, and an F1 score of 82.35%. The ROC AUC was notably high at 94.43%, indicating excellent model calibration and predictive capability. The marginally lower precision compared to Naive Bayes suggests a slightly higher rate of false positives, but overall, logistic regression exhibited strong classification performance.

The random forest (RF) and support vector machine (SVM) classifiers both achieved an accuracy of 84.91%, reflecting consistent performance. Both classifiers recorded identical precision, recall, and F1 scores of 84%, confirming their balanced classification strength. However, the ROC AUC for random forest was slightly higher at 95.64% compared to 95.29% for SVM. This indicates that random forest had a marginal advantage in distinguishing between the classes.

XGBoost produced an accuracy of 83.02%, with a precision of 83.33%, a recall of 80%, and an F1-score of 81.63%. The ROC AUC of 92% reflects strong overall model performance, although slightly lower than random forest and SVM. XGBoost demonstrated effective classification but showed a slight trade-off between precision and recall.

The decision tree (DT) classifier performed comparatively lower, with an accuracy of 67.92%. It recorded a precision of 66.67%, a recall of 64%, and an F1 score of 65.31%. The ROC AUC stood at 67.71%, indicating weaker discriminative ability. This result highlights the model's limitations in handling high-dimensional data without robust feature selection, further emphasizing the necessity of more sophisticated classifiers in conjunction with E-CAE.

Results for Parkinson's Disease Dataset

The classification performance of the proposed E-CAE method on the Parkinson's disease dataset is comprehensively analyzed and compared with existing methodologies in Table 3. This comparative analysis highlights the effectiveness of the E-CAE method in improving the classification outcomes across multiple classifiers (Figure 2).

Table 2: Comparative analysis of Alzheimer/Darwin classification results (%)

Classification techniques	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*ROC AUC (%)*
AZZALI *ET AL.*, (2024)	71	-	82	-	-
ECAE–KNN	77.36	100	52	68.42	90.71
ECAE–Naive Bayes	**86.79**	87.5	84	**85.71**	86.64
ECAE–LR	83.02	80.77	84	82.35	94.43
ECAE–RF	84.91	84	84	84	**95.64**
ECAE–SVM	84.91	84	84	84	95.29
ECAE – XGBoost	83.02	83.33	80	81.63	92
ECAE–DT	67.92	66.67	64	65.31	67.71

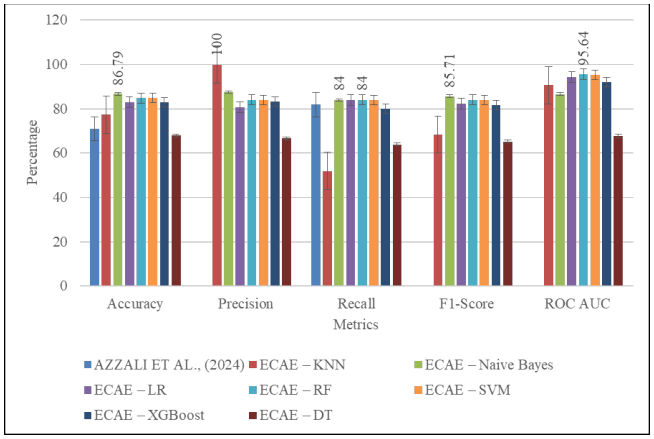

Figure 2: Comparative analysis of Alzheimer/Darwin classification results

The benchmark studies conducted by Nazmun *et al*. (2021) and Faisal *et al*. (2022) reported accuracies of 82.35 and 88.3%, respectively. Both studies demonstrated moderate performance but did not provide a comprehensive evaluation across all critical metrics, such as ROC AUC. In contrast, the E-CAE method, when paired with various classifiers, consistently yielded superior classification performance, showcasing its robustness in handling complex biomedical data.

Among the classifiers evaluated, the random forest (RF) classifier achieved the highest accuracy of 88.66%. This model also demonstrated a strong balance between sensitivity and specificity, as reflected in its recall of 96.39% and precision of 88.40%. The F1-score of 92.22% and a high ROC AUC of 92.62% further confirm the model's capacity to accurately distinguish between Parkinson's and non-Parkinson's cases. This result indicates the model's efficiency in leveraging the feature space optimized by the E-CAE method.

The XGBoost classifier closely followed, with an accuracy of 87.67%, precision of 87.91%, and recall of 96.39%. Its F1-score of 91.95% and ROC AUC of 93.42% highlight its robustness in classification tasks, slightly outperforming random forest in terms of discriminative ability. This performance suggests that XGBoost, with its gradient boosting mechanism, effectively exploits the refined features selected by E-CAE to enhance classification accuracy.

The KNN classifier also showed competitive performance, achieving an accuracy of 87.22%. Its recall was notably high at 97.59%, indicating a strong ability to detect Parkinson's cases. However, its precision was relatively lower at 86.63%, reflecting a higher false positive rate compared to Random Forest and XGBoost. The F1-score of 91.78% and ROC AUC of 90.11% affirm the model's overall effectiveness, although with a slight compromise in precision.

LR achieved an accuracy of 86.78%, with precision and recall both recorded at 90.96%, leading to a balanced F1-score of 90.96%. Its ROC AUC of 90.49% demonstrates its strong discriminative capability. The model's consistency across precision, recall, and F1-score indicates that logistic regression performs reliably with the feature set refined by the E-CAE method.

The SVM classifier attained an accuracy of 84.14%, with a precision of 82.83% and an exceptionally high recall of 98.80%. This significant recall suggests that SVM was highly effective in identifying true positive cases of Parkinson's disease. However, the lower precision resulted in an F1-score of 90.11% and an ROC AUC of 87.61%, indicating room for improvement in balancing false positives and negatives.

The DT classifier achieved an accuracy of 80.62%, with a precision of 87.65% and a recall of 85.54%. Its F1 score was 86.59%, and the ROC AUC stood at 76.38%. Although Decision Tree performance improved with the E-CAE method, it remained less effective compared to ensemble methods like Random Forest and XGBoost. This result suggests that Decision Tree models may struggle with the high-dimensional feature space, even after feature optimization.

The Naive Bayes classifier reported the lowest accuracy among the E-CAE models at 77.53%. However, it achieved a precision of 86.16% and a recall of 82.53%, resulting in an F1 score of 84.31%. Its ROC AUC was relatively low at 74%, indicating limitations in its ability to separate classes effectively. This result suggests that the Naive Bayes classifier may not fully capitalize on the complex feature interactions captured by the E-CAE method.

Results for Dyslexia Dataset

The classification performance of the E-CAE method on the dyslexia dataset is presented and compared in Table 4. The results indicate varying levels of effectiveness across different classifiers, highlighting both the strengths and limitations of the E-CAE approach for dyslexia prediction (Figure 3).

Table 3: Comparative analysis of Parkinson's classification results

Classification techniques	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*ROC AUC (%)*
Nazmun *et al*., (2021)	82.35	80	83	82	-
Faisal *et al*., (2022)	88.3	88.3	88.3	88.3	-
ECAE–KNN	87.22	86.63	97.59	91.78	90.11
ECAE–Naive Bayes	77.53	86.16	82.53	84.31	74
ECAE–LR	86.78	90.96	90.96	90.96	90.49
ECAE–RF	88.66	88.40	96.39	92.22	92.62
ECAE–SVM	84.14	82.83	98.80	90.11	87.61
ECAE–XGBoost	87.67	87.91	96.39	91.95	93.42
ECAE–DT	80.62	87.65	85.54	86.59	76.38

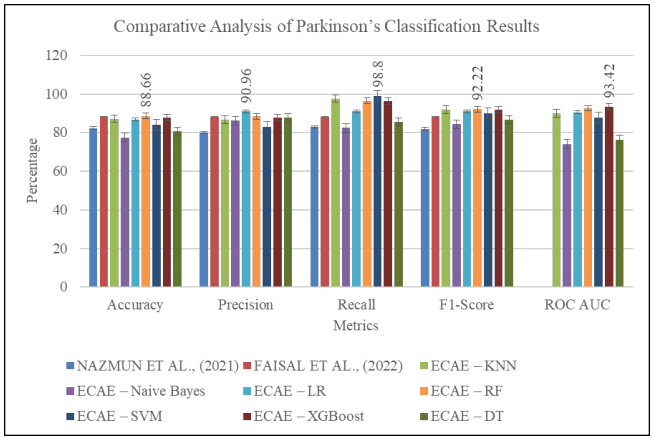

Figure 3: Comparative analysis of Parkinson's classification results

The benchmark studies conducted by Tabassum *et al*. (2023) and Vanitha & Kasthuri (2023) reported accuracies of 90.5 and 89.8%, respectively. However, the model by Vanitha & Kasthuri (2023) exhibited low precision at 54.2% and recall at 34.7%, which resulted in a modest F1-score of 42.3%. This performance suggests challenges in effectively distinguishing between dyslexic and non-dyslexic cases, despite achieving high accuracy.

The E-CAE method, when integrated with LR, outperformed the benchmark models with an accuracy of 90.86%. It achieved a precision of 67.14%, indicating a balanced rate of correct positive predictions. However, the recall stood at 37.90%, reflecting moderate sensitivity in identifying dyslexic cases. The F1-score of 48.45% and ROC AUC of 84.27% demonstrate a well-rounded performance, confirming the model's robustness in handling complex data distributions inherent in dyslexia datasets.

The XGBoost classifier closely followed, achieving an accuracy of 90.77%. Its precision was 75.56%, suggesting that most of the positive predictions were correct. However, its recall was limited to 27.42%, which slightly constrained its overall performance. Despite this, the F1-score of 40.24% and a higher ROC AUC of 86.74% highlight the model's strong discriminatory ability, making it effective for dyslexia detection.

The RF classifier also delivered competitive results with an accuracy of 90.31%. Its precision was the highest among all classifiers at 78.13%, reflecting strong performance in predicting true positive cases. However, the recall dropped significantly to 20.16%, indicating challenges in capturing all dyslexic instances. The F1-score of 32.05% and an ROC AUC of 83.35% emphasize the model's high precision but also its need for better sensitivity.

In contrast, the SVM classifier demonstrated an accuracy of 88.85%. Despite achieving a high precision of 75%, its recall was remarkably low at 2.42%, leading to a minimal F1-score of 4.69%. This outcome suggests that while SVM is highly conservative in predicting dyslexia, it fails to identify a significant portion of positive cases. The ROC AUC of 82.99% further indicates limited overall classification capability.

The performance of the KNN classifier was moderate, achieving an accuracy of 86.93%. Its precision was 36.62%, but the recall was relatively low at 20.97%, resulting in an F1-score of 26.67%. The ROC AUC of 75.80% reflects the model's limited ability to differentiate between classes. This suggests that KNN struggled with the high-dimensional nature of the dataset, even after E-CAE feature optimization.

The DT classifier showed the lowest performance among the ensemble models, with an accuracy of 84.55%. Its precision of 31.40% and recall of 30.65% resulted in a balanced but low F1 score of 31.02%. Additionally, the ROC AUC of 61.04% reveals significant limitations in its classification ability. This suggests that the decision tree model could not effectively handle the complex and high-dimensional feature space of the dyslexia dataset.

The most notable underperformance was observed with the Naive Bayes classifier. It recorded the lowest accuracy of 58.78%. Interestingly, its recall was quite high at 80.65%, which indicates that it was effective in identifying dyslexic cases. However, its precision was extremely low at 18.98%, suggesting a high false positive rate. The F1-score of 30.72% and ROC AUC of 69.20% highlight the model's severe imbalance between sensitivity and specificity, limiting its overall effectiveness.

Discussion

The overall analysis of the E-CAE method across the Alzheimer/Darwin, Parkinson's, and dyslexia datasets demonstrates its capability to improve classification performance through effective feature selection. However, the effectiveness of the feature selection is highly dependent on the classifier used. The study shows that no

Table 4: Comparative analysis of dyslexia classification results

Classification Techniques	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*ROC AUC (%)*
Tabassum *et al*. (2023)	90.5	-	-	-	-
Vanitha & Kasthuri (2023)	89.8	54.2	34.7	42.3	84.20
ECAE – KNN	86.93	36.62	20.97	26.67	75.80
ECAE – NAIVE BAYES	58.78	18.98	80.65	30.72	69.20
ECAE – LR	**90.86**	67.14	37.90	**48.45**	84.27
ECAE – RF	90.31	78.13	20.16	32.05	83.35
ECAE – SVM	88.85	75	02.42	04.69	82.99
ECAE – XGBOOST	90.77	75.56	27.42	40.24	**86.74**
ECAE – DT	84.55	31.40	30.65	31.02	61.04

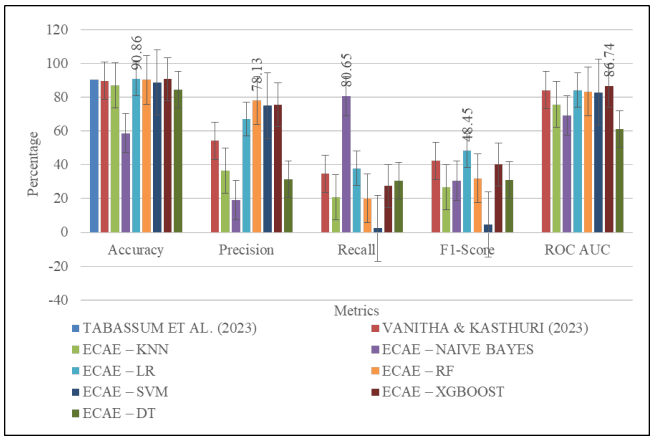

Figure 4: Comparative analysis of dyslexia classification results

single classifier consistently outperformed others across all datasets, highlighting the complexity of feature selection and classification in medical datasets.

For the Alzheimer/Darwin dataset, the Naive Bayes classifier, combined with the E-CAE method, delivered the highest accuracy of 86.79%. This performance was closely followed by RF and SVM, both achieving an accuracy of 84.91%. The high ROC AUC scores of 95.64% for RF and 95.29% for SVM indicate strong classification capabilities, suggesting that ensemble and margin-based classifiers can effectively utilize the optimized features selected by E-CAE for this dataset.

In the Parkinson's dataset, the Random Forest classifier again led with an accuracy of 88.66%, along with high precision and recall scores, demonstrating its robustness in handling high-dimensional data. XGBoost also performed exceptionally well, achieving an accuracy of 87.67% with a superior ROC AUC of 93.42%, highlighting its strength in leveraging gradient boosting for complex feature interactions. LR also demonstrated competitive performance with balanced precision and recall.

Conversely, the dyslexia dataset presented more challenges in achieving balanced classification performance. While Logistic Regression and XGBoost achieved the highest accuracies of 90.86 and 90.77%, respectively, these models struggled to maintain high recall rates. This imbalance suggests that, despite effective feature reduction, sensitivity in detecting dyslexia remains limited. Notably, Random Forest achieved the highest precision of 78.13%, but its recall was significantly lower, impacting overall detection performance.

Random Forest along with XGBoost demonstrated superior performance than other classifiers and E-CAE in all datasets because of both models' exceptional capabilities to process complex high-dimensional data sets. However, logistic regression demonstrated a strong balance of performance and interpretability, especially for the dyslexia dataset. These results emphasize the importance of aligning feature selection methods with classifier characteristics to maximize predictive accuracy and model robustness across varying datasets.

Conclusion

The study introduced the E-CAE method, which effectively addressed the challenges of high-dimensional medical datasets through robust feature selection. The proposed method integrated multiple correlation metrics to evaluate feature relevance comprehensively. This multi-perspective evaluation improved the selection of informative attributes, contributing to better classification performance across various datasets. The experimental results showed that E-CAE brought substantial performance improvements to various classification techniques. The NB classifier showed its optimal performance on the Alzheimer/Darwin dataset with 86.79% accuracy which exceeded standard methods. For the Parkinson's dataset, the RF classifier obtained an accuracy of 88.66% and a high F1 score of 92.22%, demonstrating its robustness in handling complex and high-dimensional data. Similarly, in the dyslexia dataset, logistic regression and XGBoost achieved the highest accuracies of 90.86 and 90.77%, respectively. Despite these promising results, the recall scores for dyslexia classification remained low, indicating difficulty in identifying all positive cases. The study highlighted the adaptability of the E-CAE method across diverse datasets and classifiers. Ensemble classifiers like RF and gradient boosting models such as XGBoost consistently delivered superior results, emphasizing their ability to exploit the informative features selected by E-CAE. However, classifiers like Naive Bayes and logistic regression provide a balance between performance and interpretability, which is essential for clinical applications.

One limitation of this work was the computational cost associated with calculating multiple correlation metrics, especially for large datasets with thousands of features. Additionally, the E-CAE method, while effective in feature reduction, did not fully resolve the recall imbalance observed in certain datasets, such as dyslexia. This limitation suggests the need for further optimization to balance precision and recall effectively.

Acknowledgments

The authors would really appreciate the help of Bishop Heber College (Autonomous), Tiruchirappalli, Tamil Nadu, India for their support for the research. On that account, the researcher wishes to express his gratitude to the team members for the useful insights and positive influence in the creation of this study. The authors also wish to thank other friends and colleagues who have made important inputs that have enhanced the quality of this work.

References

Akram. P, & Latha, P. H. (2020). Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification. *Health information science and systems*, *8*(1), 13.

Ali, M. Z., Abdullah, A., Zaki, A. M., Rizk, F. H., Eid, M. M., & El-Kenway, E. M. (2024). Advances and challenges in feature selection methods: a comprehensive review. *J Artif Intell Metaheuristics*, *7*(1), 67-77.

Chin, F. Y., & Goh, Y. K. (2024). Boosting Cancer Dataset Performance with Mutual Information-Based Feature Prioritization. *Journal of Statistical Modeling & Analytics (JOSMA)*, *6*(1).

Faisal Saeed, Al-Sarem, M., Al-Mohaimeed, M., Emara, A., Boulila, W., Alasli, M., & Ghabban, F. (2022). Enhancing Parkinson's disease prediction using machine learning and feature selection methods. *Computers, Materials and Continua*, *71*(3), 5639-5658.

Faiyazuddin, M., Rahman, S. J. Q., Anand, G., Siddiqui, R. K., Mehta, R., Khatib, M. N., ... & Sah, R. (2025). The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of

Advancements in Diagnostics, Treatment, and Operational Efficiency. *Health Science Reports*, *8*(1), e70312.

Gholampour, S. (2024). Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable. *Machine Learning and Knowledge Extraction*, *6*(2), 827-841.

Hider, M. A., Nasiruddin, M., & Al Mukaddim, A. (2024). Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems. *Revista de Inteligencia Artificial en Medicina*, *15*(1), 1010-1042.

Irene Azzali, Cilia, N. D., De Stefano, C., Fontanella, F., Giacobini, M., & Vanneschi, L. (2024). Automatic feature extraction with Vectorial Genetic Programming for Alzheimer's Disease prediction through handwriting analysis. *Swarm and Evolutionary Computation*, *87*, 101571.

Jha, K., & Kumar, A. (2024). Role of Artificial Intelligence in Detecting Neurological Disorders. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(02), 73-79.

Karim Gasmi, Ammar, L. B., Krichen, M., Alamro, M. A., Mihoub, A., & Mrabet, M. (2024). Optimal Ensemble Learning model for Dyslexia prediction based on an adaptive genetic algorithm. *IEEE Access*.

Kavitha, M., & Kasthuri, M. (2024). Enhanced cost-sensitive ensemble learning for imbalanced class in medical data. *J. Electrical Systems*, *20*(7s), 1043-1053.

Kasthuri, M., & Jency, M. R. (2020). Lung cancer prediction using machine learning algorithms on big data: survey. *International Journal of Computer Science and Mobile Computing*, *9*(10), 73-77.

Khalifa, M., Albadawy, M., & Iqbal, U. (2024). Advancing clinical decision support: the role of artificial intelligence across six domains. *Computer Methods and Programs in Biomedicine Update*, 100142.

Kitova, K., Ivanov, I., & Hooper, V. (2024). Stroke Dataset Modeling: Comparative Study of Machine Learning Classification Methods. *Algorithms*, *17*(12), 571.

Nazmun Nahar., Ara, F., Neloy, M. A. I., Biswas, A., Hossain, M. S., & Andersson, K. (2021). Feature selection based machine learning to improve prediction of Parkinson disease. In *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14* (pp. 496-508). Springer International Publishing.

Reddy, K. P., Satish, M., Prakash, A., Babu, S. M., Kumar, P. P., & Devi, B. S. (2023, October). Machine Learning Revolution in Early Disease Detection for Healthcare: Advancements, Challenges, and Future Prospects. In *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (pp. 638-643). IEEE.

Rello, L., Baeza-Yates, R., Ali, A., Bigham, J. P., & Serra, M. (2020). Predicting risk of dyslexia with an online gamified test. *Plos one*, *15*(12), e0241687.

Shahriar Kaisar, & Chowdhury, A. (2022). Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests. *ICT Express*, *8*(4), 563-568.

Tabassum Jan., & Khan, S. M. (2022). An Effective Feature Selection and Classification Technique Based on Ensemble Learning for Dyslexia Detection. In *Intelligent Communication Technologies and Virtual Mobile Networks: Proceedings of ICICV 2022* (pp. 413-423). Singapore: Springer Nature Singapore.

Usman, O. L., Muniyandi, R. C., Omar, K., & Mohamad, M. (2021). Advance machine learning methods for dyslexia biomarker detection: A review of implementation details and challenges. *IEEE Access*, *9*, 36879-36897.

Vanitha, G., & Kasthuri, M. (2021). Dyslexia prediction using machine learning algorithms–a review. *Int. J. of Aquatic Science*, *12*(2), 3372-3380.

Wilson, A., & Anwar, M. R. (2024). The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing. *International Transactions on Artificial Intelligence*, *3*(1), 97-107.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(1), 56-70.

N. D. Cilia, C. De Stefano, F. Fontanella, A. S. Di Freca, An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis, Procedia Computer Science 141 (2018) 466–471. https://doi.org/10.1016/j.procs.2018.10.141

Biswas, D. (2019). Parkinson's disease speech signal features. Kaggle. https://www.kaggle.com/datasets/dipayanbiswas/parkinsons-disease-speech-signal-features