STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–25.
Published online in International Academic Press (www.IAPress.org)

Minimum Regularized Covariance Distance-Based SMOTE Approaches for Zero-Inflated Datasets Augmented with High-Dimensional Features in Big Data Analytics

Keith R Musara[1*], Edmore Ranganai[2], Charles Chimedza[3], Florance Matarise[1], Sheunesu Munyira[1]

[1] *Department of Mathematics and Computational Sciences, University of Zimbabwe, Zimbabwe*
[2] *Department of Statistics, University of South Africa, South Africa*
[2] *School of Statistics and Actuarial Science, University of the Witwaterrand, South Africa*

Abstract Big data in the credit risk landscape is often characterized by zero-inflated datasets, heterogeneity, and high dimensionality. These data aberrations adversely diminish the computational efficacy of the conventional predictive classifiers. To ensure accurate and reliable predictions, it is crucial to remedy these aberrations, as they may result in bias towards the majority class, sparsity, and computational complexity. The modified Euclidean distance (MED)-based synthetic minority oversampling technique (SMOTE) approaches have been suggested in contemporary literature as countermeasures for zero-inflated datasets coupled with heterogeneity. Despite their mathematical tractability, these approaches substantially fail to effectively capture correlations and variability among features. They are also susceptible to heavy-tailed error distributed data points (outliers) and collinearity, rendering them computationally suboptimal in high-dimensional data spaces. In this study, authors present a novelty of supplanting the MED with modified Mahalanobis distance (MMD) to the variants of SMOTE, enhancing their ability to adequately capture correlations, variability, and heterogeneous features. To mitigate the intricacies posed by these multifaceted data aberrations in high-dimensional data settings, the authors propose the fast minimum regularized coefficient determinant (FMRCD) approach to estimate the parameters of the MMD measure. Therefore, this paper enhances the robustness and computational efficiency of SMOTE-based approaches, by leveraging MMD computed intrinsically to the FMRCD approach, in conjunction with classical predictive classifiers. The empirical evidence suggests that our novelty, demonstrates superior predictions and offers computational stability edge over traditional approaches. These contributions circumvent overwhelming data complexities presented by zero-inflated datasets combined with high-dimensional heterogeneity in modelling big data phenomena.

Keywords Big data, zero-inflated datasets, heterogeneous features, high dimensionality, SMOTE-based approaches, modified Mahalanobis distance (MMD), modified minimum regularized coefficient determinant (FMRCD)

DOI: 10.19139/soic-2310-5070-2964

1. Introduction

1.1. Research Motivation

Global economies are gradually adopting sophisticated spending patterns over prioritizing savings (1). This transition poses a significant risk to the credit risk landscape, especially for those institutions with suboptimal credit risk profiling (2). Consequently, banking regulations mandate the adoption of structured approaches to scrutinize and manage risks using up-to-date data and data analytics approaches that yield superior and stable predictions (3). The advent of big data has revolutionized robust credit risk assessment, greatly enhancing the risk profiling

*Correspondence to: Keith R Musara (Email: musarakeith@gmail.com). Department of Mathematics and Computational Sciences, University of Zimbabwe, Zimbabwe.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

harnessed to assess borrower's creditworthiness (4). The most widely cited predictive classifiers include the logistic regression (LR) (5) and random forest (RF) (6). However, their effectiveness is substantially compromised in real-world big data scenarios, which can be characterized by zero-inflated (class-imbalanced) datasets and high-dimensional heterogeneous features (including both continuous and nominal features) (7). The former occurs when one class, the majority class (non-default), vastly outnumbers the other class, the minority class (defaults) (8), while the latter consists of numerous heterogeneous features (7). The overwhelming data aberrations exhibited by zero-inflated datasets substantially diminish the capabilities of the predictive classifiers due to their inherent symmetric property, resulting in erroneous predictions. These conventional predictive classifiers often prioritize the majority class samples, while accurately prediction for the minority class samples is more crucial, particularly in the credit risk domain (2). The abundant heterogeneous features often leads to insurmountable sparsity known as the curse of dimensionality phenomena (7). Subsequently, this high dimensionality also exacerbates inaccurate prediction performance due to bias inclined toward the majority class. This emphasizes the necessity for statistically sound strategies, including computationally efficient fashioned SMOTE-based approaches integrated with generalizable distance measures in conjunction with the tractable predictive classifiers.

1.2. Adaptations of SMOTE-based approaches for Heterogeneous Features

To proffer some remedy for zero-inflated datasets, the synthetic minority oversampling technique (SMOTE) is broadly acknowledged in contemporary literature due to its tractability, as it exhibits desirable theoretical proprieties (8). Building on the mathematical foundations of the classical SMOTE approach, Chawla et al. (8) proposed SMOTE for nominal and continuous features (SMOTE-NC) by supplanting the conventional Euclidean distance (ED) with the modified ED (MED). The MED incorporates the medians of standard deviations for all continuous features of the minority class in its computation. Despite its robustness, this approach significantly falls short due to its inability to capture mismatched labels inherent in nominal features, as it presumes that these labels are proportional and uniformly distributed. To circumvent this deficiency, Mukherjee and Khushi (9) presented SMOTE for encoding nominal and continuous features (SMOTE-ENC) , which enables the configuration of the encoding mechanism for both feature types. Although the afore-mentioned SMOTE-based approaches are acknowledged for their statistical soundness and appealing mathematical underpinnings, they are well-known to induce distributional overlap into the data. Subsequently, this has been established as the primary source of ambiguous information that can adversely affect the generalizability conclusions (10). This motivated Fonseca and Bacao (11) to introduce the geometric SMOTE-NC features (G-SMOTE-NC), which use the hypersphere approach coupled with a geometric fashion approach as an alternative to the linear interpolation mechanism adopted by several variations of SMOTE-based approaches. This approach aims to effectively capture heterogeneous features and robustly discards ambiguous instances (distributional overlap), demonstrating superior capability against noisy information. However, G-SMOTE-NC still exhibit certain limitations of SMOTE-NC by being considerably unable to accurately capture mismatched labels intrinsic to nominal features, thereby compromising the efficacy of the predictive classifiers This prompted Musara et al. (2) to suggest the SMOTE edited nearest neighbors (SMOTEENN) for encoding both nominal and continuous features (SMOTEENN-ENC). This hybrid approach is robust in discarding noisy information and adequately captures mismatched labels inherent in nominal features. Conversely, this approach is highly computationally infeasible, particularly in handling high-dimensional features due to the complex iterations performed by the ENN approach (12).

1.3. Robust Distance Measures

Despite the statistical efficiency demonstrated by the MED for SMOTE-based approaches typically tailored for heterogeneous features in the literature, its effectiveness is often curtailed by its analytical intractability in high-dimensional feature settings (13). Additionally, the MED measure's other drawbacks are that it fails to sufficiently capture variability and correlations among features (14), as well as its susceptibility to heavy-tailed error distributed data points and collinearity (15), particularly in high-dimensional data scenarios. The MD, a scale invariant measure (16), originally designed to effectively capture correlations and variability between features is often preferred to the ED measure in the contemporary literature for the classical SMOTE approach (13), as it utilizes these crucial information exhibited by the features. This measure has garnered significant attention due

to its applicability across a wide spectrum of real-world applications. Furthermore, the MD exhibits excellent computational performance and achieves exponential convergence (16). However, the MD remains susceptible to heavy-tailed error distributed data points, collinearity and heterogeneous features as it inaccurately captures nominal features.

In this article, the authors extended the MD to the modified MD (MMD), which incorporates the median of the standard deviations for all continuous features in the minority class as a remedy for heterogeneous features. However, the MMD still preserves some of the flaws exhibited by the MD, particularly its susceptibility to heavy-tailed error distributed data points and collinearity, which adversely affect the location (mean) and scatter (covariance) parameters (17). To counter these data aberrations, Rousseeuw (18) suggested employing the minimum covariance determinant (MCD) estimator to estimate the location and scatter parameters, owing to its robustness, high efficiency, and equivariance in flagging heavy-tailed error distributed data points in low-dimensional data. This estimator is well-known for its tractability in discarding heavy-tailed error distributed data points (19). However, the classical MCD estimator is time-consuming in handling high-dimensional data, due to its computational intensity, susceptibility to collinearity, as well as singularity of the scatter matrix. To enhance the computational efficiency of the MCD estimator, Rousseeuw and Leroy (20) proposed the FAST-MCD (FMCD) estimator. Despite its computational efficiency, this estimator remains susceptible to high-dimensionality due to the ill-conditioning of the scatter matrix. This motivated Boudt et al. (21) to put-forward the minimum regularized covariance determinant (MRCD) estimator, which robustly estimates location and scatter parameters in high-dimensional data. Additionally, the MRCD is customized to mitigate the singularity (collinearity) of the covariance matrix by inducing a regularization parameter, ensuring its positive definiteness even in the high-dimensional space (21). This greatly enhances stability and reliability of the generalizability of the predictive classifiers (22). To enhance robustness and computational efficacy simultaneously, authors propose the MMD computed intrinsically to the Fast-MRCD (FMRCD) estimator in SMOTE-based approaches primarily designed to effectively heterogeneous features. This novelty ensures statistical soundness and better computational performance in handling zero-inflated datasets, particularly in tandem with high-dimensional heterogeneous features, especially in the big data domain.

1.4. Benchmark Predictive Classifiers and Suggested Methodology

To evaluate the efficacy of the suggested modelling approach, this article employed the LR model and the random forest (RF) algorithm. The popularity of LR algorithm stems from its simplicity, tractability, and interpretability, as well as its ability to capture the linear features inherent in real-world datasets (23). On the other hand, the RF algorithm is highly regarded across a wide spectrum of real-life applications owing to its greater efficiency, ability to accurately capture non-linear features, interpretability, and generalizable predictive performance (minimization of generalization errors) in high-dimensional data, resulting in statistically sound predictions (25). However, high dimensionality also exacerbates bias toward the majority class (24), which may result in diminished predictive capabilities of the classifiers, ultimately leading to the bankruptcy of financial lending institutions due to erroneous predictions. To mitigate these data complexities and feature redundancies inherent in high-dimensional data, while enhancing computational efficiency, the LR classifier is regularized with the adaptive elastic net (AE-NET) penalty (26) and the RF algorithm uses the recursive feature elimination (RFE)-based algorithm (27), resulting in the identification of key features. Thus, these hybrid approaches have been suggested in the contemporary literature due their computational efficacy. The AE-NET is well-known to remedy collinearity, minimize over-parameterization, and exhibit oracle properties (26), as well as selects groups of uncorrelated features and ensures stability in handling high-dimensional data settings unlike adaptive least absolute shrinkage and selection operator (ALASSO). The RFE-based approach has demonstrated effectiveness in real-world application as a feature selection approach, offering advantages such as dimensionality reduction, computational efficacy, and interpretability (27).

1.5. Contribution of the Study

The contributions of this study are premised on the following points:

- To the best of the authors' knowledge, there is a significant scarcity in the literature regarding the circumvention of data aberrations posed by zero-inflated datasets coupled with high-dimensional heterogeneous features in the big data space, particularly in the credit risk landscape. However, these data aberrations intricacies substantially diminish the generalizability of the conventional predictive classifiers.
- Recently, the data complexities posed by zero-inflated datasets blended with heterogeneous features have been addressed in the literature by generalization of the SMOTE-based variations leveraging the MED, which effectively capture these heterogeneous features. Additionally, the MED is well-known for its computational efficiency owing to its simplicity and robustness in handling low-dimensional data scenarios, thereby leaving a gap in the high-dimensional data scenarios.
- Despite its mathematical tractability, the MED measure falls considerably short due to its inability to capture correlations and variability among features, as well as its susceptibility to heavy-tailed error distributed data points and collinearity, especially in high-dimensional data settings. It this context, the modifications of SMOTE-based approaches, typically designed to effectively handle heterogeneous features are compromised.
- To achieve both robustness and greater efficiency, in this paper, authors propose the interplay of the MMD computed intrinsically to the FMRCD estimator in the adaptations of SMOTE designed to accurately capture correlations, variability, and heterogeneous features in high-dimensional data. Additionally, this framework robustly discards outliers, mitigates collinearity, and enhances the computational speed, especially in high-dimensional data settings.
- To evaluate the effectiveness of the proposed architecture, we employed the benchmark statistical approach, the LR model and the baseline ML approach, the RF algorithm. The former is intuitively appealing due to its simplicity and exhibition of good theoretical properties, while the latter is highly regarded due to its tractability ability to adequately capture non-linear features and greater efficacy in high-dimensional data settings.
- To counter data complexities and redundant features in high-dimensional data, the LR was regularized with AE-NET, while the RF algorithm was integrated with the RFE-based algorithm. The former with its oracle property is typically designed for linear classification, as well as for mitigating collinearity, minimizing overfitting, selection of groups of uncorrelated features and ensuring stability in handling high-dimensional data settings, while the latter is designed for optimal feature selection, minimization of complexity, enhancing computational efficacy, and improving efficiency in managing high-dimensional data scenarios.
- To enhance the robustness and effectiveness of SMOTE-based approaches, which are typically designed for heterogeneous features, we combined MMD computed intrinsically with the FMRCD estimator in conjunction with the hybridization of the LR and the RF algorithm with AE-NET and RFE feature selection approaches, respectively. The proposed framework is designed to mitigate data aberrations posed by zero-inflated data combined with high-dimensional heterogeneous features, particularly in big data scenarios.

The rest of the article unfolds as follows: Section 2 presents the theoretical properties of SMOTE-based approaches, suggested framework and traditional predictive classifiers. Section 3 discusses the data features and methodological procedures. Section 4 presents the empirical results. Section 5 discusses the conclusions of the research findings.

2. Mathematical Framework

2.1. Related Theoretical Literature

The conventional SMOTE approach intelligently generates synthetic instances of the minority class rather than replicating existing instances, thereby minimizing the risk of overfitting. This approach selects a minority class instance and generates synthetic instances using a linear interpolation mechanism that connects the instance to

its k-nearest class neighbors derived from the ED (8). The SMOTE approach is typically designed to capture continuous features exclusively, but it substantially fails to accurately capture nominal features; subsequently, this approach generates new labels for these nominal features (9). Thus, SMOTE-NC is primarily tailored to treat nominal features analogously to continuous features, while preserving data integrity by leveraging the MED (8). Building on these advancements, SMOTE-ENC is intrinsically customized to differentiate the relationships between the labels of a particular nominal feature (9), allowing for a more precise representation of the feature's contribution to distance measures through the one-hot encoding method. Similarly, G-SMOTE-NC employs a hyper-sphere mechanism to generate instances for nominal features, facilitating the non-linear generation of synthetic instances, while the continuous features are generated using the mode of each feature. Furthermore, SMOTEENN-ENC was introduced to address ambiguous instances and mismatched labels of nominal features simultaneously by leveraging the ENN approach and implementing a one-hot encoding mechanism for managing multi-label nominal features, respectively. The MED adopted by the aforementioned approaches is unable to capture correlations and variability between features (17). Subsequently, these SMOTE-based approaches may lose essential information, thereby compromising the robustness of the predictive classifiers. Additionally, MED is sensitive to multivariate heavy-tailed error distributed data, which adversely affect parameter estimation (10).

To address these deficiencies, Xie and Huang (29) suggested leveraging the MD for the most popular hybrid approach in the literature, the SMOTEENN approach. Notwithstanding the greater efficacy demonstrated by MD in practice, it is limited exclusively to continuous features; hence, we extended the MD to MMD by incorporating the median of the standard deviations of all continuous features in the minority class to adequately capture nominal features. The MMD also retains the shortcomings of MD inbeing amenable to multivariate heavy-tailed error distributed data points, through a phenomenon known as the masking effect, which increases its sensitivity to heavy-tailed error distributed data (18). Moreover, the MD is susceptible to collinearity, which can degrade the capabilities of predictive classifiers (17). To overcome these data aberrations, the MCD estimator, has been suggested in the literature to estimate the mean and covariance matrix for the MD (19). Motivated by this consideration, Jung and Choi (30) employed the MD, computed intrinsically with the MCD estimator for the classical SMOTE approach. This estimator attempts to establish a concentrated or tightly distributed subset whose classical covariance matrix has the smallest possible determinant. This minimum determinant approach is considered robust because its estimation is minimal influenced by potential heavy-tailed error distributions. Despite its statistical soundness and mathematical tractability, the brute-force approach of subset selection is computationally infeasible for high dimensional settings (31). To circumvent the complexity posed by MCD, the FMCD has been suggested in the literature, which employs the concentration steps (C-steps) constructs to minimize the computational time of subset selection (31). Albeit its theoretical backing, the estimator is susceptible to invertability in high dimensional space. To ensure a robust estimator in high-dimensional data, the MRCD has been recommended to regularize the covariance matrix by introducing a penalty to the matrix of features, thereby enhancing well-conditionedness across all dimensions (32). Table 1 summarizes the strengths and weakness of the SMOTE-based variants, distance measures, and estimators considered in this study.

2.2. Suggested Modelling Approach

The proposed novelty is outlined in six phases, including the data normalization, selection of the best subsets, regularization of the covariance matrix, estimating the MMD, elimination of heavy-tailed errors in the data distributions, and generation artificial samples. The suggested framework can be summarized as follows:

2.2.1. Data Normalization : It is fundamental to normalize and scale the features using the min-max normalization approach to preserve the underlying distribution and patterns of the data. Given a feature $\boldsymbol{x}$ consisting of samples $\{x_1, x_2, \ldots x_m\}$ where m is the total number of samples in the feature, the normalization procedure involves subtracting the minimum values $x_{minimum}$ from each sample and then dividing by the range. The min-max normalization approach is mathematically defined as follows:

$$x_{normalized} = \frac{x_k - x_{minimum}}{x_{maximum} - x_{minimum}} \quad [1]$$

Table 1. Strengths and weakness of the SMOTE-based variants, distance measures, and estimators

Approach	Merits	Demerits
	SMOTE Variants	
SMOTE	Mitigates risk of overfitting and minimizes generalization errors (9).	Susceptible to nominal features, while tailored exclusively only for continuous features (8).
SMOTE-NC	Intrinsically designed to manage heterogeneous features (both nominal and continuous features) (8).	Generates incorrect labels for nominal features due to mismatched labels and noise introduced during linear interpolation mechanism (9).
SMOTE-ENC	Effectively handles mismatched labels inherent in nominal features (8).	Susceptible to ambiguous information induced in generating artificial samples (14).
G-SMOTE-NC	Robust in discarding heavy-tailed error distributed data points arising from distributional overlap (14).	Substantially fail to accurately capture mismatched labels for nominal features (2).
SMOTEENN-ENC	Mathematically tractable in remedying heavy-tailed error distributed data points and adequately captures heterogeneous features (2).	Computationally intensive due to the ENN approach iteration complexity (12).
	Distance Measures	
ED	Ensures simplicity and computational efficiency (2).	Susceptible to heavy-tailed error distributed data points, collinearity, nominal features in high-dimensional settings (17).
MED	Effective in capturing heterogeneous features (2).	Inherits the limitations of the classical ED measure (18).
MD	Accurately captures both correlations and variability among features (17).	Amenable to nominal features, heavy-tailed error distributed data points, and collinearity (17).
MMD	Adequately captures heterogeneous features (proposed novelty)	Sensitive to heavy-tailed error distributed data points and collinearity (17).
	Estimator	
MCD	Robust estimator which mitigates heavy-tailed error distributed data points (19).	Computationally intensive, especially in high-dimensional data (31).
FMCD	Enhances computational speed in discarding heavy-tailed error distributed data points.	Susceptible to collinearity, particularly in high-dimensional space (20).
MRCD	Ensures that the estimated covariance matrix is invertible, thereby remedying collinearity (21).	Computational infeasible, particularly in high-dimensional data (22).
FMRCD	Ensures computational feasibility in high-dimensional data (proposed novelty).	Computationally demanding for convergence of the covariance matrix subjected to regularization (proposed novelty).

where $k = 1, 2, \ldots m$. The resulting normalized samples retain the original distribution shape but are confined with the range of [0,1].

2.2.2. Subset Optimization : The continuous features from the minority class are selected. Subsequently a subset H_1 that contains h elements from the normalized data matrix $\boldsymbol{x}$ is generated based the formula of h, which is defined as follows:

$$h = \left[\frac{n+p+1}{2}\right] \quad [2]$$

where n and p are number of instances and continuous features, respectively. This procedure ensures that samples are tightly distributed cloud of points in the multivariate space, whose covariance matrix has the smallest determinant such that $\hat{\mu}_1 = \frac{1}{h}\sum_{i\in H_1} x_i$, $\hat{S}_1 = \frac{1}{h}\sum_{i\in H_1}(x_i - \hat{\mu}_i)(x_i - \hat{\mu}_i)^T$.

2.2.3. Regularization of the Covariance Matrix : To counter invertability in the covariance matrix, the covariance matrix is regularized. The regularized matrix is given by :

$$\hat{\boldsymbol{K}}_1 = \lambda \boldsymbol{T} + (1-\lambda) c_\alpha \boldsymbol{S}_1 \quad [3]$$

where λ is the regularization parameter: $0 \leq \lambda \leq 1$ to ensure that $\boldsymbol{K}_1$ is positive, hence invertible and well-conditioned, c_α is a consistent factor that depends on the trimming factor $\alpha = \frac{n-h}{n}$. $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{'}$, where Λ and $\boldsymbol{Q}$ are eigen values and eigen vectors corresponding to matrix $\boldsymbol{T}$, respectively. So we can compute the MD (d_i), formulated as follows:

$$d_i(\boldsymbol{z}_i, \boldsymbol{\mu}_1, \boldsymbol{K}_1) = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{K}}_1^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1) \quad [4]$$

for $i = 1, 2, \ldots, n$.. To enhance the computational efficiency of this novelty, c-step algorithm is employed by sorting the observations in ascending order according to d_i, yielding a permutation π for which:

$$d_\pi(1)_{old} \leq d_\pi(2)_{old} \leq \cdots \leq d_\pi(n)_{old} \quad [5]$$

The next step is to get a subset H_2 which contains random h elements from an observation that has the smallest d_i and compute $\hat{\boldsymbol{\mu}}_2$, $\hat{\boldsymbol{S}}_2$, and $\hat{\boldsymbol{K}}_2$ based on H_2. Recompute $\hat{\mu}_i$, $\hat{S}_i$, $\hat{\boldsymbol{K}}_i$, and $d_i(\boldsymbol{z}_i, \boldsymbol{\mu}_i, \boldsymbol{K}_i)$, until a convergent subset with the smallest value of determinant of the covariance matrix is obtained i.e. $|\boldsymbol{K}_{k+1}| \leq |\boldsymbol{K}_k|$; with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{K}}_2 = \hat{\boldsymbol{K}}_1$, otherwise the iteration process continues until it converges.

2.2.4. Compute the Modified Mahalanobis Distance : The fundamental objective of FMRCD is to substitute a covariance estimate to the subset MMD based on covariance. The FMRCD estimators are obtained from the subset H_{FMMRCD}. This subset is obtained by solving the minimization problem formulated as follows:

$$H_{FMRCD} = \arg\min_{\boldsymbol{H}\in\mathscr{H}} [det(\boldsymbol{K}^{1/p})] \quad [6]$$

where $\mathscr{H}$ is the set which consists of all subsets with size h in the data. Once the H_{FMMRCD} is determined, the FMRCD scatter estimators are determined as follows:

$$\hat{\boldsymbol{\Sigma}}_{FMRCD} = \boldsymbol{D}_x \boldsymbol{Q} \Lambda^{1/2} [\lambda \boldsymbol{1} + (1-\lambda) c_\alpha \boldsymbol{S}_w(H_{FMRCD})] \Lambda^{1/2} \boldsymbol{Q}^T \boldsymbol{D}_x \quad [7]$$

respectively. $S_w(H_{FMMRCD})$ is calculated as follows:

$$\boldsymbol{S}_w(H_{FMRCD}) = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{Q}^T \boldsymbol{S}_z(H_{FMRCD}) \boldsymbol{Q} \boldsymbol{\Lambda}^{-1/2} \quad [8]$$

Compute the MMD measure, which incorporates the continuous features in the minority class and the median of the standard deviations of all continuous features in the minority class. The MMD is mathematically defined as follows:

$$\text{MMD}(\boldsymbol{\mu}^0_{FMRCD}, \boldsymbol{\Sigma}^0_{FMRCD}) = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{FMRCD})^{'} [\hat{\boldsymbol{\Sigma}}_{FMRCD}]^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{FMRCD}) + \hat{m}_i \quad [9]$$

where $\hat{\boldsymbol{\mu}}_{FMRCD} = \hat{\mu}_i$ and m_i is the median of standard deviations for all continuous features in the minority class.

2.2.5. Heavy-Tailed Error Distributed Data Points Detection :To discard the heavy-tailed error distributed data points, we determined the outliers based on the 97.5 percentile of the generated distance. The elimination of outliers often ensures that the predictions are reliable, stable and statistically sound.

2.2.6. Generation of Artificial Samples : To avoid overfitting and minimize generalization errors inherent in oversampling approaches, we utilized the the conventional SMOTE approach, which is well-known to mitigate such undesirable traits.

- Finally, artificial instances are generated in the classical SMOTE approach based on a linear interpolation, which is mathematically defined as follows:

$$p_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i) \quad [10]$$

where $rand(0,1)$ denoting the random numbers obtained from the uniform distribution, x_i are original instances in the minority class, and x_{ij} these are samples randomly selected by MMD after discarding the heavy-tailed error distributed data points, with $j = 1,2,\ldots n$ denoting the randomization.

Table 2 presents a schematically pseudo-code for generating artificial samples by leveraging the SMOTE approaches for heterogeneous features coupled with the MMD-based computed intrinsically to the FMRCD estimator, ensure robust predictions in handling high dimensional data. This framework is appealing from a

Table 2. Pseudocode of the Proposed Framework

Generating Samples for SMOTE-based Approaches in High-dimensional Data
Data Normalization: Initially, the min–max approach is employed to the features, thereby preserving the underlying distribution and patterns of the data.
Selection of the Subsets: The subset selection process is initiated on the continuous features of the minority class to generate a concentrated subset achieving a minimal determinant for the covariance matrix.
Regularization of the Covariance Matrix: To circumvent invertibility of the covariance matrix, a penalty is induced to shrink the matrix. The process occurs until the covariance matrix converges to a minimized determinant using c-step algorithm.
Estimate the MMD: Compute the mean based on the best subset and utilize the optimal covariance, and also incorporate the median of the standard deviations.
Elimination of Heavy-Tailed Error Distributed Data Points: The heavy-tailed error distributed data points generated in achieving the MMD computed intrinsically by the FMCD estimator are discarded.
Generation of Artificial Samples: Finally, the artificial samples are generated by leveraging the interpolation mechanism.

computational standpoint, as its estimators are affine equivariant, asymptotically efficient (32), and tend to exhibit a 50% breakdown point, indicating insensitivity to contaminated data (17). To address the deficiency of the traditional approach, particularly in selecting instances, we employed the MMD-based, computed intrinsically to the FMRCD estimator for the SMOTE-based methods customized to accurately capture correlation and variability in continuous features. Additionally, this framework's merit is to effectively manage heterogeneous features, robustly discard heavy-tailed error distributions, and mitigate collinearity. The FMRCD scatter estimate is location invariant and scale equivariant due to the initial standardization of the process.

2.3. Logistic Regression

Generalized linear models (GLMs) are an extension of the ordinary linear regression model, relaxing the distribution of the response variables to include members of the exponential family (33). The GLM approach

is defined as a monotonic and twice-differentiable function $g(.)$, called a link function, such that:

$$g(\boldsymbol{\pi}(x_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{11}$$

where $\boldsymbol{\pi}(x_i)$ is the probability of success. Let the random variable be defined as $x_i = 1$ if there is a default, and $x_i = 0$ if there is no default, with $p(x_i = 1) = p_i$ and $p(x_i = 0) = 1 - p_i$, such that x_i follows a Bernoulli distribution (p_i). Next let $Y = \sum_{i=1}^{n} x_i$ so that Y is the number of defaults in n independent trials. Then Y has a Binomial distribution, Bin $(n,\ p)$, if all p_i are equal. The vector of probability estimates is subjected to a logistic transformation, and the relationship can be represented by a linear function of heterogeneous features that has undergone a logit transformation:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{12}$$

Maximum likelihood estimation (MLE) has been employed extensively in the literature to estimate the parameters of the LR classifier. Hence, in this study, we used the MLE approach to obtain estimates of the parameters of the classifier. The likelihood function can be expressed as:

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i} \tag{13}$$

The logarithm of the $L(\beta)$ gives the log-likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) - ln(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}})\right] \tag{14}$$

Parameter estimates are obtained by differentiating the log-likelihood function and setting it to zero, followed by solving the equation using the Newton-Raphson method. However, this algorithm can be compromised in high-dimensional data scenarios (34).

2.3.1. Penalization of Logistic Regression : To counter the shortcomings of LR in handling high-dimensional features, we suggest exploring the penalization procedures that have been acknowledged in the literature as remedies since they continue to profilerate due to their effectiveness. Arguably, the mostly predominantly employed penalization procedure in the literature is the LASSO, due to its sound and attractive theoretical backing of executing both continuous shrinkage and automatic feature selection simultaneously (35). LASSO is typically designed to shrink some coefficients toward zero while leaving others not entirely shrunk to zero, resulting in the creation of a sparse classifier. The LASSO-penalized LR procedure, denoted as LR-LASSO is a feature selection and penalization procedure that leverages LASSO based on the l_1-norm penalty term. The LR-LASSO procedure is given by a minimization problem:

$$\hat{\boldsymbol{\beta}}^{LR-LASSO} = \arg\min_{\boldsymbol{\beta}}\left\{-l(\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{15}$$

where $\lambda\sum_{j=1}^{p}|\beta_j|$ is the l_1-penalty, and $\lambda \in [0,1]$ is the tuning parameter that controls the amount of shrinkage in the l_1 penalty. This procedure is well-known for significantly outperforming ridge regression, which is typically designed stochastically to shrink all parameters estimates towards zero but not entirely to zero, thereby failing to minimize overfitting unlike the former. However, LASSO is amenable to high-dimensional features as it tend to select a group of highly correlated features (36). To circumvent this gap, Zou and Hastie (36) introduced the elastic net (E-NET), a hybrid penalization procedure that synthesizes the desirable properties of both LASSO regression (l_1-penalty) and ridge regression (l_2-penalty). This procedure has been extensively employed in the literature owing to its computational efficacy compared to LASSO. The E-NET-penalized LR procedure, denoted as LR-E-NET is a feature selection and penalization procedure that utilizes the E-NET procedure based on the l_1-norm and l_2-norm

squared penalty terms. The LR-E-NET procedure is given by a minimization problem:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LR-E-NET}} = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\beta) + \lambda [\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2] \right\} \quad [16]$$

where $\alpha \in [0,1]$ is the parameter that balances the effects of the l_1 and l_2 (from ridge regression) penalties. Clearly, when $\alpha = 0$, LR-E-NET reduces to LR penalized by ridge regression, and when $\alpha = 1$, it reduces to LR-LASSO. Additionally, it offers more statistically sound solutions to the weaknesses of LASSO by selecting groups of uncorrelated features and minimizing over-penalization by including a hyper-parameter (36). E-NET offer a better superior edge over the conventional LASSO. However, these procedures penalize parameter coefficient estimates equally, leading to inconsistencies in feature selection and do not exhibit the oracle properties (37), (38). Additionally, LASSO and E-NET are susceptible to high-dimensional data scenarios (36).

To achieve oracle properties, the adaptive versions of penalization procedures have been suggested in the literature. This prompted, Zou (39) and Zou and Zhang (40) propose adaptive LASSO (ALASSO) and adaptive E-NET (AE-NET) respectively, which aims to eliminate bias in the LASSO estimator using data-driven weights to apply different amount of shrinkage on different coefficients. The ALASSO-penalized LR procedure is referred to as LR-ALASSO, where by ALASSO is the feature selection and penalization procedure combined with LR. The LR-ALASSO procedure is formulated as a minimization problem:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LR-ALASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\beta) + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \quad [17]$$

where $\boldsymbol{w}$ is the adaptive weight such that $\boldsymbol{w} = (w_1, ..., w_p)^{'}$ is a $p \times 1$ weight vector, and the tuning parameter $\lambda_j = w_j \lambda$, for $j = 1, 2, ...p$. The weights depend on the consistent initial values of $\hat{\boldsymbol{\beta}}$ and are given by $w_j = (|\hat{\beta}_j|)^{\gamma}$, where γ is a positive constant. This procedure enjoys oracle properties of the smoothly clipped absolute deviation (SCAD), which guarantees outstanding performance in high-dimensional data settings, as well as a computational advantage due to the different penalization of the parameter estimates that rely on suitably chosen data-driven weights (37). However, this procedure inherits instability from LASSO in handling high-dimensional features (41). On the other hand, the AE-NET integrates the desirable properties of quadratic penalization and adaptively weighted LASSO shrinkage. The AE-NET penalized LR procedure, is referred to as LR-AE-NET, is a feature selection and penalization procedure that utilizes the AE-NET. The LR-AE-NET procedure is formulated as a minimization problem:

$$\hat{\boldsymbol{\beta}}^{\mathrm{AE-NET}} = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\beta) + \lambda \left[\alpha \sum_{j=1}^{p} w_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} w_j \beta_j^2 \right] \right\} \quad [18]$$

where weights are also constructed as $w_j = (|\hat{\beta}_j|)^{\gamma}$ similar to Equation 17. LR-AE-NET reduces to LR-ALASSO for $\alpha = 1$ and to the LR penalized by adaptive ridge regression for $\alpha = 0$. The AE-NET achieves the oracle property, which guarantees its robustness in handling high-dimensional features (40). Hence, the AE-NET is well known to greatly outperform A-LASSO and E-NET, particular in high-dimensional data scenarios. To ensure computational efficiency in penalization of coefficients of parameters in this study, we employed the tuning parameter, $\lambda = \alpha = 0.5$.

2.4. Random Forest

Recently, in the literature, machine learning (ML) algorithms have been regarded as the gold standard in the predictive data analytics domain due to their efficacy, tractability, and efficiency, particularly ensemble-based ML algorithms compared to single-based ML algorithms. The RF algorithm is typically an ensemble-based ML algorithm, which has been widely cited as a remedy for the high-dimensional feature space and offering robust

statistical statistical efficacy among ML approaches (42). This algorithm has received significant attention in the literature due to its ability to effectively handle heterogeneous features, large-scale datasets, and complex data structures as well as mitigate against heavy-tailed error distributed data points (43). This algorithm generates numerous decision trees through random sampling of data using bootstrap samples and by randomly selecting input features. Each decision tree is considered a simple decision tree. Particularly, each tree, T_b, built on a bootstrapped sample with feature randomization at each split, produces a prediction $\hat{\mu}_b(\boldsymbol{x})$ for a new point $\boldsymbol{x}$. This prediction is the fitted value of the leaf node in a tree, T_b, where $\boldsymbol{x}$ resides, i.e., $\hat{\mu}_b(\boldsymbol{x}) = \hat{\mu}_{R_m}(\boldsymbol{x})$. The RF algorithm's prediction is the ensemble average across all trees, defined as follow:

$$\mu(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(\boldsymbol{x}) \quad [19]$$

where B denotes the randomized trees. This aggregation minimizes the variance by leveraging the predictions. The prediction can also be contextualized as a kernel-weighted average, expressed as:

$$\hat{\mu}(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i(\boldsymbol{x})y_i \quad [20]$$

where α_i denotes the weights, which are determined by the number of times the observation, i, appears in the same leaf as $\boldsymbol{x}$ across all trees in the forest, normalized by the number of trees. The RF algorithm has demonstrated stability and flexibility, offering accurate and reliable predictions in this domain (44). Despite its intuitive appeal and popularity, this algorithm is amenable to zero-inflated datasets, which compromises its capabilities (47).

2.4.1. Recursive Feature Elimination Despite the robustness of the RF algorithm in high-dimensional data analysis, it may be amenable to complexities due to redundant features. To circumvent these data complexities, RFE-based approach was employed with RF algorithm, which selects the key important features (27). The RFE algorithm estimates the feature importance from the training model by leveraging the mean decrease in accuracy (MDA) or mean decrease in Gini (MDG). The former demonstrates the algorithm's accuracy decreases from permuting the values of each feature, whereas the latter exhibits means of features total impurity, weighted by the proportion of instances reaching that node in each single-based decision tree. In this study, MDA was employed for features selection due to its robustness and efficiency in a real-world applications. The MDA is mathematically defined as:

$$W_{RF}(x_i) = \frac{\sum_{i\in B} VI^i(x_i)}{Number\ of\ estimators} \quad [21]$$

where B denotes the out-of-bag instances of a tree t, and VI is the importance of feature x_i in tree t. The RF algorithm combined with RFE approach is denoted as RF-RFE-based algorithm which is a feature selection procedure that leverages RF algorithm based on the MDA approach. The RFE algorithm is recognized as a greedy search strategy, as it does not explore all possible feature combinations exhaustively but rather selects locally optimal features at each iteration aiming toward a globally optimal feature subset (45). Consequently, this greed methodology substantially enhances computational efficiency compared to exhaustive evaluations, which can quickly become computationally intensive for high-dimensional data (46).

2.5. Hyperparameter Selection

In this article, we adopted 5-fold cross-validation (CV) for the random search approach to select the optimal hyperparameters for the LR algorithm and RF algorithm, which are used in practice (47). Table 3 presents the search spaces of the hyperparameters leveraged for the aforementioned algorithms. Hyperparameters are significant in controlling the algorithm's behaviour, including its complexity and learning speed. As a result, this greatly enhances the predictive performance of the classifier, thereby leading to accurate and reliable predictions.

Table 3. Hyperparameters for the LR algorithm and RF algorithm

Algorithm	Hyperparameter	Search Space	Reference
LR	Solver	{liblinear, lbfgs}	(48)
	Maximum iteration	{100, 200, 300, 400, 500}	
RF	Number of estimator	{100, 200, 300, 400, 500}	(49)
	Maximum features	{'none', 'sqrt', 'log2'}	
	Maximum depth	{10, 20, 30, 40, 50}	
	Minimum samples split	{2, 4, 6, 8, 10}	
	Minimum samples	{1, 2, 3, 4, 5}	

3. Data and Methods

3.1. Benchmark Datasets

The real-world benchmark datasets suitable for this study consist of zero-inflated binary classes characterized by a minority class representing less than 40% of the total (50). These datasets feature high-dimensional heterogeneous characteristics, with more than 100 features (51), and are classified as big data scenarios, containing a minimum of 20,000 instances (52). In this study, we adopted three widely referenced credit risk datasets in the literature: two from the consumer lending domain and one from the fraud detection landscape. Table 4 reports the data description and characteristics after the data cleaning processes, including the number of instances, the number of features, the skewed ratio and the source from which these datasets were utilized.

Table 4. Data Description

Dataset	Instances	Features		Skewed Ratio	Reference	Source
		Nominal	Continuous			
P2P	340, 571	15	88	3.60%	(53)	(54)
LDCD	20, 000	4	674	8.85%	(55)	(56)
IEEE-CIS	590,540	18	329	3.50%	(57)	(58)

3.2. Performance Evaluation Metrics

The proposed framework was evaluated using results derived from a confusion matrix (CM), where entries (i, j) represent the number of correct and incorrect classifications. Table 5 shows a 2×2 table representing the CM, where the columns represent the predicted results of the proposed architecture and the rows represent the actual classes. The elements on the main diagonal represent the correct number of predictions for the negative and positive

Table 5. Confusion Matrix (CM)

	Predicted negative	Predicted positive
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

classes, while the other entries represent prediction errors. Typically, the accuracy metric is the most extensively employed in the literature for evaluating the efficacy of the predictive classifiers. However, this metric is susceptible to evaluating the predictive classifiers when confronted with zero-inflated binary class. The most widely studied evaluations performance metrics in this domain, include the area under the curve (AUC), geometric mean (GM) and F1-score (59). The AUC metric is mathematically defined as follows:

$$AUC = \frac{S_o - \frac{n_o(n_o+1)}{2}}{n_o n_1} \quad [22]$$

where n_o and n_1 are the numbers of negative and positive instances, respectively, and $S_o = \sum r_i$, with r_i denoting the rank of the i^{th} positive instance. The GM metric is formulated as follows:

$$\text{GM} = \sqrt{Specificity \times Sensetivity} \times 100\% \quad [23]$$

where $specificity = \frac{TN}{TN+FP}$.The F-measure metric is expressed as follows:

$$F1 - score = \frac{(1+\beta^2)}{\beta^2} \times \frac{(precision \times recall)}{precision + recall} \times 100\% \quad [24]$$

where $precision = \frac{TP}{TP+FP}$, $sensitivity\ or\ recall = \frac{TP}{TP+FN}$, and $\beta \in (0,1]$ but β is generally is generally set to one in this study. These evaluation metrics are employed in study, due to their capabilities in handling zero-inflated binary class (60). In this paper, we employed the average macro weights for the evaluation metrics adopted in evaluating the efficacy of the proposed novelty.

3.3. Performance Stability Evaluation Approach

To substantial presents statistically sound evidence on the merit of the proposed approach, we comprehensively considered the combination of the three aforementioned metrics, denoted by AGF, defined as:

$$AGF = (w_1 \times GM + w_2 \times F1 - score + w_3 \times AUC) \quad [25]$$

where $w_1 = w_2 = w_3 = 1$, in this study. The average of the aggregated metrics is susceptible to outliers; hence, the combination of coefficient of variation (CV) and AFG were utilized to assess the dispersion of the proposed approaches.

$$P_{AFG} = \{AFG_0, AFG_1, \ldots, AFG_{r-1}\} \quad [26]$$

$$CV_{AFG} = (\sigma_{AFG}/\mu_{AFG}) \times 100 \quad [27]$$

where $\mu_{AFG} = \sum_{i=0}^{r-1} AFG_i/r$ (denotes the average of all AFG in the P_{AGF} set) and $\sigma_{AFG} = \sqrt{\sum_{i=0}^{r-1}(AFG_i - \mu_{AFG})^2/r}$ (denotes the standard deviation of all AFG in the P_{AGF} set) and r denotes the number of zero-inflated datasets.The larger the CV_{AFG}, the greater the degree of dispersion of the performance metric. Generally, when the CV_{AFG} exceeds 10.00%, the performance of the approach, depicts instability, while when CV_{AFG} is less than 10.00%, the performance of the approach is considered stable (59).

3.4. Experimental Flow Design

The real-world datasets in credit risk landscape are often characterized by zero-inflated observations coupled with high-dimensional heterogeneous features. These complexities often compromised the predictive capabilities of the predictive classifiers due to their inherent symmetric property. To circumvent these data aberrations, contemporary studies have suggested variants of SMOTE-based mechanism typically designed for heterogeneous features through utilization of MED. However, the approaches are susceptible to high-dimensional data contexts; as a result, the predictive classifiers performance are suboptimal. Figure 1 displays the experimental flow design that counter the complexities emanating from high-dimensionality. The proposed framework outlined data partitioning (70% training set based on stratified approach and 30% for testing test), data pre-processing, training of the predictive classifiers, performance evaluations of the predictive and statistical tests.

3.5. Exploratory Data Analysis and Statistical Tests

Initially, we presented the averages of the performance evaluation metrics, which depict the effectiveness of the proposed framework. Nevertheless, the averages of these metrics demonstrate instability due to their susceptibility to heavy-tailed error distributed data points. Thus, we calculated the medians of the performance evaluation metrics, which are less amenable to heavy-tailed error distributed data points. Additionally, we computed the standard

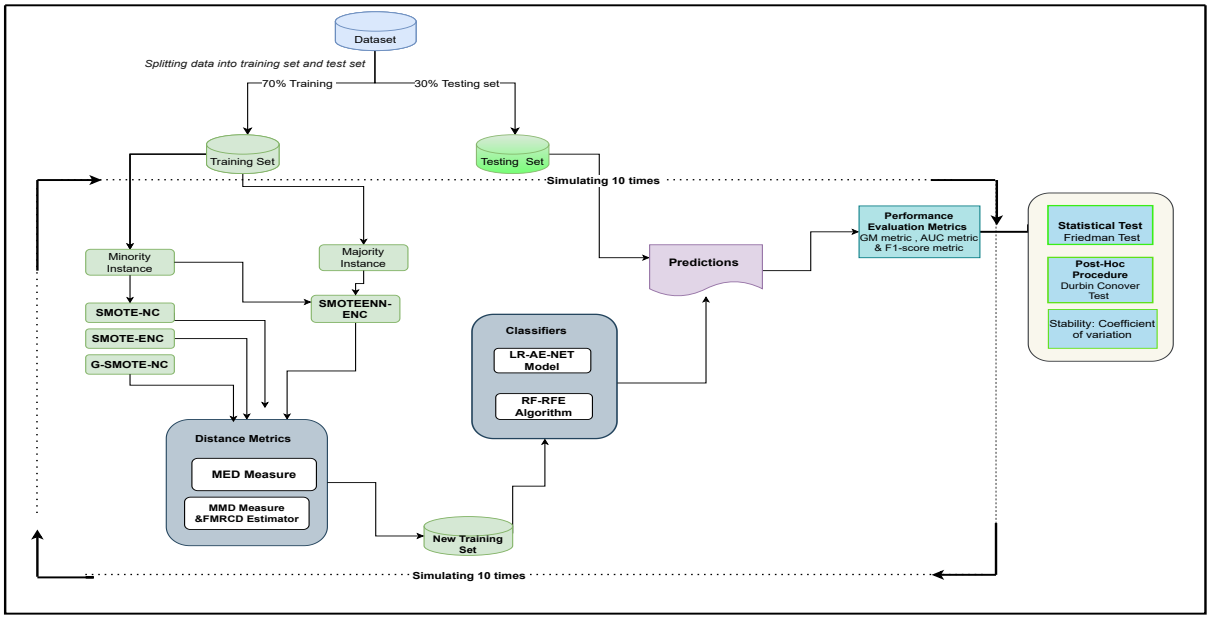

Figure 1. Schematic representation of the proposed framework.

deviations (SDs) for the performance evaluation metrics to assess the variability and stability of the predictions. Finally, we presented the medians of the performance evaluation metrics on violin plots, which display the densities and distributions of these metrics. These evaluation metrics often exhibit non-normal distribution, which violate the assumptions of parametric approach (61). For this reason, the authors validated the efficacy of the proposed architecture, by employing the Friedman test, a non-parametric approach (59). The technique investigates the significant differences in the medians of the performance evaluation metrics for multiple datasets at the 5% level of significance. If the null hypothesis of the Friedman test is rejected, the pairwise comparisons are performed using the Durbin-Conover test (2), while the Bonferroni approach was utilized for adjusting the pairwise *p-values* (62).

3.6. Computational Tools

The experiments were implemented using Python programming (version 3.11), while R programming (version 4.5.2) was employed for data analysis and visualization. All the algorithms were trained and evaluated on a Lenovo T-series notebook featuring an AMD Ryzen 5 Pro 5650U processor with Radeon Graphics at 2.30 GHz and 24 GB of RAM, running Windows 11 Pro.

4. Empirical Results and Analysis

This section presents an empirical generalizable overview of the predictive performance of the LR-AE-NET, and RF-RFE-based algorithm coupled with SMOTE-based approaches typically designed for heterogeneous features, combined with either MED (the conventional approach) or MMD computed intrinsically to the FMRCD approach (the new approach), leveraging zero-inflated binary classes and high-dimensional features.

4.1. Evaluating the efficacy of the LR-AE-NET leveraging the variants of SMOTE-based approaches typically customized for heterogeneous features

Table 6 presents the averages and SDs for AUC metrics achieved by the LR-AE-NET, leveraging datasets pre-processed by adaptations of SMOTE-based approaches primarily tailored for heterogeneous features,

integrated with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive

Table 6. Mean±SD of the AUC metric attained by LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
P2P	MED	0.9837±0.0002	0.9705±0.0066	0.9685±0.0081	0.9732±0.0083
	MMD-FMRCD	0.9814±0.0006	0.9885±0.0024	**0.9898**±0.0007	0.9891±0.0012
IEEE-CIS	MED	0.7822±0.0092	0.7502±0.0421	0.7915±0.0131	0.7657±0.0190
	MMD-FMRCD	0.7910±0.0103	0.7989±0.0109	**0.8199**±0.0047	0.7905±0.0153
LDCD	MED	0.8056±0.0220	0.8910±0.1249	**0.9999**±0.0000	0.9205±0.0708
	MMD-FMRCD	0.7770±0.0147	0.9694±0.0362	**0.9999**±0.0000	0.9647±0.0395

classifier demonstrated its tractability by obtaining the highest average AUC metric of 98.98% (±0.0007) and 81.99% (±0.00047) for the P2P and IEEE-CIS datasets pre-processed by SMOTE-NC blended with the MMD computed intrinsically to the FMRCD estimator. On the other hand, the LR-AE-NET exhibited superior predictive performance by attaining the highest AUC metric of 99.99% (±0.0000) for the LDCD dataset pre-processed by SMOTE-NC, which was integrated either with the MED approach or the MMD computed intrinsically to the FMRCD estimator. To validate the suggested architecture, Figure 2 showcases the median AUC metrics for the three aggregated datasets attained by the LR-AE-NET utilizing datasets pre-processed by variants of SMOTE-based approaches typically designed for heterogeneous features, combined with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive classifier demonstrated premier predictive performance by achieving a maximum median AUC metric of 99.00% using datasets pre-processed with SMOTE-NC combined with the MMD computed intrinsically to the FMRCD estimator. Further analysis of

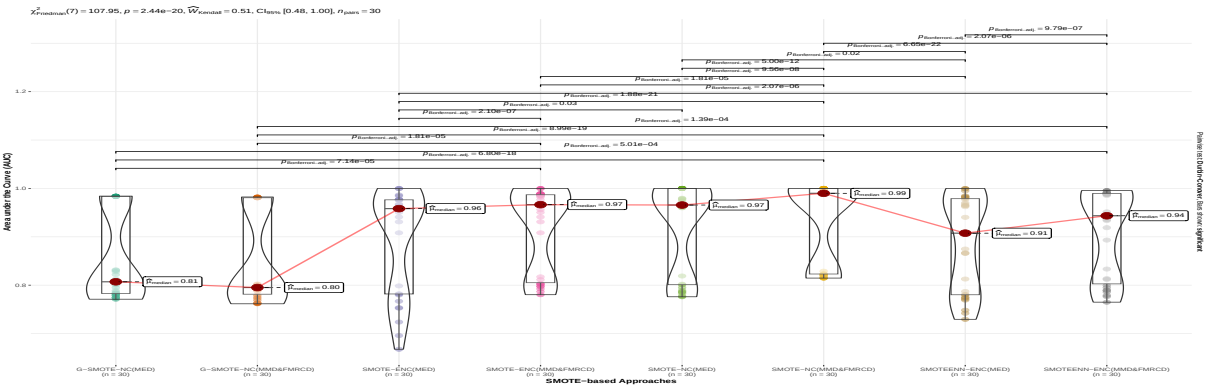

Figure 2. Friedman and Durbin-Conover tests based on AUC metric obtained by the LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

the Friedman test recorded a *p-value* of $2.44 \times 10^{-20} < 0.05$ ($\chi_F^2 = 107.95 > \chi_7^2(\alpha = 0.05) = 2.1670$), indicating that the generalizability of the adaptations of SMOTE-based approaches is statistically significantly different at 5% level of significance. Subsequently to that, the Durbin-Conover test revealed that the superior approach is statistically significantly different from all the variations of the SMOTE-based approaches.

Table 7 reports the averages and SDs for GM metrics attained by the LR-AE-NET using datasets pre-processed by modifications of SMOTE-based approaches typically designed for heterogeneous features, coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive classifier exhibited substantial predictive capabilities by achieving a maximum average GM metric of 97.77% (±0.0025) and 99.91%

(±0.0004) for the P2P and LDCD datasets pre-processed by SMOTE-ENN-ENC and SMOTE-NC, respectively, both combined with the MMD computed intrinsically to the FMRCD estimator. However, the LR-AE-NET demonstrated superior predictive performance by achieving a maximum GM metric of 75.90% (±0.0159) for the IEEE-CIS dataset pre-processed by SMOTE-ENC integrated with the MED approach.

Table 7. Mean±SD of the GM metric achieved by LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
P2P	MED	0.9546±0.0006	0.9524±0.0120	0.9508±0.0135	0.9567±0.0157
	MMD-FMRCD	0.9504±0.0014	0.9773±0.0026	0.9742±0.0028	**0.9777**±0.0025
IEEE-CIS	MED	0.7206±0.0208	0.7069±0.0392	**0.7590**±0.0159	0.7289±0.0182
	MMD-FMRCD	0.7285±0.0078	0.7306±0.0212	0.7574±0.0065	0.7309±0.0106
LDCD	MED	0.7128±0.0173	0.7242±0.1496	0.9979±0.0006	0.8074±0.1239
	MMD-FMRCD	0.6955±0.0115	0.8927±0.0877	**0.9991**±0.0004	0.8814±0.0990

To further scrutinize the merits of the proposed framework, Figure 3 displays the median GM metrics achieved by the LR-AE-NET for the three combined datasets leveraging datasets pre-processed by adaptations of SMOTE-based approaches primarily customized for heterogeneous features, integrated with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive classifier exhibited outstanding predictive capabilities by obtaining the highest median GM metric of 97.00% utilizing datasets pre-processed with SMOTE-NC combined with the MMD computed intrinsically to the FMRCD estimator. Further robust

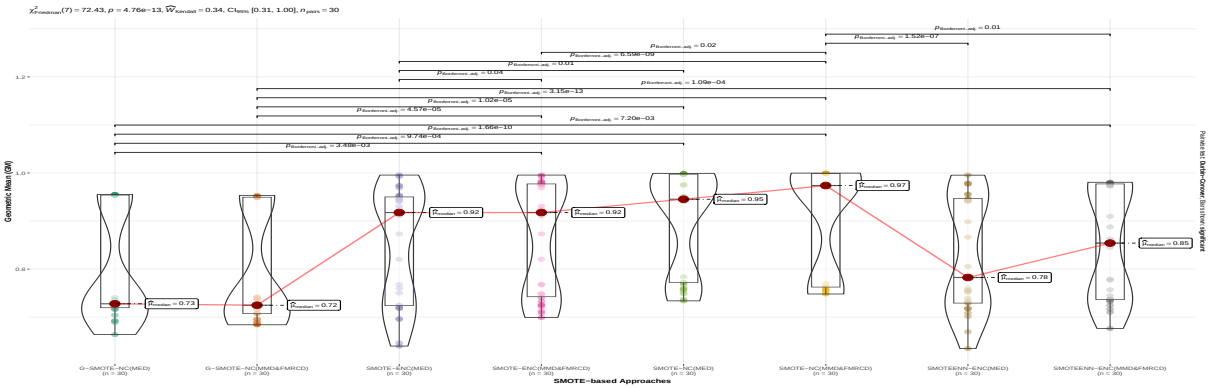

Figure 3. Friedman and Durbin-Conover tests based on GM metric exhibited by the LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

analysis was performed employing the Friedman test, which reported a *p-value* of $4.76 \times 10^{-13} < 0.05$ ($\chi_F^2 = 72.43 > \chi_7^2(\alpha = 0.05) = 2.1670$), suggesting that the efficacy of the variants of SMOTE-based approaches is statistically significantly different at 5% level of significance. Although the effective approach demonstrated premier performance, the Durbin-Conover test showed that it is statistically insignificant different to SMOTE-NC combined with MED approach.

Table 8 shows the averages and SDs for F1-score metrics achieved by the LR-AE-NET, leveraging datasets pre-processed by adaptations of SMOTE-based approaches primarily customized for heterogeneous features, integrated with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive classifier exhibited superior predictive performance by attaining the highest average F1-score of 97.00% (±0.0025) and 99.85% (±0.0004) for the P2P and LDCD datasets pre-processed by SMOTEENN-ENC and SMOTE-NC,

respectively, both integrated with the MMD computed intrinsically to the FMRCD estimator. In contrast, the LR-AE-NET demonstrated a significant predictive performance by obtaining a maximum F1-score metric of 52.68% (±0.0134) for the IEEE-CIS dataset pre-processed by SMOTE-ENC coupled with the MED approach. To substantiate the effectiveness of the proposed novelty, Figure 4 reports the median F1-score metrics for the

Table 8. Mean±SD of the F1-score metric attained by LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	GSMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
P2P	MED	0.9363±0.0004	0.9340±0.0167	0.9317±0.0188	0.9401±0.0222
	MMD-FMRCD	0.9304±0.0019	0.9695±0.0038	0.9650±0.0042	**0.9700**±0.0037
IEEE-CIS	MED	0.4931±0.0173	0.4906±0.0379	**0.5268**±0.0134	0.5014±0.0151
	MMD-FMRCD	0.4988±0.0068	0.4998±0.0109	0.5204±0.0061	0.5023±0.0091
LDCD	MED	0.5759±0.0156	0.6123±0.1752	0.9967±0.0011	0.6960±0.1640
	MMD-FMRCD	0.5604±0.0101	0.8023±0.1475	**0.9985**±0.0006	0.7844±0.1332

three aggregated datasets exhibited by the LR-AE-NET using datasets pre-processed by variants of SMOTE-based approaches primarily tailored for heterogeneous features, combined with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive classifier demonstrated computational efficacy by achieving a maximum median F1-score of 95.00% by utilizing datasets pre-processed with SMOTE-NC integrated with the MMD computed intrinsically to the FMRCD estimator. A comprehensive analysis of the Friedman test recorded

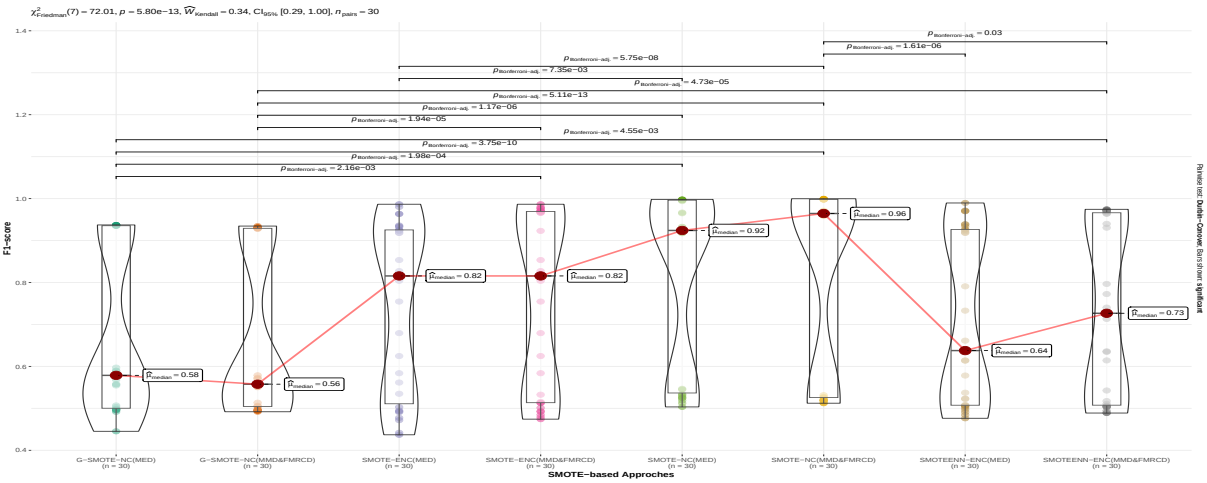

Figure 4. Friedman and Durbin-Conover tests based on GM metric exhibited by the LR-AE-NET leveraging datasets pre-processed by the variants of SMOTE-based approaches typically customized for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator.

a *p-value* of $5.80 \times 10^{-13} < 0.05$ ($\chi^2_F = 72.01 > \chi^2_7(\alpha = 0.05) = 2.1670$), indicating that the effectiveness of the adaptations of SMOTE-based approaches is statistically significantly different at 5% level of significance. Despite the efficacy of the robust approach, the Durbin-Conover test established that it is statistically insignificant to SMOTE-NC combined with the MED approach and SMOTE-ENC integrated with the MMD computed intrinsically to the FMRCD estimator at 5% level of significance.

4.2. Evaluating the robustness of the RF-RFE-based algorithm employing the modifications of SMOTE-based approaches generally tailored for heterogeneous features

Table 9 presents the averages and SDs for AUC metrics exhibited by the RF-RFE-based algorithm, utilizing datasets pre-processed by variants of SMOTE-based approaches typically tailored for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive

Table 9. Mean±SD of the AUC metric achieved by RF-RFE-based algorithm utilizing datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
P2P	MED	0.9572±0.0002	0.9563±0.0017	0.9561±0.0020	0.9586±0.0009
	MMD-FMRCD	**0.9599**±0.0005	0.9536±0.0014	0.9552±0.0029	0.9502±0.0012
IEEE-CIS	MED	0.5768±0.0054	**0.7713**±0.1150	0.7494±0.0069	0.5713±0.0095
	MMD-FMRCD	0.5956±0.0061	0.7605±0.0130	0.7692±0.0062	0.7665±0.0083
LDCD	MED	0.6598±0.0863	0.6618±0.1007	0.7683±0.0148	0.7022±0.0425
	MMD-FMRCD	0.6269±0.0828	0.7542±0.0646	**0.8150**±0.0209	0.7844±0.0496

algorithm demonstrated greater efficacy by achieving a maximum average AUC metric of 95.99% (±0.0005) and 81.50% (±0.0004) for the P2P and LDCD datasets pre-processed by G-SMOTE-NC and SMOTE-NC, respectively, both synergized with the MMD computed intrinsically to the FMRCD estimator. On the other hand, the RF-RFE-based algorithm exhibited excellent predictive performance by attaining the highest AUC metric of 77.13% (±0.1150) for the IEEE-CIS dataset pre-processed by SMOTE-ENC integrated with the MED approach. To substantiate these empirical findings, Figure 5 showcases the median AUC metrics for the three datasets aggregated exhibited by the RF-RFE-based algorithm utilizing datasets pre-processed by variants of SMOTE-based approaches typically designed for heterogeneous features integrated with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive algorithm demonstrated superior efficacy by achieving a maximum median AUC metric of 82.00% by utilizing datasets pre-processed with SMOTE-NC combined with the MMD computed intrinsically to the FMRCD estimator. Further analysis of the Friedman test

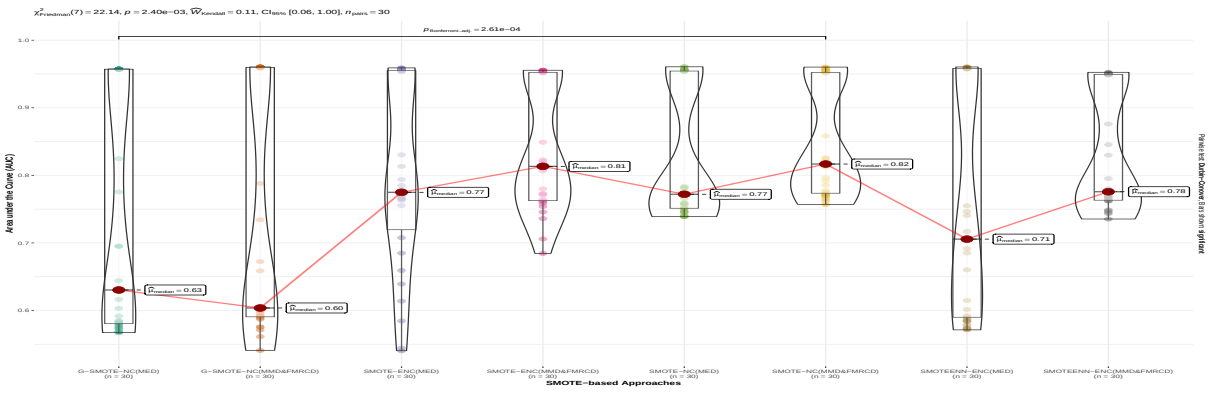

Figure 5. Friedman and Durbin-Conover tests based on AUC metric exhibited by the RF-RFE-based algorithm leveraging datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

recorded a *p-value* of $2.40 \times 10^{-03} < 0.05$ ($\chi^2_F = 22.14 > \chi^2_7(\alpha = 0.05) = 2.1670$), suggesting that the robustness of the variations of SMOTE-based approaches is statistically insignificant different at 5% level of significance. Despite the effectiveness demonstrated by the robust approach, the Durbin-Conover test highlighted that it is statistically insignificant different only to G-SMOTE-NC combined with the MED approach.

Table 10 reports the averages and SDs for GM metrics achieved by the RF-RFE-based algorithm using datasets pre-processed by adaptations of SMOTE-based approaches primarily designed for heterogeneous features combined with either the MED or the MMD computed intrinsically to the FMRCD estimator. The

Table 10. Mean±SD of the GM metric attained by RF-RFE-based algorithm utilizing datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTE-ENN-ENC
P2P	MED	0.9561±0.0020	0.9563±0.0017	0.9586±0.0009	0.9572±0.0002
	MMD-FMRCD	**0.9596**±0.0005	0.9536±0.0014	0.9552±0.0029	0.9502±0.0012
IEEE-CIS	MED	0.7494±0.0069	**0.7713**±0.0115	0.5713±0.0095	0.5768±0.0054
	MMD-FMRCD	0.7692±0.0062	0.7605±0.0130	0.7665±0.0083	0.5956±0.0061
LDCD	MED	0.7683±0.0148	0.6618±0.1007	0.7022±0.0425	0.6598±0.0863
	MMD-FMRCD	0.8150±0.0209	0.7543±0.0646	**0.7844**±0.0496	0.6269±0.0828

predictive algorithm demonstrated premier predictive performance by obtaining a maximum average GM metric of 95.96% (±0.0005) and 78.44% (±0.0496) for the P2P and LDCD datasets pre-processed by G-SMOTE-NC and SMOTE-NC, respectively, both integrated with the MMD computed intrinsically to the FMRCD estimator. Conversely, the RF-RFE-based algorithm exhibited outstanding predictive performance by achieving a maximum GM metric of 77.13% (±0.1150) for the IEEE-CIS dataset pre-processed by SMOTE-ENC coupled with the MED approach. To affirm these results, Figure 6 displays the median GM metrics for the three datasets attained by the RF algorithm. The predictive algorithm demonstrated its effectiveness by achieving the highest median GM metric of 82.00% using datasets pre-processed with SMOTE-NC integrated with the MMD computed intrinsically to the FMRCD estimator. To substantiate these findings, the Friedman test reported a *p-value* of $8.91 \times 10^{-03} < 0.05$

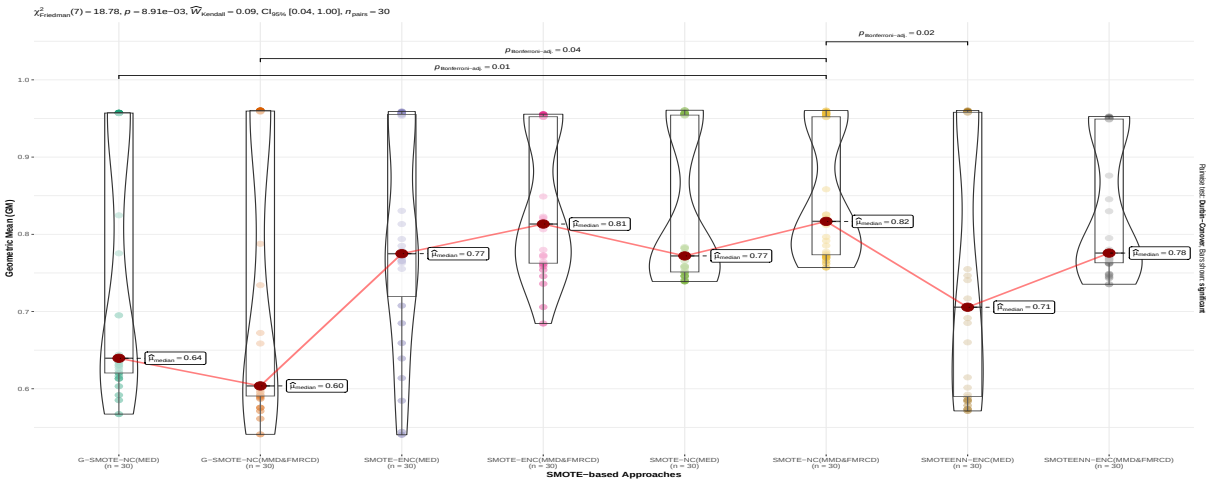

Figure 6. Friedman and Durbin-Conover tests based on GM metric obtained by the RF-RFE-based algorithm leveraging datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

($\chi_F^2 = 18.78 > \chi_7^2(\alpha = 0.05) = 2.1670$), suggesting that the efficacy of the modifications of SMOTE-based approaches is statistically insignificant different at 5% level of significance. However, the Durbin-Conover test revealed that the effective approach is statistically insignificant different to G-SMOTE-NC integrated with MED approach and SMOTEENN-ENC blended with MMD computed intrinsically to the FMRCD estimator.

Table 11 presents the averages and SDs for F1-score metrics achieved by the RF-RFE-based algorithm

using datasets pre-processed by the adaptations of SMOTE-based approaches typically tailored for heterogeneous features, combined with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive algorithm demonstrated robust performance by achieving the highest average F1-score metric of 97.00% (±0.0009) and 87.41% (±0.0209) for the P2P and LDCD datasets pre-processed by G-SMOTE-NC and SMOTE-NC, respectively, both combined with the MMD computed intrinsically to the FMRCD estimator. Alternatively, the RF-RFE-based algorithm attained superior predictive performance by achieving the highest F1-score metric of 65.49% (±0.0074) for the IEEE-CIS dataset pre-processed by SMOTE-ENC integrated with the MED approach.

Table 11. Mean±SD of the F1-score metric achieved by RF-RFE-based algorithm utilizing datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
P2P	MED	0.9690±0.0003	0.9464±0.0045	0.9463±0.0048	0.9698±0.0007
	MMD-FMRCD	**0.9700**±0.0009	0.9397±0.0034	0.9459±0.0076	0.9245±0.0018
IEEE-CIS	MED	0.6249±0.0080	0.6459±0.0206	**0.6549**±0.0074	0.6168±0.0090
	MMD-FMRCD	0.6516±0.0084	0.6237±0.0061	0.6293±0.0021	0.6121±0.0082
LDCD	MED	0.7128±0.0976	0.5500±0.1318	0.8359±0.0137	0.7204±0.0341
	MMD-FMRCD	0.6707±0.1022	0.7408±0.0494	**0.8741**±0.0209	0.6466±0.0992

To ascertain the effectiveness of the proposed framework, Figure 7 displays the median F1-score metrics achieved by the RF-RFE-based algorithm for the three aggregated datasets utilizing datasets pre-processed by the variants of SMOTE-based approaches typically designed for heterogeneous features, integrated with either the MED or the MMD computed intrinsically to the FMRCD estimator. The predictive algorithm demonstrated computational efficacy, achieving the highest median F1-score of 88.00% by utilizing datasets pre-processed with SMOTE-NC combined with the MMD computed intrinsically by the FMRCD estimator. Further analysis

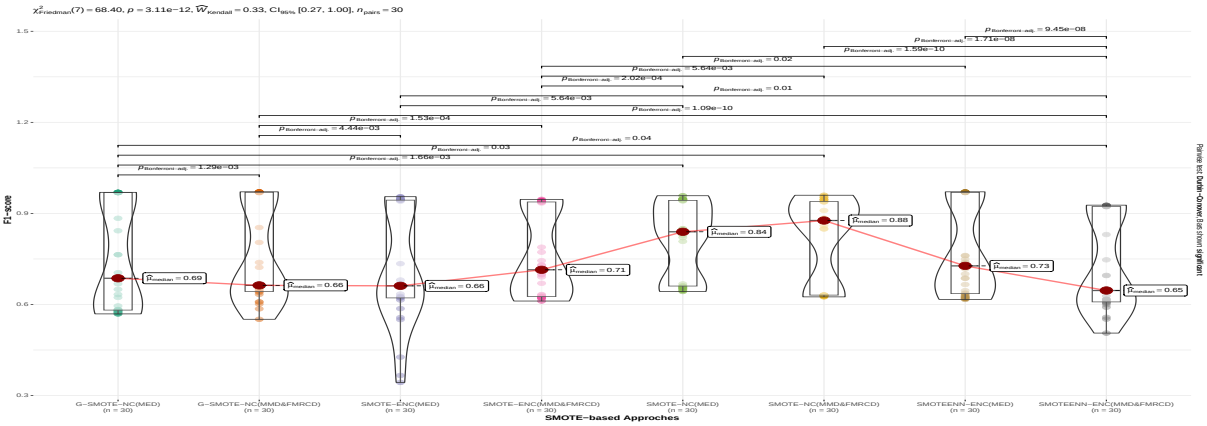

Figure 7. Friedman and Durbin-Conover tests based on F1-score metric achieved by the RF-RFE-based algorithm leveraging datasets pre-processed by the modifications of SMOTE-based approaches generally tailored for heterogeneous features coupled with either the MED or the MMD computed intrinsically to the FMRCD estimator.

employing the Friedman test reported a *p-value* of $3.11 \times 10^{-12} < 0.05$ ($\chi_F^2 = 68.40 > \chi_7^2(\alpha = 0.05) = 2.1670$), indicating that the SMOTE-based approaches are statistically significant different. Supplementary analysis using the Durbin-Conover test highlighted that the statistically optimal approach is statistically insignificant to SMOTE-ENC combined with the MED approach and SMOTEENN-ENC integrated with the MMD computed intrinsically to the FMRCD estimator as well as the former

4.3. Evaluating the Computational Stability of the Proposed Framework

Table 12 presents the coefficient of variations (based on aggregated evaluation metrics) exhibited by the predictive approach in conjunction with the adaptations of SMOTE-based approaches, either combined with the MED approach or with the MMD computed intrinsically to the FMRCD estimator. The least coefficient of variations

Table 12. Mean±SD of the coefficient of variation exhibited by the LR-AE-NET, and RF-RFE-based algorithm using datasets pre-processed by the versions of SMOTE-based approaches mainly configured for heterogeneous features blended with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Algorithm	Estimators	G-SMOTE-NC	SMOTE-ENC	SMOTE-NC	SMOTEENN-ENC
LR-AE-NET	MED	12.4965	12.7367	10.9331	12.2359
	MMD-FMRCD	12.8194	11.2308	**10.5299**	11.5236
RF-RFE	MED	22.4258	17.0680	11.5628	21.5334
	MMD-FMRCD	23.9102	10.9768	**9.6163**	10.6310

were evident from the interplay of LR-AE-NET and RF-RFE-based algorithm utilizing datasets pre-processed by SMOTE-NC combined with the MMD computed intrinsically to the FMRCD estimator. These findings suggest that our novelty offers superior stability to the predictive approach's performance.

4.4. Evaluation the Computational Efficiency of the Suggested Architecture

The computational efficiency of the conventional approaches and the proposed architecture was assessed across the three datasets, with averages runtimes measured in minutes. Table 13 reports the averages and SDs of the time taken by the adaptations of the SMOTE-based approaches either combined with the MED approach or with MMD computed intrinsically by the FMRCD estimator to pre-process the datasets.

Table 13. Mean±SD of the computational time (minutes) attained the SMOTE-based approaches predominantly formulated for heterogeneous features synthesized with either the MED or the MMD computed intrinsically to the FMRCD estimator.

Datasets	Estimators	SMOTE-NC	G-SMOTE-NC	SMOTE-ENC	SMOTEENN-ENC
P2P	MED	3.8790±0.0489	6.8950±0.0199	10.5920±0.0206	12.7073±0.0180
	MMD-FMRCD	20.4540±0.3800	31.3895±0.7693	34.9865±0.5743	49.1989±0.5467
IEEE-CIS	MED	7.9024±0.0650	18.0119±0.4948	25.1942±0.5230	19.2565±0.1575
	MMD-FMRCD	11.2911±0.5500	27.8630±0.5144	36.6375±0.6389	35.5567±0.9061
LDCD	MED	5.6565±0.0426	10.9237±0.0848	29.5140±0.8549	14.4246±0.1475
	MMD-FMRCD	25.5901±8.4467	35.4256±5.9163	38.9447±6.0074	58.0692±15.1509

The modifications of the SMOTE-based approaches typically customized for heterogeneous features integrated with the MED approach achieved the minimal computational time compared to the proposed framework. Despite the computational efficacy demonstrated by the suggested framework, it is computationally intensive, particularly in handling high-dimensional feature space.

5. Concluding Remarks

Big datasets in the credit risk landscape are often characterized by zero-inflated instances coupled with high-dimensional heterogeneous features some of which may be redundant. Consequently, these data aberrations adversely affect the conventional predictive classifiers, thereby exhibiting inaccurate and unreliable predictions. Additionally, these aberrations are exacerbated by the heavy-tailed error distributed data points and collinearity often inherent in the high-dimensional data. Therefore, in this study, the authors proposed robust distance-based SMOTE approaches typically designed for heterogeneous features by leveraging the MMD computed intrinsically

to the FMRCD estimator, which offer remedies to these multifaceted data aberrations. The proposed framework is tailored to effectively capture correlations and variability among features, which enhances the modelling strategy. Furthermore, this framework enhances robustification and efficacy of SMOTE-based approaches primarily customized for heterogeneous features, by combining them with the MMD computed intrinsically with the FMRCD estimator in conjunction with the feature selection procedures, LR-AE-NET and the RF-RFE-based algorithm. The empirical results exhibited by this framework demonstrate superior and stable predictive performance. These findings suggest that our novel approach significantly outperform traditional approaches in enhancing the predictive capabilities of the classifiers. Despite the tractability and stability demonstrated by the proposed framework, there is potential for enhancement in terms of its computational efficiency, especially in the convergence of the covariance (scatter) matrix, which is computationally demanding. This framework is suitable for adoption by practitioners and academia in enhancing credit risk management strategies that effectively circumvent data complexities, which are substantially posed by zero-inflated datasets with inherent high-dimensional heterogeneous features across a wide array of big data applications.

Nomenclature

Abbreviation	Full Term
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-NC	SMOTE for Nominal and Continuous Features
SMOTE-ENC	SMOTE for Encoding Nominal and Continuous Features
GSMOTE-NC	Geometric SMOTE-NC Features
SMOTEENN	SMOTE Edited Nearest Neighbors (ENNs) for Encoding Both Nominal and Continuous Features
ML	Machine Learning
LR	Logistic Regression
RF	Random Forest
MED	Modified Euclidean Distance
MMD	Modified Mahalanobis Distance
MMRCD	Modified Minimum Regularized Coefficient Determinant
ALASSO	Adaptive Least Absolute Shrinkage and Selection Operator
AE-NET	Adaptive Elastic Net
CV	Cross Validation
AUC	Area Under the Curve
GM	Geometric Mean
SDs	Standard Deviations
P2P	Peer to Peer
LDCD	Loan Default Credit Dataset
IEEE-CIS	IEEE-Computational Intelligence Society

Funding

The authors did not receive support from any organization for the submitted work.

REFERENCES

[1] X. Shi, F. Kong, and H. Li, *Research on credit evaluation model for high-dimensional imbalanced data*, 2021 IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA), IEEE, 2021, pp. 162–166.

[2] K.R. Musara, E. Ranganai, C. Chimedza, F. Matarise, and S. Munyira, *Robust hybrid data-level approach for handling skewed fat-tailed distributed datasets and diverse features in financial credit risk*, Sciendo, vol. 2025, pp. 230–270.

[3] Basel, II, *International convergence of capital measurement and capital standards: a revised framework*, vol. 107, 2004.

[4] A. Karami and C. Igbokwe, *The impact of big data characteristics on credit risk assessment*, Int. J. Data Sci. Anal, vol. 2025, pp. 1–21.

[5] C. Bulut and E. Arslan, *Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment*, Artificial Intelligence Review, vol. 57, no. 9, Springer, 2024, pp. 252.

[6] H. Guamán-Lloacana, A. Muzo-Bombón, C. Sánchez-Briceño, and J. Varela-Aldás, *A Literature Review on Enterprise Credit Assessment Using Random Forest*, 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM), IEEE, 2024, pp. 1–8.

[7] S. R. Lenka, S. K. Bisoy, and R. Priyadarshini, *Multiple optimized ensemble learning for high-dimensional imbalanced credit scoring datasets*, Knowledge and Information Systems, vol. 66, no. 9, Springer, 2024, pp. 5429–5457.

[8] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique*, J. Artif. Intell. Res, vol. 16, pp. 1–4, 2002.

[9] M. Mukherjee and M. Khushi, *SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features*, Appl. Syst. Innov, vol. 4, pp. 18, 2021.

[10] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, *RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification*, J. King Saud Univ.-Comput. Inf. Sci., vol. 34, pp. 5059–5074, 2022.

[11] J. Fonseca and F. Bacao, *Geometric SMOTE for imbalanced datasets with nominal and continuous features*, Expert Syst. Appl, vol. 234, 121053, 2023.

[12] G. Husain, D. Nasef, R. Jose, J. Mayer, M. Bekbolatova, T. Devine, and M. Toma, *SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models*, Algorithms, vol. 18, pp. 37, 2025.

[13] A. Arputharaj, S. Datta, and K.S. Hasan, Impact of distance measures on imbalanced classes for rule extraction, in *2019 6th Int. Conf. Soft Comput. Mach. Intell. (ISCMI)*, IEEE, 2019, pp. 29–34.

[14] H. Ghorbani, *Mahalanobis distance and its application for detecting multivariate outliers*, Facta Universitatis, Ser.: Math. Inform, vol. 2019, pp. 583–595.

[15] C. Leys, O. Klein, Y. Dominicy, and C. Ley, *Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance*, J. Exp. Soc. Psychol, vol. 74, pp. 150–156, 2018.

[16] S. Xiang, F. Nie, and C. Zhang, *Learning a Mahalanobis distance metric for data clustering and classification*, Pattern Recognit, vol. 41, pp. 3600–3612, 2008.

[17] D.L. Donoho, *The notion of breakdown point*, A Festschrift for Erich Lehmann/Wadsworth, 1983.

[18] P. Rousseeuw, *Multivariate estimation with high breakdown point*, Math. Stat. Appl. B, 1985.

[19] P.J. Rousseeuw and K.V. Driessen, *A fast algorithm for the minimum covariance determinant* estimator, Technometrics, vol. 41, pp. 212–223, 1999.

[20] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, 2003.

[21] K. Boudt, P. J. Rousseeuw, S. Vanduffel, and T. Verdonck, *The minimum regularized covariance determinant estimator*, Statistics and Computing, vol. 30, no. 1, pp. 113–128, 2020, Springer.

[22] S. Ma and Y. Duan, *An improved robust algorithm for the Fisher discriminant model with high-dimensional data*, PLoS One, vol. 20, no. 6, pp. e0322741, 2025, Public Library of Science San Francisco, CA, USA.

[23] D. Bertsimas and A. King, *Logistic regression: From art to science*, Statistical Science, JSTOR, 2017, pp. 367–384.

[24] D. Breskuvienė and G. Dzemyda, *Enhancing credit card fraud detection: highly imbalanced data case*, Journal of Big Data, vol. 11, no. 1, Springer, 2024, pp. 182.

[25] H. Guamán-Lloacana, A. Muzo-Bombón, C. Sánchez-Briceño, and J. Varela-Aldás, *A Literature Review on Enterprise Credit Assessment Using Random Forest*, 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM), IEEE, 2024, pp. 1–8.

[26] N. Li, H. Yang, and J. Yang, *Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data*, Communications in Statistics-Simulation and Computation, vol. 50, no. 12, Taylor & Francis, 2021, pp. 4263–4279.

[27] V. A. Phan, J. Jerabek, and L. Malina, *Comparison of Multiple Feature Selection Techniques for Machine Learning-Based Detection of IoT Attacks*, Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–10.

[28] S. Pirenne and G. Claeskens, *Parametric programming-based approximate selective inference for adaptive lasso, adaptive elastic net and group lasso*, Journal of Statistical Computation and Simulation, vol. 94, no. 11, Taylor & Francis, 2024, pp. 2412–2435.

[29] Z. Xie and X. Huang, *A credit card fraud detection method based on Mahalanobis distance hybrid sampling and random forest algorithm*, IEEE Access, 2024.

[30] J. Jung and Y.S. Choi, *SMOTE by Mahalanobis distance using MCD in imbalanced data*, Korean J. Appl. Stat, vol. 37, pp. 455–465, 2024.

[31] A. P. R. Sembada, M. Ahsan, and W. Wibawati, *Advanced process monitoring in ordinary portland cement (OPC) data using robust Max-Half-Mchart control charts with Fast-MCD and Dxcet-MCD estimators*, AIP Conference Proceedings, vol. 3301, no. 1, AIP Publishing LLC, 2025, pp. 050006.

[32] F. Fahri, M. Ahsan, and W. Wibawati, *Simultaneous robust Max-Half-Mchart control charts based on minimum regularized covariance determinant (MRCD)*, AIP Conference Proceedings, vol. 3301, no. 1, AIP Publishing LLC, 2025, pp. 050008.

[33] I. Sadok and M. Zribi, *Bayesian GLM: A non-informative approach for parameter estimation in exponential dispersion regression models*, Reliability: Theory & Applications, vol. 20, no. 1 (82), - Gnedenko Forum, 2025, pp. 715–727.

[34] X. Shi, F. Kong, and H. Li, *Research on credit evaluation model for high-dimensional imbalanced data*, 2021 IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA), IEEE, 2021, pp. 162–166.

[35] O. Ajibola-James and F. I. Okeke, *An approach for good modelling and forecasting of sea surface salinity in a coastal zone using machine learning LASSO regression models built with sparse satellite time series datasets*, Advances in Space Research, Elsevier, 2025.

[36] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Statist. Soc. Ser. B: Stat. Methodol, vol. 67, pp. 301–320, 2005.

[37] E. Mozafari-Majd and V. Koivunen, *The Adaptive τ-Lasso: Robustness and Oracle Properties*, IEEE Transactions on Signal Processing, IEEE, 2025.

[38] Y. Yang, J. Zou, X. Zhao, and P. Zhao, *The adaptive elastic net variable selection for linear mixed effects models based on the orthogonal projection*, Communications in Statistics - Theory and Methods, vol. 54, no. 17, Taylor & Francis, 2025, pp. 5381–5405.

[39] H. Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association, vol. 101, no. 476, pp. 1418–1429, 2006. Taylor & Francis.

[40] H. Zou and H. H. Zhang, *On the adaptive elastic-net with a diverging number of parameters*, Annals of Statistics, vol. 37, no. 4, 2009, pp. 1733.

[41] M. Nouraie, H. Zhu, and S. Muller, *A Stable Lasso*, arXiv preprint arXiv:2511.02306, 2025.

[42] H. Guamán-Lloacana, A. Muzo-Bombón, C. Sánchez-Briceño, and J. Varela-Aldás, *A Literature Review on Enterprise Credit Assessment Using Random Forest*, 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM), IEEE, 2024, pp. 1–8.

[43] Z. Mustaffa and M. H. Sulaiman, *Random forest based wind power prediction method for sustainable energy system*, Cleaner Energy Systems, Elsevier, 2025, pp. 100210.

[44] R. Masini and M. Medeiros, *Balancing Flexibility and Interpretability: A Conditional Linear Model Estimation via Random Forest*, arXiv preprint arXiv:2502.13438, 2025.

[45] H. Ghinaya, R. Herteno, M. R. Faisal, A. Farmadi, and F. Indriani, *Analysis of Important Features in Software Defect Prediction Using Synthetic Minority Oversampling Techniques (SMOTE), Recursive Feature Elimination (RFE), and Random Forest*, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 3, 2024, pp. 276–288.

[46] C. Dewi, R. A. Andika, E. Haryani, D. Riantama, A. Sajid, M. M. Alam, and M. M. Su'ud, *Feature Selection for Financial Data Classification Using Random Forest, Boruta, and Recursive Feature Elimination*, Ingenierie des Systemes d'Information, vol. 30, no. 8, 2025, pp. 2165.

[47] F. I. Sarky and E. Poerwandono, *Optimization of SMOTE Application for Classification Accuracy of Heart Disease Risk Using Artificial Neural Network*, Journal Innovations Computer Science, vol. 4, no. 2, 2025, pp. 103–110.

[48] M. Isangediok and K. Gajamannage, *Fraud detection using optimized machine learning tools under imbalance classes*, 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4275–4284.

[49] R. Belferik, F. M. Sinaga, F. Ferawaty, M. A. S. Manullang, and T. Sinaga, *Addressing class imbalance in stunting classification using SMOTE enhanced random forest*, Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 9, no. 4, 2025, pp. 2108–2116.

[50] V. Kumar, G.S. Lalotra, P. Sasikala, D.S. Rajput, R. Kaluri, K. Lakshmanna, M. Shorfuzzaman, A. Alsufyani, and M. Uddin, *Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques*, Healthcare, vol. 10, 1293, MDPI, 2022.

[51] L. Yu, L. Yu, and K. Yu, *A high-dimensionality-trait-driven learning paradigm for high dimensional credit classification*, Financ. Innov, vol. 7, pp. 1–20, 2021.

[52] E. Rendon, R. Alejo, C. Castorena, F.J. Isidro-Ortega, and E.E. Granda-Gutierrez, *Data sampling methods to deal with the big data multi-class imbalance problem*, Appl. Sci, vol. 10, no. 4, pp. 1276, MDPI, 2020.

[53] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, *Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending*, Inf. Sci, vol. 525, pp. 182–204, 2020.

[54] Fiddler Labs, *P2P Lending Data: Accepted Loans (2007 to 2018 Q3)*, Available at: https://github.com/fiddler-labs/p2p-lending-data/blob/master/raw$_data/accepted_2007_to_2018Q3.csv.gz$ (Accessed: 2025-10-17).

[55] T. Kreienkamp and A. Kateshov, *Credit risk modeling: Combining classification and regression algorithms to predict expected loss*, , vol. 8, pp. 4–10, 2014.

[56] Imperial College London, *Loan Default Prediction*, Kaggle, available at: `https://www.kaggle.com/c/loan-default-prediction` (Accessed: 2025-10-17).

[57] C. V. Sai, D. Das, N. Elmitwally, O. Elezaj, and M. B. Islam, *Explainable AI-driven financial transaction fraud detection using machine learning and deep neural networks*, Available at SSRN, 2023.

[58] IEEE-CIS Fraud Detection, *Kaggle*, Available at: https://www.kaggle.com/c/ieee-fraud-detection/data (Accessed: 2025-10-17).

[59] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, and M. Tang, *A method for analyzing the performance impact of imbalanced binary data on machine learning models*, Axioms, vol. 11, no. 11, pp. 607, 2022. Available at MDPI.

[60] B. Mirza, D. Haroon, B. Khan, A. Padhani, and T. Q. Syed, *Deep generative models to counter class imbalance: A model-metric mapping with proportion calibration methodology*, IEEE Access, vol. 9, pp. 55879–55897, 2021. Available at IEEE.

[61] R. Ghorbani, R. Ghousi, A. Makui, and A. Atashi, *A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset*, IEEE Access, vol. 8, pp. 141066–141079, 2020. Available at IEEE.

[62] T. H. K. Dong, L. S. Canas, J. Donovan, D. Beasley, N. T. Thuong-Thuong, N. H. Phu, N. T. Ha, S. Ourselin, R. Razavi, G. E. Thwaites, and others, *Convolutional neural network using magnetic resonance brain imaging to predict outcome from tuberculosis meningitis*, PloS One, vol. 20, no. 5, pp. e0321655, 2025. Available at Public Library of Science.