

https://dspace.org/

Robust Hybrid DataLevel Approach for Handling Skewed FatTailed Distributed Datasets and Diverse Features in Financial Credit Risk

Keith R Musara; Ranganai Edmore; Charles Chimedza; Florence Matarise; Sheunesu Munyira

https://hdl.handle.net/10539/46308
WOS

Downloaded from a DSpace repository. https://dspace.org/

FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 50 (2025) No. 2
ISSN 0867-6356
sciendo DOI: 10.2478/fcds-2025-0009 e-ISSN 2300-3405

Robust Hybrid Data-Level Approach for Handling Skewed Fat-Tailed Distributed Datasets and Diverse Features in Financial Credit Risk

Musara Keith R, *, Edmore Ranganai †, Charles Chimedza ‡,
Florence Matarise §, Sheunesu Munyira ¶

Abstract. Skewed fat-tailed distributed (imbalance or class-imbalance) datasets pose over- whelming aberrations in numerous machine learning (ML) algorithms, particularly in real-life ap- plications, especially in the domain of credit risk modelling, where default cases (minority-classes) are often outnumbered by non-default cases (majority-classes) cases or vice versa. Data-level (DL) approaches have been suggested in the recent literature as remedies for skewed fat-tailed distributed datasets. The popularized DL approach in contemporary studies is the synthetic minority oversampling technique (SMOTE) and its variants that are capable of mitigating the risk of overfitting and minimizing the generalization errors. However, these approaches can introduce noisy instances that adversely diminish the robustness of the ML algorithms. Also, they are often amenable to the presence of nominal features with mismatching labels that are inherent in real-world datasets. To bridge these gaps, we proposed a hybrid innovation framework that effectively mitigates the aberrations presented by nominal features with mismatching labels and noisy instances simultaneously. The proposed approach is the SMOTE-edited nearest neighbors-encoding nominal and continuous (SMOTEENN-ENC) features. The efficacy of our novelty was evaluated against DL approaches suggested in the literature, orchestrated to handle skewed fat-tailed distributed datasets with inherent diverse features. This approach was coupled with widely employed ensemble algorithms, namely the random forest (RF) and the extreme gradient boost (XGBoost). The results suggested that our novelty, SMOTEENN-ENC, integrated with the XGBoost algorithm demonstrated superiority and stability in the predictive performance when applied to skewed fat-tailed distributed datasets with inherent diverse features.

*Faculty of Science, University of Zimbabwe, Zimbabwe, {musarakeith}@gmail.com
†Faculty of Science, University of South Africa, South Africa, {rangae}@unisa.ac.za
‡School of Statistics, University of the Witwatersrand, South Africa, {charles.chimedza}@wits.ac.z
§Faculty of Science, University of Zimbabwe, Zimbabwe, {matarise7}@gmail.com
¶Faculty of Science, University of Zimbabwe, Zimbabwe,{munyirask}@gmail.com

Keywords: Skewed fat-tailed distributed datasets, Diverse features, Noisy instances.

1. Introduction

Motivation of the Study
The classification tasks in financial credit risk have gained significant attention in scientific research [70], especially in the context of globalization, where financial institutions are expected to navigate complex and interconnected markets to mitigate potential losses from borrower defaults [17]. Financial credit risk assessment is obligatory for financial lending institutions to comply with international financial reporting standard (IFRS) 9 [86], which mandates provisions based on expected credit losses (ECL) [68]. Therefore, credit risk assessment using machine learning (ML) algorithms has become the cornerstone for financial lending institutions so as to minimize risks. However, real-world datasets are characterized with skewed fat-tailed distributed (class-imbalance or imbalanced) datasets ([54],[38]), with inherent diverse features ([20], [84]) that adversely affect the predictive performance of ML algorithms. The former phenomenon occurs when at least one class is insufficiently represented, such as defaults (minority class), while the other class is overrepresented, such as non-defaults (majority class) [67], and the latter setback consists of a mixture of nominal and continuous features [42]. The efficiency and predictive power of the predominantly leveraged ML algorithms are susceptible to the afore-mentioned undesirable data features [20]. They often tend to incline towards the majority class (non-defaults), which leads to suboptimal predictive performance [4]. Consequently, it becomes futile for algorithms to discriminate between the defaults and non-defaults, as this results in erroneous predictions, and ultimately leads to significant financial losses incurred by any financial lending institution ([25]). Therefore, it is fundamental to circumvent the aberrations presented by real-world datasets simultaneously, allowing algorithms to attain superior predictive performance with desirable and reliable predictions.

To overcome the afore-mentioned limitations of ML algorithms, three main approaches have been suggested in the literature: data-level (DL) techniques (external), algorithm-level (AL) techniques (internal), and cost sensitive (CS) techniques (external) [23]. The DL techniques are orchestrated to alter (balancing) the training data to ensure its suitability for the ML algorithms [99], whereas the AL techniques hinge on modification of the existing algorithm to alleviate the bias inclined towards the majority class ([119], [103]). Moreover, CS techniques introduce a cost matrix that imposes greater expense penalties to the algorithm so as to avoid skewed predictive performance [59]. Contemporary studies have shown a stronger inclination towards the DL techniques due to their versatility and easy implementation to the ML algorithms [95]. These techniques are further classified into three categories: over-sampling approaches, under-sampling approaches and hybrid approaches [111]. Over-sampling approaches involve replicating the instances of the minority class [20], while under-sampling approaches minimize the majority class [109], and the hybrid approaches

combine the strengths of the afore-mentioned approaches while repelling their flaws. [58].

Generalization of Synthetic Minority Oversampling Technique (SMOTE) Mechanisms
The most extensively used DL approach in the literature is the synthetic minority oversampling technique (SMOTE) [33], which has led to the introduction of a plethora of variations (mechanisms) [32]. SMOTE-based approaches are robust due to their capability of generating synthetic instances interspersed among existing minority class instances by using a linear interpolation method, thereby preserving data integrity ([133], [66]). Additionally, they are designed to significantly reduce overfitting [20], a common issue in conventional approaches, such as the random oversampling [85]. SMOTE is also well-known to minimize generalization errors [84]. However, this approach often introduces irrelevant or uninformative samples, and increases class overlap, a primary source of noise, which may diminish the predictive performance of ML algorithms [3]. To address the shortcomings of SMOTE, He et al. [53], suggested the adaptative synthetic (ADASYN) procedure, an over-sampling technique that creates more synthetic instances for classes that are harder to learn based on their relative difficulty. Nonetheless, this approach tends to introduce noise into the dataset, that can adversely hinder predictive performance of the ML algorithms [49]. This prompted Han et al. [49], to further propose the borderline-SMOTE (B-SMOTE), which primarily focuses on difficult instances at class boundaries. However, this approach is inefficient when classes are highly overlapping [1].

Despite the predominant use of over-sampling approaches, they fail to effectively discard the distributional overlap. Alternative methods have been suggested in the literature such as the under-sampling approaches, which aim to address redundant noisy data points and improve the separation between classes [110]. The most widely used under-sampling approaches include: the edited nearest neighbor (ENN) and Tomek-link, as suggested by Wilson [121] and Tomek [115], respectively. ENN systematically removes majority instances that are incorrectly categorized [121], while the Tomek link discards instances from overlapping regions between classes [115]. Nevertheless, under-sampling approaches have a limitation of discarding a substantial number of majority class samples [92], which can lead to inadequately trained algorithms due to small data points (instances) [130]. Currently, two hybrid approaches have been extensively used in the literature. These include integration of SMOTE and ENN to form SMOTEENN [10], and a combination of SMOTE and Tomek to form SMOTE-Tomek [10]. These hybrid approaches have proven to be robust and computationally efficient in handling skewed fat-tailed distributed datasets due to their ability to generate synthetic instances (avoiding overfitting) [20], eliminate noisy instances and remove potentially ambiguous instances near class boundaries [72]. However, in handling skewed fat-tailed distributed datasets SMOTEENN has consistently demonstrated superior predictive performance compared to SMOTE-Tomek convincingly, which significantly enhances the predictive capabilities of ML algorithms [107]. Despite the notable improvements offered by SMOTE-based approaches in predicting skewed fat-tailed distributed datasets, their efficacy remains limited due to their in-

ability to adequately provide a better remedy for diverse features [38]. The limitations of these SMOTE mechanisms arise from their reliance on the Euclidean distance (ED) metric, which is limited for continuous features exclusively [84].

Generalization of SMOTE Mechanism for Diverse Features
To counter the shortcomings of SMOTE-based approaches afore-mentioned, Chawla et al. [20], suggested the SMOTE for nominal and continuous (SMOTE-NC) features. This approach derives its proficiency from adopting a modified ED metric used for computing distances across diverse features based on k-nearest neighbors [38]. However, SMOTE-NC often falls short in effectively capturing mismatched labels of nominal features [84], as it assumes uniformity in these labels. Consequently, this limitation can compromise the predictive performance of the ML algorithms when applied to mismatching labels of nominal features. To tackle the drawbacks of SMOTE-NC, Mukherjee and Khushi [84], proposed the SMOTE for encoding nominal continuous (SMOTE-ENC) features, a generalization of SMOTE approach. This approach offers a better computational performance compared to the SMOTE-NC in combating the mismatched labels of nominal features due to its sophisticated encoding process based on associations between nominal features using the chi-principle [84]. However, SMOTE-ENC shares a similar limitation with SMOTE, potentially introducing noisy instances [94], which leads to poor predictive performance of the ML algorithms [3]. To mitigate such undesirable features, Fonseca and Bacao [38], presented a geometric SMOTE-NC (G-SMOTE-NC) which generates synthetic instances in a geometric fashion called hypersperiod instead of linear interpolation. The efficiency of this approach lies in eliminating the distributional overlap [38]. Additionally, the approach is successful in handling skewed fat-tailed distributed datasets with inherent diverse features against SMOTE-NC [38]. Nevertheless, it overlooks the principle of managing mismatched labels of nominal features, which are inherent in real-world datasets. Notwithstanding the significant improvement of SMOTE-based mechanisms for handling skewed fat-tailed datasets coupled with diverse features, a critical gap remains in resolving noisy instances and mismatched labels of nominal features simultaneously. In this paper, we proposed a generalization of the hybrid framework to effectively handle skewed fat-tailed distributed datasets with inherent diverse features, particularly nominal features with mismatching labels and avoiding noisy instances by permitting the, SMOTEENN to encode the nominal and continuous (SMOTEENN-ENC) features .

Motivation of the Selected Machine Learning (ML) Algorithms
The field of predictive analytics domain has evolved rapidly from conventional approaches to ML algorithms, which have demonstrated a superior computational efficiency and predictive performance ([125], [21], [118]). ML algorithms, have been extensively used for their robust predictive power [116], as well as for automating the learning process [15]. Moreover, they exhibit flexibility by leveraging massive amounts of data [91], enabling them to adapt and provide accurate predictions across a wide spectrum of real-life applications [101]. However, these algorithms are susceptible to aberrations posed by skewed fat-tailed distributed datasets [79], primarily because

they are typically designed to handle symmetric binary datasets [101]. Consequently, the predictions of ML algorithms are inclined towards the majority class [60]. Hence, most misclassifications cost in the domain of skewed fat-tailed distributed datasets emanate from the minority class rather than the majority class [114]. These complications are further exacerbated when the skewed fat-tailed distributed datasets are coupled with diverse features. Therefore, corrective measures, such as DL approaches have been suggested in the literature as well as our novelty proposition. Our approach is capable of effectively handling the data phenomena in this study, which ultimately enhances the efficacy of the ML algorithms.

ML algorithms are segmented into dual categories namely; single algorithms and ensemble algorithms (aggregate tree-based algorithms). The latter algorithms have proven to be successful in predictive performance of skewed fat-tailed distributed datasets ([63], [40], [13]). Hence, in this study, we further delved into exploring the robustness of these algorithms in resolving the overwhelming aberrations presented by skewed fat-tailed distributed datasets with inherent diverse features. The extensively used ensemble algorithms in the literature for skewed fat-tailed distributed datasets include, random forest (RF) ([63], [83], [13]) and extreme gradient boost (XGBoost) ([130], [16], [12]). Breiman [14] proposed the RF algorithm, which is a generalization of the decision trees. It is frequently used due to its capabilities in capturing complex non-linear features that are inherent in real-world applications [49], as well as its effectiveness in managing outliers and noisy instances [14]. However, the algorithm tends to be susceptible to aberrations of skewed fat-tailed distributed datasets [44]. The XGBoost algorithm, proposed by Chen and Guestrin [21] is recognized for its effectiveness in combining the gradient boosting optimization strategy with decision trees classifiers [134]. Thus, the algorithm is well-known for its superior predictive capabilities, and excellent computing speed [90]. Albeit the merits of the XGBoost, the algorithm is amenable when presented to skewed fat-tailed distributed datasets. The algorithm tends to incline the prediction towards the majority class leading to inefficient prediction, particularly for the minority class [69]. Therefore, in this research paper, we aim to explore the efficacy of integrating our novelty, SMOTEENN-ENC with robust ensemble algorithms. This becomes a critical step in effectively handling skewed fat-tailed distributed datasets with inherent diverse features.

Research Gap and Contribution of the Study

Our contribution in this paper aims to address a crucial gap in handling skewed fat-tailed distributed datasets with inherent diverse features, particularly in the presence of noisy instances and mismatching labels of nominal features. To achieve this, we investigated the computational efficiency of our novelty SMOTEENN-ENC against DL approaches tailored for skewed fat-tailed distributed datasets with inherent diverse features by leveraging robust ensemble algorithms. The efficacy of the suggested framework was assessed using real-world datasets that are characterized by skewed fat-tailed distributed datasets with inherent diverse features. The real-world datasets were classifed into three categories namely; mildly ($20\% \leq$ minority class ≤ 35), moderately ($1\% \leq$ minority class < 20) and extremely (minority class $< 1\%$) [67]. The

contributions of this paper are premised on the following points:

- to the best of the author's knowledge, there is scarcity of research studies that has comprehensively explored the three categories of skewed fat-tailed distributed datasets; mildly, moderately and extremely in assessing the efficacy of DL approaches suitable for skewed fat-tailed distributed datasets with inherent diverse features.
- there is a void of studies that have suggested a hybrid DL approach for handling skewed fat-tailed distributed datasets with inherent diverse features. Therefore, in this paper, we introduced SMOTEENN-ENC, which mitigates overfitting through SMOTE, removes noisy instances with the use of ENN, and effectively manages both continuous and nominal features including those with mismatched labels by employing an encoding principle.
- we adopted the most widely used ensemble algorithms in credit assessment, namely RF and XGB coupled with SMOTEENN-ENC, a new hybrid DL approach for skewed fat-tailed distributed datasets with inherent diverse features.
- this paper serves as a guide to data analytics practitioners and academia in selecting the most suitable strategy for skewed fat-tailed distributed datasets with inherent diverse features based on the ensemble algorithms.

The paper is organized as follows: Section 2 provides the theoretical properties of DL approaches tailored for this study and our novelty contribution. Furthermore, we included a comprehensive review of commonly employed ensemble algorithms. In Section 3, we described data features, the evaluation metric, and a review of Friedman and Durbin Conover tests used for comparative analysis. Section 4 presents the results of our novelty, SMOTEENN-ENC, against DL approaches that are in existence in the literature for handling skewed fat-tailed distributed datasets with inherent diverse features blended with RF algorithm and XGBoost algorithm using the real-world datasets. Finally, Section 5 explains the results and implications of the research.

2. Theoretical Framework

In this section, we presented the theoretical foundations for the DL approaches tailored to handle skewed fat-tailed distributed datasets with inherent diverse features and the author's proposed hybridization framework. We further provided a detailed theoretical review for the widely referenced ensemble classification approaches which are capable of capturing non-linear, complex diverse features that are inherent in real-world datasets.

2.1. Theory of SMOTE Mechanisms for diverse features

SMOTE has emerged as a de facto standard and a robust DL approach for effectively enhancing the predictive performance of ensemble algorithms when applied to

skewed fat-tailed distributed datasets. It achieves this by balancing the datasets and mitigating the risk of overfitting [20]. The SMOTE approach is outlined as follows:

- Initially, the ED metric is computed for the k nearest neighbors of the feature samples. The distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$ is estimated by the formula given:

$$\mathbf{S}_{k\mathrm{NN}} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad [1]$$

  where, $\mathbf{S}_{k\mathrm{NN}}$ is considered feature samples k nearest neighbor.

- The linear interpolation mechanism used for generating artificial instances based on the formula:

$$\mathbf{S}_{\mathrm{syn}} = \mathrm{r}(\mathbf{S}_{k\mathrm{NN}} - \mathbf{S}_{\boldsymbol{f}}) + \mathbf{S}_{\boldsymbol{f}} \quad [2]$$

  where, $\mathbf{S}_{\mathrm{syn}}$ denotes the generated synthetic samples, r is a random number between 0 and 1, and $\mathbf{S}_{\mathrm{f}}$ is the feature sample.

- Finally, we combine the original minority class instances with newly generated synthetic instances to create the dataset.

Notwithstanding the substantial effectiveness of SMOTE, this approach flops in handing skewed fat-tailed distributed datasets with inherent diverse features effectively [71]. To address this limitation, SMOTE-NC has been suggested to improve the efficiency of the SMOTE mechanism by adopting modified ED metric that effectively manage the diverse features [54], that are inherent in real-world datasets. The SMOTE-NC approach is presented in detailed below:

- Calculate the median of the standard deviations of all continuous features for the minority class. If the nominal features differ between the sample and its potential nearest neighbor, this median is included in the modified ED metric calculation. The median serves to exclude the differences in nominal features by an amount associated with specific differences in continuous features.

- Utilize the continuous feature space to calculate the modified ED metric between a feature vector (minority class sample) that identifies the nearest neighbor and another feature vector (minority class sample). Each nominal feature is evaluated for deviation between the feature vector considered and its nearest potential, incorporating the median of the previously calculated standard deviation into the modified ED metric calculation.

- Synthetic instances for continuous features are created using the classical SMOTE approach. For the nominal feature, the synthetic instances are obtained from the mode of the k nearest neighbors. Finally the instances are integrated to form a complete dataset.

Despite the advancement of SMOTE-NC, the approach has received criticism in the literature. He and Garcia [53] emphasized its inability to effectively capture mismatched labels of nominal features. This approach assumes uniformity in labels of

nominal features [84]. To address this shortcoming, Mukherjee and Khushi [84] proposed the SMOTE-ENC, which effectively manages the mismatched labels of nominal features. The SMOTE-ENC approach is outlined as follows:

- Calculate the imbalance ratio (ir) of the training set, which is the ratio of the number of instances of the minority class to the total number of instances in the training set.
- For each nominal feature (l), compute the expected number of labeled minority class instances in the training set (e') using the formula: $e' = e \times ir$, where e is the total number of l labeled instances in the training set.
- Compute the chi (χ) for each nominal feature: $\chi = \frac{o-e'}{e'}$, where o is the number of l labeled minority class instances in the training set.
- If the number of continuous (c) features in the training set is greater than zero for each label l in distinct labels we multiply the χ with the median of the standard deviation of continuous features. Otherwise, the number of l labeled minority instances in training equals χ.
- Create synthetic data for continuous features using SMOTE. For nominal features, the synthetic data is represented by the mode of the k nearest neighbors.
- The final phase of the SMOTE-ENC algorithm involves reversing the encoding of nominal values to their original labels.

The afore-mentioned SMOTE-based approaches often amplify the generation of noise instances because the approaches consider the distribution of minority class samples while neglecting the majority class samples in the neighborhood when generating artificial samples [94]. This oversight, can lead to increased class distribution overlap, resulting in generation of noisy instances. To alleviate this challenge, Fonseca and Bacao [38] introduced G-SMOTE-NC, which eliminates the noisy instances by the use of a geometric region that truncates the hyper-spheroid in generating sythetic instances. The G-SMOTE-NC is presented in detail as follows:

- Compute the expected number of samples to be generated, difference between the number of samples in the majority and minority classes.
- Initially set the generated sample to zero, then generate N samples.
- Randomly select one instance from the minority class and its nearest neighbor.
- Generate a new synthetic instance in a geometric fashion known as hyper-spheroid rather than a linear approach.
- Combine the newly generated synthetic sample to the initial set of samples generated.

The G-SMOTE-NC demonstrates its effectiveness of discarding noisy instances over the SMOTE-NC [38]. However, this approach substantially fails to eliminate the detrimental effects of noisy instances and falls short in capturing the mismatching labels of nominal features. Therefore, in this paper, we present the first hybridization approach, SMOTEENN-ENC, which robustly handles the aberrations posed by skewed fat-tailed distributed datasets with inherent diverse features. The procedure is outlined as follows:

- The initial procedure of this algorithm involves the encoding process, for each label of a nominal feature as a numeric value. A higher value of a label indicates a stronger association with the minority class, based on Pearson's chi-squared principle.
- After the encoding process, the SMOTE approach is initiated to generate artificial samples while maintaining data intergrity.
- Re-calculate the k-nearest neighbors of the instances with others using the artificial samples generated, returning the majority class from the k-nearest neighbors.
- If the neighboring instances have the same class as the current sample, they will be returned; otherwise, they will be eliminated.
- The final phase of the SMOTEENN-ENC algorithm involves reversing the encoding of nominal values to their original labels.

2.2. Ensemble Classification Algorithms

Ensemble classification algorithms are extensively used for modelling skewed fat-tailed distributed datasets due to their robustness and computational efficiency [63]. Two popular classes of ensemble algorithms are bagging (Bootstrap Aggregation) [14], and boosting [100]. The bootstrapping process involves training an algorithm in parallel to each bootstrap sample generated from the training set. The predictive outcomes from each algorithm are systematically analyzed, and a single decision is made using the majority voting [40]. On the other hand, boosting involves training algorithms sequentially, with each succeeding algorithm being trained on a reweighted data subset that was misclassified in the preceding phase [34]. The most widely used conventional bagging algorithm is the RF algorithm, while the predominantly employed boosting algorithm is XGBoost, which is an extension of the gradient boost [21].

2.2.1. Random Forest (RF) Algorithm

The RF algorithm is well-known for providing a robust basis for modeling the skewed fat-tailed distributed datasets because of its robustness and versatility, excellent processing speed [98], and effective handling of multicollinearity [11]. Moreover, during

the tree-building process, some data instances may be utilized multiple times or not at all, which helps to minimize correlation among the trees [82]. This algorithm is a non-parametric technique used for classification purposes, offering several advantages, including the capability to reduce overfitting, and ease of implementation [47]. Figure 1 presents the architecture for the RF algorithm of this study adopted from Brienman [14].

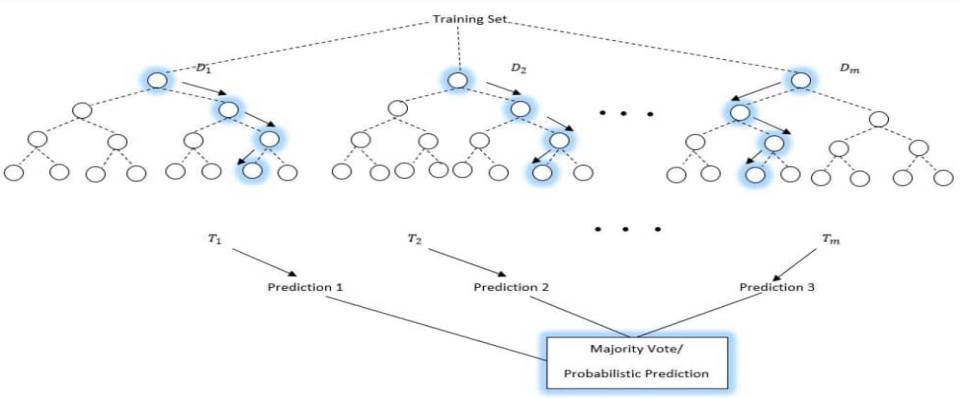


Figure 1. Architecture representation of an RF algorithm.

The RF algorithm initially generates boostrapped datasets $D_1, D_2, ..., D_M$ from the training set D. The trees are constructed without pruning based on the bootstrap datasets and from those M trees are generated such as $T_1, T_2, ..., T_M$. This enables the extraction of M predicted trees $T_1(z), T_2(z), ..., T_M(z)$. Finally, the prediction for all the M trees is determined as: $T(z)$ = majority vote $\{T_i(z)\}_{i=1}^{M}$. The algorithm has gained popularity in various application scenarios, particularly involving skewed fat-tailed distributed datasets [87]. Moreover, the RF algorithm is efficient on large datasets and even for missing data points [132], and is designed to minimize the classification error rate [26]. It has been extensively utilized for its accuracy and computational feasibility in a wide spectrum of scenarios [11]. However, Gu (2007) [44] emphasized that the algorithm is susceptible to skewed fat-tailed distributed datasets. To overcome this data aberration, applications of DL approaches can be crucial for enhancing the efficiency of the RF algorithm when dealing with the skewed fat-tailed distributed datasets with inherent diverse features.

2.2.2. Extreme Gradient Boost (XGBoost) Algorithm

The XGBoost algorithm is an advanced supervised algorithm proposed by Chen and Guestrin [21] within the boosting framework. It has been widely recognized due to its merits of high flexibility [89], strong predictability [5], strong generalization ability [31], high scalability [21], high model training efficiency, and great robustness [78]. The XGBoost algorithm can particularly involve base learner; objective function, ad-

dictive training and optimal tree structure selection.

Step 1: Base Learner
The classification and regression tree (CART) as the base learner for the boosting approach [14]. CART is known for its unique structure that allows it to capture diverse features effectively. Moreover, the approach is based on a decision tree which also offers better interpretability and transparency of the decisions constructed by CART. The base learner demonstrates its robust capabilities in handling sparse features, enhancing the computational efficiency compared to conventional tree-based algorithms [122]. CART is also versatile [61], and enhances training speed, particularly in bulk power grids with large data streams. This capability facilitates real-time monitoring and analysis, helping to counter significant threats.

Step 2: Objective Function
Consider a dataset $D=\{(x, y)\}$ with a samples and m features, the base learner allows the classification decision to occur. The objective function (θ) of the XGB is defined as:

$$\theta = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{t} \Omega(f_k), \quad f_k \in \mathcal{F} \qquad [3]$$

where, $l(\hat{y}_i, y_i)$ is the loss function representing the difference between the predicted label $\hat{y}_i$ and target y_i. $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ represents the complexity of each CART, where T is the number of leafs in a CART so that the k^{th} base learner f_k denotes a tree structure with leaves value of ω. The coefficients, γ and λ are the penalization terms that are obtained automatically. $\mathcal{F} = \{f(x) = \omega_{q(x)}\}$ is called the CART space, where $\omega \in \mathbf{R}^T$ and $q : \mathbf{R}^m \rightarrow T$ are the tree structures that map a sample to a leaf node in a tree. The effectiveness of XGBoost algorithm stems from its capability to include this penalization term, which reduces the algorithm's complexity by mitigating the risk of overfitting, while optimizing its computational efficiency.

Step 3: Additive Training
The XGBoost algorithm is trained in an additive manner, meaning optimization occurs at each iteration level rather than at the end of the training process. Let $\hat{y}_i^{(t)} = 0$, the addictive training process is expressed as:

$$\hat{y}_i^{(t)} = \sum_{i=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

where $f_t(x_i)$ denotes the predicted results of the i^{th} sample at the t^{th} iteration. To find the optimized tree structure that minimizes the objective function, we considered the square loss function $l(\hat{y}_i, y_i)$ and utilized the Taylor expansion of θ, which is defined as:

$$\begin{aligned} \theta^t &= \sum_{i=1}^{n} l(\hat{y}_i^{(t-1)} + f_t(x_i), y_i) + \Omega(f_t) + C \\ &= \sum_{i=1}^{n} \left[l(\hat{y}_i^{(t-1)}, y_i + g_i f(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + C \end{aligned} \tag{5}$$

where $g_i = 2(\hat{y}^{(t-1)} - y_i)$, $h_i = 2$ and C is a constant that can be be removed to obtain the simplified objective function t^{th} iteration below:

$$\theta^t = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i \right] + \Omega(f_i) \tag{6}$$

Let the sample set of leaf j be $I_j = \{i | q(x_i) = j\}$. Regroup the objective function and further reduce it to:

$$\begin{aligned} \theta^t &= \sum_{i=1}^{n} \left[g\omega_q(x_i) + \frac{1}{2} h_i \omega_{q(x_i)} \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T \end{aligned} \tag{7}$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. Supposed we have a fixed tree structure $q(x)$, the optimal weight in each leaf and resulting objective function could be calculated by substituting $\omega^* = -\frac{G_j}{H_j + \lambda}$ such that the simplified objective function is defined as:

$$\theta = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{8}$$

Step 4: Optimal Tree Structure Selection
Tree-based algorithms consist of an infinite number of structures. To determine the optimal decision tree, the score function is utilized. The gain score is expressed as:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L^2 + G_R^2)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{9}$$

where $\frac{G_L^2}{H_L+\lambda}$ and $\frac{G_R^2}{H_R+\lambda}$ the score of the left and right side after splitting a node on a tree respectively. This approach provides robust predictive accuracy through iterative boosting and regularization techniques, which avoids overfitting by minimizing variance whilst improving accuracy [105]. However, the algorithm's predictive performance can become subtle when dealing with skewed fat-tailed distributed datasets [118]. This necessitates the implementation of DL approaches as an effective strategy to enhance the robustness of the XGBoost algorithm.

3. Methodological Procedures

In this section, we discuss the features of real-world datasets characterized with skewed fat-tailed distribution with inherent diverse features, as well as widely referenced evaluation metric measures for evaluating the proposed framework. Additionally, the evaluation metrics for the hybridization DL approaches coupled with robust ensemble algorithms are scrutinized through statistical methods, including the Friedman and Durbin-Conover tests. This study aims to ascertain the efficacy of the suggested strategy in handling skewed fat-tailed distributed datasets with inherent diverse features.

Data Features

In this paper we investigated the three scenarios of skewed fat-tailed distributed datasets based on the proportion of the minority class; mildly ($20\% \leq$ minority class ≤ 35), moderately ($1\% \leq$ minority class < 20) and extremely (minority class $< 1\%$) as suggested by Kumar et al. [67]. Table 1 a summary of the data features, including the type of datasets, the total number of instances, minority proportion, the number of nominal features and continuous features as well as the total number of features. The features consist of features of a customer such as credit history, personal information, and details of the credit requested.

Table 1. Summary of Datasets

Dataset	Instances	Features		Minority Class	Class-Imbalance
		Nominal	Continuous		
ACD	511	9	6	25.05%	
DCCC	30,000	17	3	30.01%	Mild
GCCD	1,000	3	20	22.12%	
CCCP	5, 960	2	10	19.95%	
GMSD	8, 899	5	5	6.00%	Moderate
HMEQ	10, 127	9	10	16.07%	
CCFD	1, 048, 575	3	5	0.50%	
FT	1, 048, 575	2	5	0.10%	Extreme
XFDC	95,662	4	10	0.20%	

The mildly skewed fat-tailed distributed datasets consisted of the Australia credit dataset (ACD) (http://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+ap\proval)), and default of credit card clients (DCCC) (https://archive.ics.uci.edu/dataset/350/default+of+credit+card+client) and Germany credit card data (GCCD) (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data).
For the moderately skewed fat-tailed distributed datasets consisted of three namely; credit card card churn prediction (CCCP) (https://www.kaggle.com/code/msafi04/creditcard-churn-\prediction-extensiveeda/input), get me some data (GMSD) (https://www.kaggle.com/c/GiveMeSomeCredit/data) and home equity (HMEQ) (https://www.kaggle.com/datasets/ajay1735/hmeq-data/data). Lastly, the extremely skewed fat-tailed distributed datasets included; credit card fraud detection (CCFD) (https://www.kaggle.com/code/krippanandhini/credit-card-fraud-detection/input), fraud transactions (FT) https://www.kaggle.com/data//sets/shayalvaghasiya/fraud-transactions and xente fraud detection challenge (XFDC) (https://www.kaggle.com/datasets/ajay1735/hmeq-data/data).

3.1. Evaluation Metric Measures

Evaluation metric measures such as accuracy have been extensively employed to assess the generalization ability of the RF algorithm in the literature [55]. However, when modeling skewed fat-tailed distributed datasets, accuracy tends to incline towards the majority class [77]. Therefore, we employed four evaluation metric measures that are predominantly applied for skewed fat-tailed distributed datasets. These include area the under the curve (AUC) [64], and geometric mean (GM) [99]. These metric measures aim to strike a balance between specificity (correct classification of the majority class) and sensitivity (correct classification of the minority class), making them suitable for skewed fat-tailed distributed datasets.

Table 2 shows the confusion matrix used to assess the efficacy of hybridization of DL approaches integrated with robust ensemble algorithms. The matrix consists of rows representing actual classes and columns representing predicted classes after applying the hybridization of DL approaches blended with robust ensemble algorithms. True negative (TN) refers to the number of negative instances classified correctly, while false positive (FP) refers to the number of negative instances incorrectly classified as positive. The false negative (FN) refers to the number of positive instances incorrectly classified as negative while true positive (TP) refers to the number of positive instances correctly classified as positive.

Table 2. Confusion matrix for evaluating ML techniques

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

3.1.1. Geometric Mean (GM)

The GM is a prominent evaluation metric for the classification of skewed fat-tailed distributed datasets [53]. This metric estimates the recalls for both positive and negative classes [66]. The GM can be expressed as;

$$\text{GM} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad [10]$$

where sensitivity measures the number of positive instances correctly classified as positive (true positive rate); Sensitivity $= \frac{\text{TP}}{\text{TP+FN}}$. Specificity measures the number of negative instances correctly classified as positive (true negative rate);
Specificity $= \frac{\text{TN}}{\text{TN+FP}}$. The GM metric is crucial for avoiding overfitting to the negative class and underfitting to the positive class. [2].

3.1.2. Area under the Curve (AUC)

The AUC is a commonly used metric measure for evaluating an algorithm's predictive performance, particularly in the context of skewed fat-tailed distributed datasets [52]. The AUC is an alternative metric to accuracy, which exhibits several desirable properties compared to accuracy characteristics, such as statistical consistency and efficient discriminant power [75]. Hand and Till [50] presented a simplified approach to compute the AUC, which is defined as:

$$\text{AUC} = \frac{S_o - \frac{n_o(n_o+1)}{2}}{n_o n_1} \quad [11]$$

where, n_o and n_1 are numbers of positive and negative instances respectively, and $S_o = \sum r_i$, with r_i denoting the the rank of the i^{th} positive instance. This metric does not disproportionately prioritize one class over the other [65], meaning that the AUC is not biased against the minority class in a binary classification problem. However, a notable drawback of AUC is that it can present an overly optimistic view of the algorithm's predictive performance [27].

3.2. Statistical Evaluation Approaches

The predictive results need to be statistically examined; without employing robust statistical tests, the results may be misleading. In the literature, two approaches are widely used for statistical verification: parametric and non-parametric tests. However, datasets often do not conform to the normality assumption, which is a prerequisite for parametric tests. Thus, we adopted a non-parametric Friedman test [37], and the Nemenyi test [37] as alternatives to the parametric tests that include the analysis of variance (ANOVA) test and Tukey test. The Friedman test is employed for different DL approaches coupled with an ensemble algorithm to establish statistical differences based on GM, and AUC metrics at 95% confidence interval. The null hypothesis of the

Friedman test posits that DL approaches coupled with robust ensemble algorithms have similar predictive performance. The Friedman test statistic is based on the average ranked (AR) performance of the DL approach coupled with an ensemble algorithm on each dataset and is calculated as follows:

$$\chi^2 = \frac{12D}{k(k+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{k(k+1)^2}{4}\right] \quad [12]$$

where, $AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^2$, with D denoting the number of datasets in this study, k represents the total number of DL approaches coupled with ensemble algorithms, and r_i^j denotes the rank of ensemble algorithm blended with DL approach j on dataset j. The χ_F^2 follows a chi-square distribution with $K-1$ degrees of freedom. A small *p-value* (typically less than 5% level of significance) indicates evidence against the null hypothesis, suggesting that significant differences exist among the algorithms being evaluated [41]. The Friedman statistic is particularly effective for this type of data analysis procedure because it is less susceptible to outliers. If the null hypothesis is rejected, a post hoc procedure is necessary, typically conducted using the Durbin-Conover test. The post hoc test is a statistical method used for pairwise comparisons to identify specific differences between groups after establishing that overall differences exist among them [29]. After rejecting the null hypothesis of the Friedman test, the Durbin-Conover test assesses the significant differences between pairs of group medians of GM and AUC [**?**]. The test statistic for each pair can be expressed as:

$$D_{ij} = \frac{R_i - R_j}{\sqrt{\frac{(n(n+1)}{12}}} \quad [13]$$

where, ranks of group i and j, R_i, R_j= rank sums for groups i and j. The p-values are adjusted using methods such as Holm-Bonferroni corrections to account for multiple comparisons [46].

3.3. Experimental Flow Design

Figure 2 presents the framework of this study, aiming to investigate the efficacy of hybridized DL approach, specifically our novelty SMOTEENN-ENC, for handling skewed fat-tailed distributed datasets with inherent diverse features. Each dataset was split into two sets using a stratified approach; the training set constituted 70% of the dataset, while the remaining 30% was allocated to the test set. The training sets were used for modelling purposes, and the test set was employed for evaluation of the robust ensemble algorithms. For each training set, a preprocessing step was applied using SMOTEENN-ENC, along with other DL techniques suggested in the literature for effectively managing skewed fat-tailed distributed datasets with inherent diverse features.

The pre-processed training sets were then modelled 100 times to avoid bias [131], using the widely referenced ensemble classification algorithms, the RF and XGBoost.

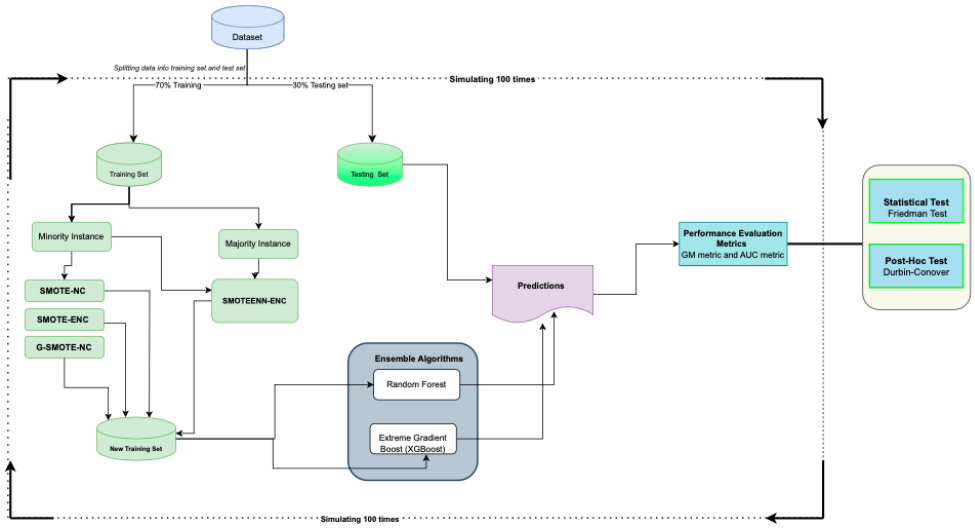


Figure 2. Schematic representation of the proposed framework.

In this study, we selected five parameters for the RF algorithm and three parameters for the XGBoost algorithm, aiming to optimize predictions for skewed fat-tailed distributed datasets with diverse features. The parameter selection employed a k-fold grid search, a well-known cross-validation technique. For the RF algorithm, the parameters included the number of trees {100, 200, 300, 400, 500}, maximum depth {10, 20, 30, 50}, minimum sample split {2, 5, 10}, minimum sample leaf {1, 2, 4, 5, 10}, and maximum leaf nodes {10, 20, 30, 40}. The XGBoost algorithm considered the number of trees {100, 200, 300, 400, 500}, maximum depth {10, 20, 30, 50}, and learning rates {0.01, 0.001, 0.0001}. Their performance was evaluated based on GM and AUC using the test data for each dataset. Statistical tests such as Friedman, and Durbin Conover were employed to validate the computational efficiency of our proposed framework.

4. Predictive Analysis

In this section, we presented the results of datasets characterized by skewed fat-tailed distributed datasets coupled with inherent diverse features. The results are categorized into three scenarios namely, mildly, moderately and extremely. We primarily focused on the hybridization of DL approaches with robust ensemble algorithms.

4.1. Results for mildly skewed fat-tailed distributed datasets

The predictive performance, measured based on the averages and standard deviations (SD) of the GM and AUC metrics for the hybridization of DL approaches integrated with robust ensemble algorithms, is recorded in Table 3. These methods were applied to mildly skewed fat-tailed distributed datasets with inherent diverse features, including the ACD, DCCC, and GCD datasets. The DL approaches and ensemble algorithms with better predictive performance are highlighted in bold for each evaluation metric in each dataset.

Table 3. Mean ± SD of GM and AUC metric for hybridization of DL approaches and ensemble algorithms applied to mildly skewed fat-tailed distributed datasets.

Algorithm	Metric	Datasets	SMOTE-NC	SMOTE-ENC	G-SMOTE-NC	SMOTEENN-ENC
RF	GM	ACD	0.8442 ± 0.0084	$\mathbf{0.8680} \pm 0.0089$	0.8120 ± 0.0026	0.8557 ± 0.0001
		DCCC	0.6965 ± 0.0010	0.6982 ± 0.0008	0.6050 ± 0.0015	$\mathbf{0.7012} \pm 0.0007$
		GCCD	0.6797 ± 0.0060	0.6740 ± 0.0065	0.5412 ± 0.0079	0.6933 ± 0.0060
	AUC	ACD	0.9257 ± 0.0029	0.9317 ± 0.0022	0.9180 ± 0.0024	0.9115 ± 0.0028
		DCCC	0.7755 ± 0.0003	0.7781 ± 0.0003	0.7725 ± 0.0003	0.7738 ± 0.0003
		GCCD	0.7791 ± 0.0017	0.7740 ± 0.0019	$\mathbf{0.7802} \pm 0.0019$	0.7660 ± 0.0017
XGBoost	GM	ACD	0.8251 ± 0.0110	0.8396 ± 0.0107	0.8380 ± 0.0137	0.8591 ± 0.0075
		DCCC	0.6073 ± 0.0025	0.6708 ± 0.0026	0.6055 ± 0.0023	0.6884 ± 0.0019
		GCCD	0.6442 ± 0.0103	0.6879 ± 0.0117	0.6525 ± 0.0118	$\mathbf{0.7028} \pm 0.0116$
	AUC	ACD	0.9402 ± 0.0035	0.9337 ± 0.0032	$\mathbf{0.9451} \pm 0.0033$	0.9342 ± 0.0020
		DCCC	0.7776 ± 0.0012	0.7654 ± 0.0013	$\mathbf{0.7806} \pm 0.0012$	0.7631 ± 0.0012
		GCCD	0.7747 ± 0.0069	0.7776 ± 0.0060	0.7794 ± 0.0071	0.7846 ± 0.0051

The GM metrics varied from an average of 54.12% to 86.80% for DL approaches coupled with robust ensemble algorithms. Only 41.67% of the DL approaches integrated with robust ensemble algorithms exhibited an average GM greater than 70.00%. These results suggest that the majority of the DL approaches combined with robust ensemble algorithms performed fairly. The RF algorithm combined with SMOTE-ENC and SMOTEENN-ENC achieved maximum averages of 86.80% and 70.12%, respectively, for the ACD and DCCC datasets. Conversely, the XGBoost algorithm blended with SMOTEENN-ENC attained the highest GM metric of 70.28% for the GCCD datasets. The range of AUC metrics for DL approaches integrated with robust ensemble algorithms is from 76.31% to 94.51%. Considering that the DL approaches coupled with robust ensemble algorithms exhibited 100% GM metrics greater than 70.00%, the results of the AUC metrics demonstrate the effectiveness of coupling DL approaches with robust ensemble algorithms. The maximum average AUC metrics were obtained by the XGBoost algorithm coupled with SMOTEENN-ENC, achieving 94.51% and

78.06% for the ACD and DCCC datasets, respectively. In contrast, the RF algorithm integrated with SMOTE-ENC yielded the best AUC metric of 78.02% for the GCCD dataset.

Thus, the SMOTEENN-ENC offered better computational efficiency in the predictive capabilities of ensemble algorithms for the three mildly skewed fat-tailed distributed datasets with inherent diverse features. However, the proposed DL approach may not always exhibit the best metrics in some instances; thus, the conclusions may be inconclusive. To address this, we further performed the Friedman test and the Durbin-Conover (post-hoc) test to draw well-informed conclusions about the computational efficiency of the proposed framework by leveraging the mildly skewed fat-tailed distributed datasets with inherent diverse features. Figure 3 illustrates the violin diagrams that visually depict the median predictive capability of the DL approaches and ensemble algorithms on mildly skewed fat-tailed distributed datasets with inherent diverse features in terms of the GM metric. The median GM metrics vary from 65.00% to 70.00%, with only 50.00% of the DL approaches and ensemble algorithms attaining a median greater than or equal to 70.00%. The highest median GM metric of 70.00% was achieved by both SMOTE-ENC and SMOTEENN-ENC combined with either the RF or XGBoost algorithm. However, the Friedman test yielded a *p-value* of $3.31 \times 10^{-211} < 0.05$ ($\chi^2_{\mathrm{F}} = 997.97 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$), suggesting a statistically significant difference between the DL approaches and ensemble algorithms on mildly skewed fat-tailed distributed datasets with inherent diverse features with respect to the GM metric. These results prompted the researchers to perform the Durbin-Conover test, which identifies pairs of DL approaches and ensemble algorithms that are significantly different at the 5% level of significance. The Durbin-Conover test results indicated that the DL approaches and ensemble algorithms that obtained a median of 70.00% are statistically significant different in terms of the GM metric, except for the RF and XGBoost algorithm coupled with SMOTE-ENC and SMOTEENN-ENC, respectively. Notably, SMOTEENN-ENC integrated with the RF algorithm demonstrated consistent variability (see Table 3), and exhibited a statistically significance difference from the RF and XGBoost algorithm coupled with SMOTE-ENC and SMOTEENN-ENC, respectively, for the GM metric across the three aggregated mildly skewed fat-tailed distributed datasets. Violin diagrams presented in Figure 4 visually depict the superior predictive performance (highest median AUC metric) of DL approaches and ensemble algorithms on mildly skewed fat-tailed distributed datasets. The median AUC metrics range from 77.00% to 79.00%, indicating good predictive performance by the DL approaches and ensemble algorithms on these datasets. The XGBoost algorithm integrated with SMOTEENN-ENC achieved the best median AUC metric of 79.00%, which marginally differs from the DL approaches and ensemble algorithms that obtained an AUC metric of 78.00%. The Friedman test shown in Figure 4 reported a *p-value* of $3.31 \times 10^{-211} < 0.005$ ($\chi^2_{\mathrm{F}} = 394.07 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$), implying that there is a statistically significant difference between the DL approaches and ensemble algorithms on mildly skewed fat-tailed distributed datasets with inherent diverse features using the GM metric. The Durbin-Conover test displayed pairs of DL approaches and ensemble algorithms that are significantly different at the 5% level of significance. The results

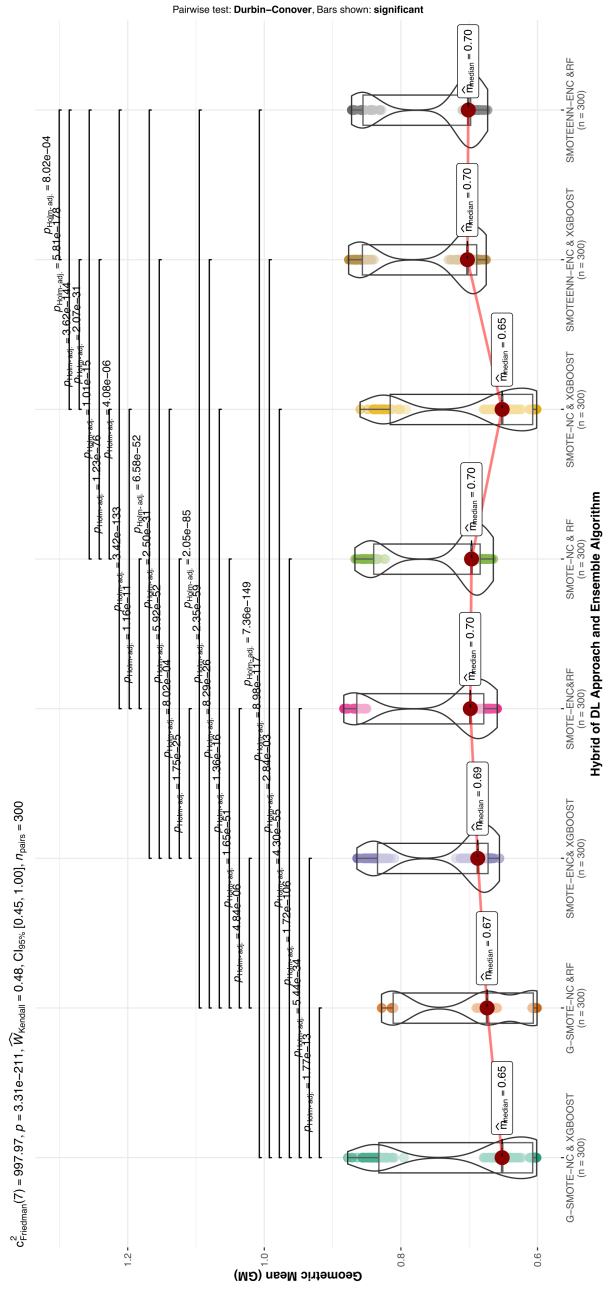


Figure 3. Friedman and Durbin-Conover tests based on GM metric, for hybridization of DL approaches and ensemble algorithms applied to mildly skewed fat-tailed distributed datasets.

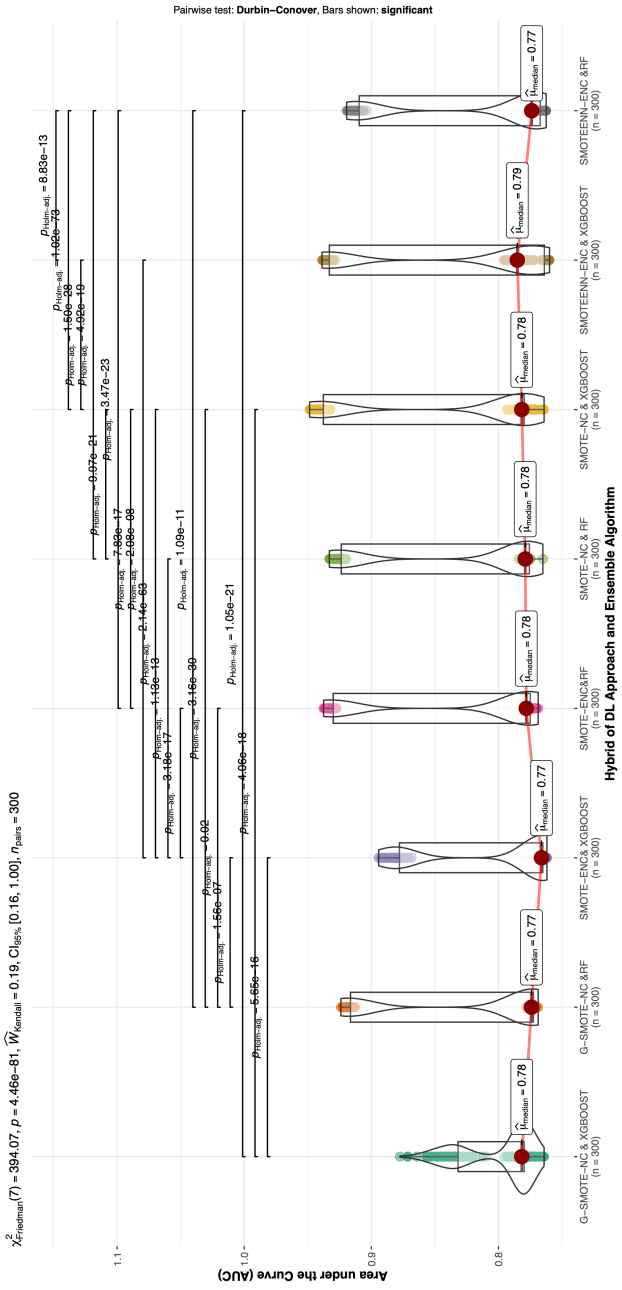


Figure 4. Friedman and Durbin-Conover tests based on AUC metric for hybridization of DL approaches and ensemble algorithms applied to mildly skewed fat-tailed distributed datasets.

illustrated that SMOTEENN-ENC integrated with the XGBoost algorithm may not significantly differ from SMOTE-ENC coupled with the XGBoost algorithm. The results for the XGBoost algorithm blended with SMOTEENN-ENC, shown in Table 3, demonstrated minimal variability (stable prediction) of AUC metrics compared to the DL approaches and ensemble algorithms that obtained a median AUC metric of 78.00

Overall, the GM and AUC metric results suggest that the XGBoost algorithm integrated with our novel SMOTEENN-ENC significantly enhanced the prediction of mildly skewed fat-tailed distributed datasets with inherent diverse features.

4.2. Results for moderately skewed fat-tailed distributed datasets

Analysis in Table 4 summarizes the averages and standard deviations (SD) of GM and AUC metrics to evaluate the efficacy of the hybridization of DL approaches integrated with robust ensemble algorithms. These hybridization techniques were examined using moderately skewed fat-tailed distributed datasets, which include the CCCP, HMEQ, and GMSD. The results indicate that the DL approaches coupled with robust ensemble algorithms demonstrated excellent predictive performance on moderately skewed fat-tailed distributed datasets, as highlighted in Table 4.

Table 4. Mean ± SD of GM and AUC metric for hybridization of DL approaches and ensemble algorithms applied to moderately skewed fat-tailed distributed datasets.

Algorithm	Metric	Datasets	SMOTE-NC	SMOTE-ENC	G-SMOTE-NC	SMOTEENN-ENC
RF	GM	CCCP	0.9287 ± 0.0014	0.9253 ± 0.0024	0.8291 ± 0.0030	0.9272 ± 0.0021
		HMEQ	0.8335 ± 0.0024	0.8321 ± 0.0018	0.7997 ± 0.0050	0.8313 ± 0.0019
		GMSD	0.3276 ± 0.0037	0.7269 ± 0.0033	0.5324 ± 0.0033	0.7274 ± 0.0002
	AUC	CCCP	0.9777 ± 0.0003	0.9778 ± 0.0003	0.9728 ± 0.0004	0.9743 ± 0.0003
		HMEQ	0.9093 ± 0.0006	0.9050 ± 0.0007	0.9298 ± 0.0006	0.9043 ± 0.0006
		GMSD	0.8569 ± 0.0004	0.8007 ± 0.0018	0.8024 ± 0.0006	0.7893 ± 0.0029
XGBoost	GM	CCCP	0.9289 ± 0.0026	**0.9465** ± 0.0024	0.9314 ± 0.0029	0.9413 ± 0.0020
		HMEQ	**0.8581** ± 0.0043	0.8527 ± 0.0047	0.8421 ± 0.0043	0.8504 ± 0.0040
		GMSD	0.4367 ± 0.0038	0.7243 ± 0.0010	0.5378 ± 0.0028	**0.7244** ± 0.0009
	AUC	CCCP	**0.9922** ± 0.0003	0.9919 ± 0.0003	0.9918 ± 0.0003	0.9891 ± 0.0003
		HMEQ	0.9433 ± 0.0014	0.9436 ± 0.0015	**0.9470** ± 0.0013	0.9422 ± 0.0016
		GMSD	**0.8632** ± 0.0004	0.7339 ± 0.0024	0.8082 ± 0.0009	0.7380 ± 0.0017

The average GM metrics ranged from 32.76% to 94.65%, with 83.33% of the GM metrics exceeding 70.00%. The XGBoost algorithm combined with SMOTE-ENC, SMOTE-NC, and SMOTEENN-ENC achieved maximum AUC metrics of 94.65%, 85.81%, and 72.44%, leveraging the CCCP, HMEQ, and GMSD datasets, respectively.

The average AUC metrics varied from 73.39% to 99.22%, with all DL approaches coupled with robust ensemble algorithms attaining an AUC metric greater than 70.00%. This indicates better predictive capabilities of DL approaches coupled with robust ensemble algorithms on moderately skewed fat-tailed distributed datasets. The highest average AUC metrics were achieved by the XGBoost algorithm integrated with SMOTE-NC, yielding 99.22% and 86.32% for the CCCP and GMSD datasets, respec-

tively, while G-SMOTE-NC attained 94.70% for the HMEQ dataset.

The GM and AUC metrics displayed conflicting results for DL approaches coupled with robust ensemble algorithms on moderately skewed fat-tailed distributed datasets with inherent diverse features. The Friedman and Durbin-Conover tests were performed to validate the computational efficiency of the proposed framework. Figure 5 presents the results of these two tests and displays the violin diagrams for DL approaches coupled with robust ensemble algorithms on moderately skewed fat-tailed distributed datasets based on GM metrics. The median GM metrics ranged from 80.00% to 86.00%, indicating that DL approaches coupled with robust ensemble algorithms excelled in efficiently predicting the moderately skewed fat-tailed distributed datasets with inherent diverse features. The highest median GM metric of 86.00% was obtained by the XGBoost algorithm combined with SMOTE-NC. However, this hybridization demonstrated inconsistency in predictive performance, with a minimum average GM metric of 43.67% and a maximum average GM of 92.89% across the three datasets (see Table 4). The second-best median GM metric of 85.00% was achieved by SMOTE-ENC and SMOTEENN-ENC blended with the XGBoost algorithm. The Friedman test attained a *p-value* of $8.28 \times 10^{-191} < 0.05$ ($\chi^2_{\mathrm{F}} = 903.54 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$), indicating a statistically significant difference in their predictive power. The Durbin-Conover results illustrated that there is no significant difference in terms of predictive capabilities on moderately skewed fat-tailed distributed datasets. Notably, the SMOTEENN-ENC integrated with the XGBoost algorithm demonstrated minimal variability compared to SMOTE-ENC using the same algorithm, as evidenced by minimal variability (see Table 4) across the three combined moderately skewed fat-tailed distributed datasets.

Figure 6 reports the results of the Friedman and Durbin tests and visually illustrates the violin diagrams for DL approaches coupled with robust ensemble algorithms on moderately skewed fat-tailed distributed datasets with inherent diverse features, utilizing the AUC metrics. The median AUC for DL approaches coupled with robust ensemble algorithms ranged from 90.00% to 95.00%. The highest median AUC metric was obtained by the G-SMOTE-NC blended with the XGBoost algorithm, with only a marginal difference compared to the XGBoost algorithm integrated with either SMOTE-NC, SMOTE-ENC or SMOTEENN-ENC. The Friedman test secured a *p-value* of $3.91 \times 10^{-240} < 0.05$ ($\chi^2_{\mathrm{F}} = 1,131.82 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$). The results suggest that there is statistically significant difference in their predictive performance. The Durbin-Conover test further indicated that the aforementioned techniques are statistically significantly different in terms of AUC metrics. Thus, G-SMOTE-NC combined with the XGBoost algorithm significantly outperforms the alternative DL approaches coupled with robust ensemble algorithms.

Overall, the SMOTEENN-ENC and SMOTE-ENC, coupled with the XGBoost algorithm, demonstrated effectiveness in predicting moderately skewed fat-tailed distributed datasets with inherently diverse features, using the GM and AUC metrics, respectively. This suggests that our proposed novelty is a viable option for mitigating the aberrations presented in real-world datasets.

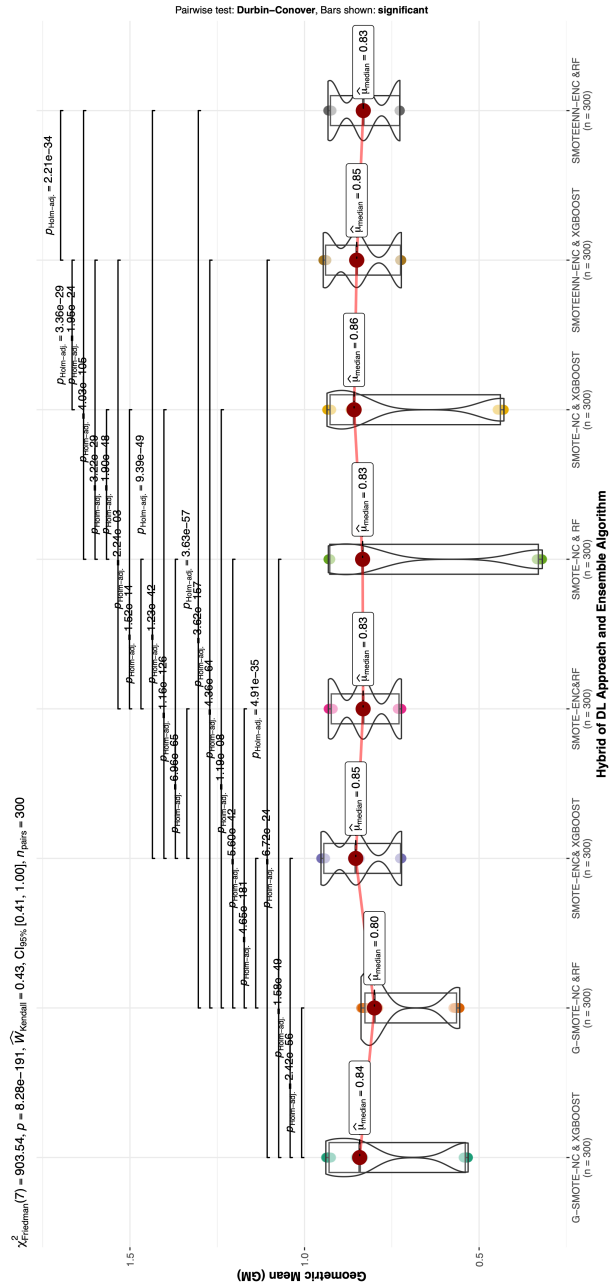


Figure 5. Friedman and Durbin-Conover tests based on GM metric, for hybridization of DL approaches and ensemble algorithms applied to moderately skewed fat-tailed distributed datasets.

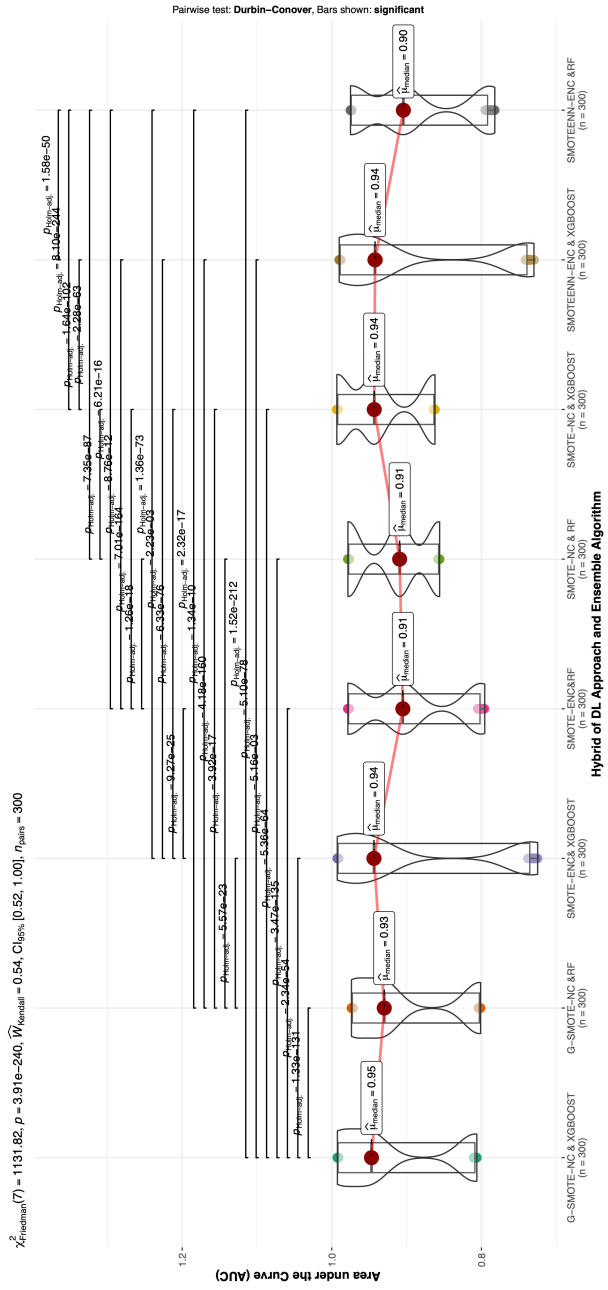


Figure 6. Friedman and Durbin-Conover tests based on AUC metric, for hybridization of DL approaches and ensemble algorithms applied to moderately skewed fat-tailed distributed datasets.

4.3. Results for extremely skewed fat-tailed distributed datasets

The robustness of DL approaches integrated with ensemble algorithms was examined on extremely skewed fat-tailed distributed datasets using GM and AUC metrics, as shown in Table 4. This framework utilized real-world datasets, including the CCFD, FT, and XFDC datasets. The DL approaches and ensemble algorithms that achieved the best predictive performance are highlighted in bold in Table 4.

Table 5. Mean ± SD of GM and AUC metric for hybridization of DL approaches and ensemble algorithms applied to extremely skewed fat-tailed distributed datasets.

Algorithm	Metric	Datasets	SMOTE-NC	SMOTE-ENC	G-SMOTE-NC	SMOTEENN-ENC
RF	GM	CCFD	0.6582 ± 0.0212	0.5849 ± 0.0036	0.5878 ± 0.0294	0.5900 ± 0.0040
		FT	0.5739 ± 0.0048	0.5798 ± 0.0055	0.5430 ± 0.0022	0.5593 ± 0.0022
		XFDC	0.9076 ± 0.0098	0.8578 ± 0.0077	0.7634 ± 0.0307	0.8601 ± 0.0064
	AUC	CCFD	0.6439 ± 0.0012	0.7370 ± 0.0233	0.6489 ± 0.0002	0.7396 ± 0.0018
		FT	0.6101 ± 0.0021	0.5764 ± 0.0009	0.5491 ± 0.0054	0.5798 ± 0.0008
		XFDC	0.9946 ± 0.0028	0.9521 ± 0.0008	0.9936 ± 0.0021	0.9522 ± 0.0008
XGB	GM	CCFD	0.4992 ± 0.0081	0.9211 ± 0.0011	0.3887 ± 0.0125	$\mathbf{0.9221} \pm 0.0011$
		FT	0.9021 ± 0.0021	$\mathbf{0.9605} \pm 0.0021$	0.8910 ± 0.0014	0.9599 ± 0.0023
		XFDC	$\mathbf{0.9735} \pm 0.0007$	0.9698 ± 0.0033	0.6902 ± 0.0179	0.9730 ± 0.0017
	AUC	CCFD	$\mathbf{0.9832} \pm 0.0002$	0.9778 ± 0.0002	0.9760 ± 0.0017	0.9777 ± 0.0002
		FT	0.9972 ± 0.0002	0.9907 ± 0.0007	$\mathbf{0.9973} \pm 0.0003$	0.9901 ± 0.0007
		XFDC	$\mathbf{0.9998} \pm 0.0000$	0.9995 ± 0.0001	0.9988 ± 0.0001	0.9995 ± 0.0001

The GM metrics for DL approaches combined with robust ensemble algorithms fluctuate from 38.87% to 97.35%, with 45.83% of GM metrics exceeding 70.00%. These results suggest fair predictive performance for DL approaches combined with robust ensemble algorithms on extremely skewed fat-tailed distributed datasets. The highest average GM metrics were achieved by the XGBoost algorithm blended with SMOTEENN-ENC (92.21%), SMOTE-ENC (96.05%), and SMOTE-NC (97.35%) for the CCFD, FT, and XFDC datasets, respectively.

DL approaches coupled with robust ensemble algorithms exhibited AUC metrics ranging from from 54.91% to 99.98%, with 75.00% of AUC metrics exceeding 70.00%. The proposed framework thrived in efficiently predicting of extremely skewed fat-tailed distributed datasets leveraging the DL approaches coupled with robust ensemble algorithms. The XGBoost algorithm integrated with SMOTE-NC achieved maximum averages of 98.32% and 99.98% for the CCFD and XFDC datasets, respectively, while the G-SMOTE-NC combined with the XGBoost algorithm attained the highest AUC metric for the FT datasets.

Additionally, we performed the Friedman and Durbin-Conover tests to validate the results pre sented in Table 5. Although the metric measures depicted various predictive performances for extremely skewed fat-tailed distributed datasets with inherent diverse features, the Friedman test and Durbin-Conover test results are presented in 7. The violin diagrams also display the median and variability of the GM metrics. The median of the GM metrics varies from 58.00% to 96.00%, with the highest value obtained by the XGBoost algorithm integrated with either SMOTE-ENC or SMOTEENN-ENC. The results of the Friedman test displayed a *p-value* of $0.00 < 0.05$ ($\chi^2_{\mathrm{F}} = 1,631.36 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$), indicating a signifi-

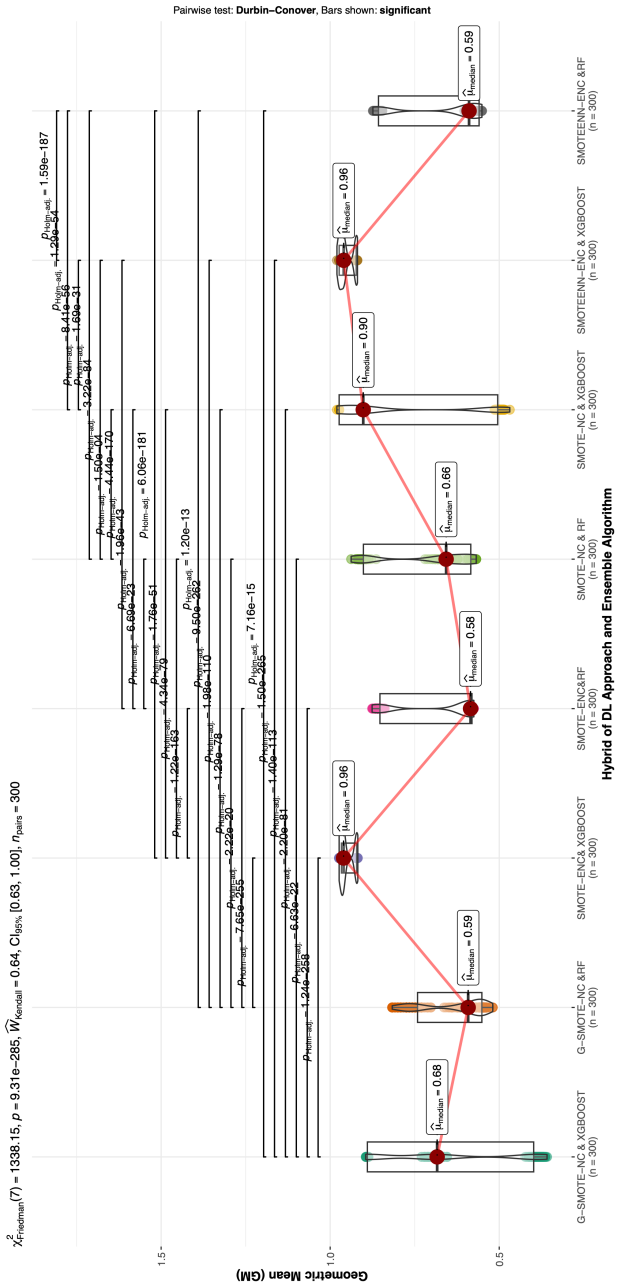


Figure 7. Friedman and Durbin-Conover tests based on GM metric, for hybridization of DL approaches and ensemble algorithms applied to extremely skewed fat-tailed distributed datasets.

cant difference among the DL approaches integrated with ensemble algorithms. The Durbin-Conover test revealed no significant difference in the predictive performance for the best approaches in terms of AUC. However, the SMOTEENN-ENC coupled with the XGBoost algorithm exhibited consistent GM metrics (less volatile GM metrics) when evaluated against SMOTE-ENC for the same algorithm using the three aggregated, extremely skewed fat-tailed distributed datasets (see Table 5). The median AUC metrics ranged from 65.00% to 100%. This prompted further investigation through the Friedman test to determine if there was a significant difference in predictive performance for the DL approaches combined with ensemble algorithms. The results of the Friedman test are shown in Figure 8, were the *p-value* is of $0.00 < 0.05$ ($\chi^2_{\mathrm{F}} = 1,338 > \chi^2_{(7)}(\alpha = 0.05) = 2.1670$).8, These results suggest a statistically significant significant difference in the predictive performance of DL approaches combined with ensemble algorithms. Despite this, the maximum median was attained by G-SMOTE-NC and SMOTE-ENC blended with the XGBoost algorithm. The Durbin-Conover results in Figure 8 demonstrated a statistically insignificant difference in their predictive performance based on AUC metrics. However, SMOTE-ENC integrated with the XGBoost algorithm exhibited minimal variability (less volatile AUC metrics) compared to G-SMOTE-NC for the same algorithm (see Table 5). Researchers may not rely solely on AUC metrics for extremely skewed fat-tailed distributed datasets with inherent diverse features due to inconsistency exhibited by DL approaches integrated with ensemble algorithms. Therefore, our conclusion was based on GM metrics, which depicted that SMOTEENN-ENC coupled with XGBoost algorithm is the most stable for extremely skewed fat-tailed distributed datasets with inherent diverse features.

5. Conclusion Remarks

To tackle the adverse effects posed by skewed fat-tailed distributed datasets coupled with diverse features that are inherent in real-world applications, recent researchers have suggested various SMOTE mechanisms, which include the SMOTE-NC, G-SMOTE-NC, and SMOTE-ENC. The strength of SMOTE-based approaches resides in their capability to mitigate the risk of overfitting due to synthetic generation of artificial instances. However, these approaches substantially fail to capture effectively both nominal and continuous features due to the ED metric, complicating the determination of the k-nearest neighbors for the nominal features. In response, SMOTE-NC, a generalization of SMOTE, adopts a modified ED metric in computation of k-nearest neighbors which guarantees a better prediction power for the ML algorithm. Also, literature revealed that SMOTE-based approaches often introduce noisy instances, compromising the computational efficiency of the ML algorithms. To address this issue, G-SMOTE-NC was proposed to alleviate the noisy information by generating instances in a geometric fashion called hyper-spheriod instead of the linear interpolation method. Despite the effectiveness of all the aforementioned approaches, they overlooked the aspect of nominal features consisting of mismatching labels which may lead to diminished predictive performance of ML algorithms. SMOTE-ENC was

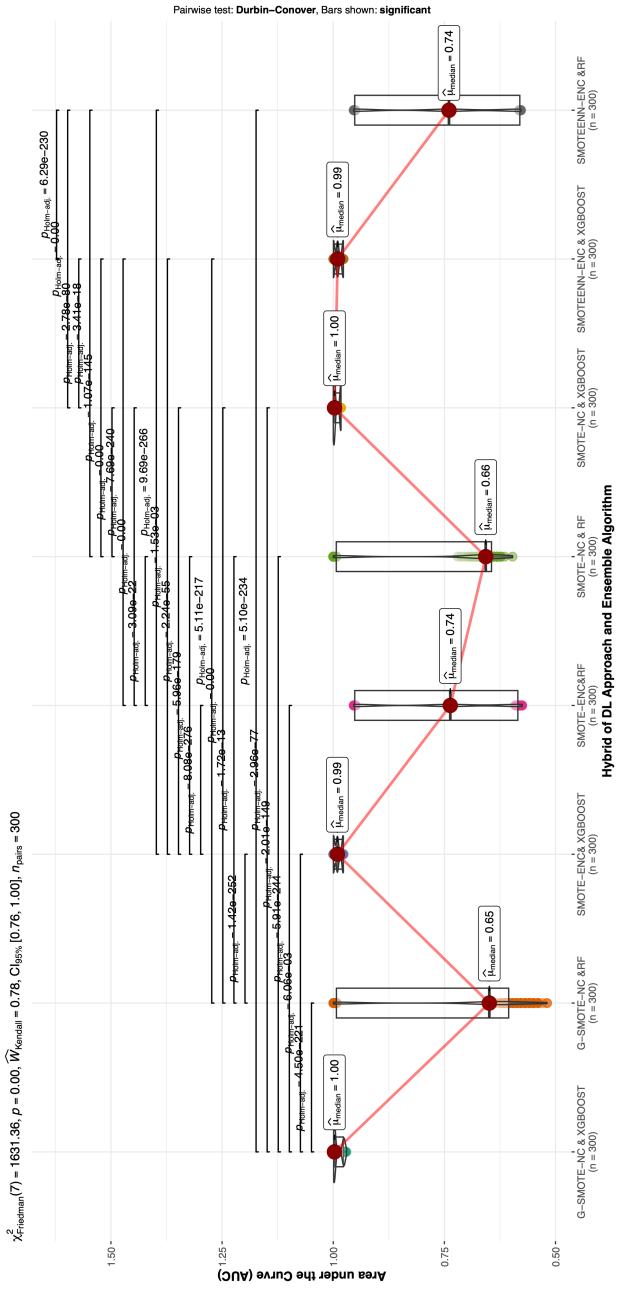


Figure 8. Friedman and Durbin-Conover tests based on AUC metric, for hybridization of DL approaches and ensemble algorithms applied to extremely skewed fat-tailed distributed datasets.

introduced to resolve this issue. However, to best of the author's knowledge there is a void in research studies in the literature that simultaneously addresses aberrations exhibited by nominal features consisting of mismatching labels and eliminating noisy instances.

To overcome the drawbacks presented by noisy instances introduced by SMOTE mechanism, hybrid DL approaches such as SMOTE-Tomek and SMOTEENN have been broadly employed as remedies for noisy data points. The SMOTEENN offers superior predictive capability compared to SMOTE-TOMEK in effectively discarding noisy information and has also demonstrated robustness in handling skewed fat-tailed distributed datasets. However, this approach is tailored for continuous features exclusively while ignoring the nominal features which are inherent in real-world applications. Building upon its success, we extended the principle of encoding of nominal and continuous features to SMOTEENN, resulting in our proposed hybridization novelty called SMOTEENN-ENC. This approach effectively handles skewed fat-tailed distributed datasets with inherent diverse features, particularly with nominal features that consist of mismatching labels, while avoiding overfitting as well as eliminating the noisy data points.

The efficacy of our proposed hybridization framework was coupled with the most successful ensemble algorithms in the domain of skewed fat-tailed distributed datasets that include the RF algorithm and XGBoost algorithm. These algorithms have capabilities of capturing complex non-linear features, managing outliers and noisy instances. Additionally, they are well-known for their superior predictive capabilities and excellent computational speed. However, these algorithms tend to be susceptible to aberrations presented by skewed fat-tailed distributed datasets which are further exacerbated by diverse features that are inherent in real-world datasets. Therefore, we proposed the hybridization of DL approaches coupled with the most efficient ensemble algorithms. The proposed framework was verified using nine (9) real-world credit risk datasets. Three datasets represented each category of skewed fat-tailed distributed datasets that include; mildly, moderately and extremely scenarios. The evaluation metrics that were adopted for this study are GM and AUC, which effectively balance the prediction of defaults and non-defaults. Furthermore, we performed robust statistical tests to compare the significance of differences (median of the GM and AUC metrics separately) between the hybrid DL approaches coupled ensemble algorithms which include the Friedman test for repeated measures and Durbin Conover pairwise comparison test.

The results suggest that SMOTEENN-ENC coupled with XGBoost algorithms demonstrate superior predictive performance when applied to mildly skewed fat-tailed distributed datasets based on the GM and AUC metrics, respectively. For the moderately skewed fat-tailed distributed datasets, SMOTEENN-ENC and SMOTE-NC blended with the XGBoost algorithm offered better predictive power. Lastly, SMOTEENN-ENC integrated with the XGBoost algorithm yielded stable predictive power for extremely skewed fat-tailed distributed datasets. Overall, our novelty SMOTEENN-

ENC offered the best computational efficiency in enhancing the predictive performance when exposed to skewed fat-tailed distributed with inherent diverse features. Future research could extend the application of SMOTEENN-ENC for handling skewed fat-tailed distributed datasets with inherent diverse features coupled with high-dimensional features.

References

[1] Ahmad M., Aftab S., Muhammad S. S., Ahmad S., Machine learning techniques for sentiment analysis: A review, *International Journal of Multidisciplinary Sciences and Engineering*, 2017, 8(3), 27.

[2] Akosa J., Predictive accuracy: A misleading performance measure for highly imbalanced data, in *Proceedings of the SAS Global Forum*, 2017, 12, 1–4. SAS Institute Inc. Cary, NC, USA.

[3] Arafa A., El-Fishawy N., Badawy M., Radad M., RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification, *Journal of King Saud University-Computer and Information Sciences*, 2022, 34(8), 5059–5074.

[4] Araf I., Idri A., Chairi I., Cost-sensitive learning for imbalanced medical data: a review, *Artificial Intelligence Review*, 2024, 57(4), 1–72.

[5] Asselman A., Khaldi M., Aammou S., Enhancing the prediction of student performance based on the machine learning XGBoost algorithm, *Interactive Learning Environments*, 2023, 31(6), 3360–3379.

[6] Awaji B., Senan E.M., Olayah F., Alshari E.A., Alsulami M., Abosaq H.A., Alqahtani J., Janrao P., Hybrid techniques of facial feature image analysis for early detection of autism spectrum disorder based on combined CNN features, *Diagnostics*, 2023, 13(18), 2948.

[7] Azhar N.A., Pozi M.S.M., Din A.M., Jatowt A., An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis, *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[8] Bader-El-Den M., Teitei E., Perry T., Biased random forest for dealing with the class imbalance problem, *IEEE Transactions on Neural Networks and Learning Systems*, 2018, vol. 30, no. 7, pp. 2163–2172, IEEE.

[9] Bansal A., Jain A., Analysis of focussed under-sampling techniques with machine learning classifiers, *2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2021, 91–96.

[10] Batista G.E.A.P.A., Prati R.C., Monard M.C., A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD Explorations Newsletter*, 2004, 6(1), 20–29.

[11] Belgiu M., Drăguț L., Random forest in remote sensing: A review of applications and future directions, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, 114, 24–31.

[12] Bi J., Zhang C., An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme, *Knowledge-Based Systems*, 2018, vol. 158, pp. 81–93, Elsevier.

[13] Brown I., Mues C., An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Systems with Applications*, 2012, 39(3), 3446–3453.

[14] Breiman L., Random forests, *Machine Learning*, 2001, 45, 5–32.

[15] Budjač R., Nikmon M., Schreiber P., Zahradníková B., Janáčová D., Automated machine learning overview, *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, 2019, 27(45), 107–112.

[16] Carmona P., Climent F., and Momparler A., Predicting failure in the US banking sector: An extreme gradient boosting approach, *International Review of Economics & Finance*, 2019, vol. 61, pp. 304–323, Elsevier.

[17] Caouette J.B., Altman E.I., Narayanan P., Nimmo R., Managing credit risk: The great challenge for global financial markets, *John Wiley & Sons*, 2011.

[18] Carrington A.M., Fieguth P.W., Qazi H., Holzinger A., Chen H.H., Mayr F., Manuel D.G., A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, *BMC Medical Informatics and Decision Making*, 2020, 20, 1–12.

[19] Chandra W., Suprihatin B., Resti Y., Median-KNN Regressor-SMOTE-Tomek links for handling missing and imbalanced data in air quality prediction, *Symmetry*, 2023, 15(4), 887.

[20] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, 2002, 16, 321–357.

[21] Chen T., Guestrin C., XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785–794.

[22] Chen N., Ribeiro B., Chen A., Financial credit risk assessment: a recent review, *Artificial Intelligence Review*, 2016, 45, 1–23, Springer.

[23] Chen W., Liu C., Yu W., Wen J., Zhou L., Asymmetric Loss-Oriented Cost Sensitive Learning Model for Credit Risk Assessment, *(Journal Name Pending)*, 2024.

[24] Cheng K., Zhang C., Yu H., Yang X., Zou H., Gao S., Grouped SMOTE with noise filtering mechanism for classifying imbalanced data, *IEEE Access*, 2019, 7, 170668–170681.

[25] Cheng L.-C., Wu Y.-T., Chao C.-T., Wang J.-H., Detecting fake reviewers from the social context with a graph neural network method, *Decision Support Systems*, 2024, vol. 179, p. 114150, Elsevier.

[26] Couronné R., Probst P., Boulesteix A.-L., Random forest versus logistic regression: A large-scale benchmark experiment, *BMC Bioinformatics*, 2018, 19, 1–14.

[27] Davis J., Goadrich M., The relationship between Precision-Recall and ROC curves, in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, 233–240.

[28] Demšar J., Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research*, 2006, 7, 1–30.

[29] De Corso E., Pasquini E., Trimarchi M., La Mantia I., Pagella F., Ottaviano G., Garzaro M., Pipolo C., Torretta S., Seccia V., et al., Dupilumab in the treatment of severe uncontrolled chronic rhinosinusitis with nasal polyps (CRSwNP): a multicentric observational phase IV real-life study (DUPIREAL), *Allergy*, 2023, 78(10), 2669–2683.

[30] Ding H., Sun Y., Wang Z., Huang N., Shen Z., Cui X., RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification, *Information Processing & Management*, 2023, 60(2), 103235.

[31] Dong J., Chen Y., Yao B., Zhang X., Zeng N., A neural network boosting regression model based on XGBoost, *Applied Soft Computing*, 2022, 125, 109067.

[32] Douzas G., Bacao F., Geometric SMOTE: a geometrically enhanced drop-in replacement for SMOTE, *Information Sciences*, 2019, 501, 118–135.

[33] Elreedy D., Atiya A.F., A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance, *Information Sciences*, 2019, 505, 32–64.

[34] Ferreira A.J., Figueiredo M.A.T., Boosting algorithms: A review of methods, theory, and applications, *Ensemble Machine Learning: Methods and Applications*, 2012, 35–85.

[35] Fernández A., García S., Herrera F., Chawla N.V., SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, *Journal of Artificial Intelligence Research*, 2018, 61, 863–905.

[36] Fränti P., Mariescu-Istodor R., Soft precision and recall, *Pattern Recognition Letters*, 2023, 167, 115–121.

[37] Friedman M., A comparison of alternative tests of significance for the problem of m rankings, *The Annals of Mathematical Statistics*, 1940, 11(1), 86–92.

[38] Fonseca J., Bacao F., Geometric SMOTE for imbalanced datasets with nominal and continuous features, *Expert Systems with Applications*, 2023, 234, 121053.

[39] Fu G.-H., Yi L.-Z., Pan J., Tuning model parameters in class-imbalanced learning with precision-recall curve, *Biometrical Journal*, 2019, 61(3), 652–664.

[40] Galar M., Fernández A., Barrenechea E., Bustince H., Herrera F., A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2011, 42(4), 463–484.

[41] García S., Fernández A., Luengo J., Herrera F., Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, *Information Sciences*, 2010, 180(10), 2044–2064.

[42] Ganganwar V., An overview of classification algorithms for imbalanced datasets, *International Journal of Emerging Technology and Advanced Engineering*, 2012, 2(4), 42–47.

[43] Gnip P., Vokorokos L., Drotár P., Selective oversampling approach for strongly imbalanced data, *PeerJ Computer Science*, 2021, 7, e604.

[44] Gu J., Random Forest Based Imbalanced Data Cleaning and Classification, 2007, Citeseer.

[45] Gu Q., Cai Z., Zhu L., Huang B., Data mining on imbalanced data sets, in *2008 International Conference on Advanced Computer Theory and Engineering*, 2008, 1020–1024, IEEE.

[46] Gür E., Duyan Y.A., Balcı F., Mice make temporal inferences about novel locations based on previously learned spatiotemporal contingencies, *Animal Cognition*, 26(3), 771–779, 2023.

[47] Guo K., Wan X., Liu L., Gao Z., Yang M., Fault diagnosis of intelligent production line based on digital twin and improved random forest, *Applied Sciences*, 2021, 11(16), 7733.

[48] Han S., Williamson B.D., Fong Y., Improving random forest predictions in small datasets from two-phase sampling designs, *BMC Medical Informatics and Decision Making*, 2021, 21, 1–9, Springer.

[49] Han H., Wang W.-Y., Mao B.-H., Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, *International Conference on Intelligent Computing*, 2005, 878–887, Springer.

[50] Hand D. J., Till R. J., A simple generalisation of the area under the ROC curve for multiple class classification problems, *Machine Learning*, 2001, 45, 171–186.

[51] Hancock J.T., Khoshgoftaar T.M., Johnson J.M., Evaluating classifier performance with highly imbalanced big data, *Journal of Big Data*, 2023, 10(1), 42.

[52] Hanley J.A., McNeil B.J., The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 1982, 143(1), 29–36.

[53] He H., Garcia E.A., Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(9), 1263–1284, IEEE.

[54] He H., Ma Y., Imbalanced learning: foundations, algorithms, and applications, 2013, John Wiley & Sons.

[55] Hossin M., Sulaiman M.N., A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process*, 2015, 5(2), 1.

[56] Islam A., Belhaouari S.B., Rehman A.U., Bensmail H., KNNOR: An oversampling technique for imbalanced datasets, *Applied Soft Computing*, 2022, 115, 108288.

[57] Ishwaran H., Lu M., Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival, *Statistics in Medicine*, 2019, 38(4), 558–582.

[58] Jiang C., Lu W., Wang Z., Ding Y., Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring, *Expert Systems with Applications*, 2023, 213, 118878, Elsevier.

[59] Johnson J.M., Khoshgoftaar T.M., Survey on deep learning with class imbalance, *Journal of Big Data*, 2019, 6(1), 1–54.

[60] Kaur H., Pannu H.S., Malhi A.K., A systematic review on imbalanced data challenges in machine learning: Applications and solutions, *ACM Computing Surveys (CSUR)*, 2019, 52(4), 1–36.

[61] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., LightGBM: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, 2017, 30.

[62] Khalilia M., Chakraborty S., Popescu M., Predicting disease risks from highly imbalanced data using random forest, *BMC Medical Informatics and Decision Making*, 2011, 11, 1–13.

[63] Khoshgoftaar T.M., Golawala M., and Van H.Ja, An empirical study of learning from imbalanced data using random forest, *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, 2007, vol. 2, pp. 310–317, IEEE.

[64] Kim T., Lee J.-S., Maximizing AUC to learn weighted naive Bayes for imbalanced data classification, *Expert Systems with Applications*, 2023, 217, 119564.

[65] Kotsiantis S., Kanellopoulos D., Pintelas P., Handling imbalanced datasets: A review, *GESTS International Transactions on Computer Science and Engineering*, 2006, 30(1), 25–36.

[66] Kraiem M.S., Sánchez-Hernández F., Moreno-García M.N., Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models, *Applied Sciences*, 2021, 11(18), 8546.

[67] Kumar V., Lalotra G.S., Sasikala P., Rajput D.S., Kaluri R., Lakshmanna K., Shorfuzzaman M., Alsufyani A., Uddin M., Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques, *Healthcare*, 2022, 10(7), 1293.

[68] Kyiu A., Tawiah V., IFRS 9 implementation and bank risk, in *Accounting Forum*, 2023, 1–25, Taylor & Francis.

[69] Lai S.B.S., Shahri N.H.N.B.M., Mohamad M.B., Rahman H.A.B., Rambli A.B., Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data, *Mathematics and Statistics*, 2021, 9(3), 379–385.

[70] Laker J., The Evolution of Risk and Risk Management–A Prudential Regulator's Perspective, *Conference–2007*, 2007, Reserve Bank of Australia.

[71] Le T., Vo M.T., Vo B., Lee M.Y., Baik S.W., A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction, *Complexity*, 2019, 2019, 8460934.

[72] Li X., Liu Q., A hybrid sampling algorithm for imbalanced and class-overlap data based on natural neighbors and density estimation, *Knowledge and Information Systems*, 2024, 1–32.

[73] Liang X.W., Jiang A.P., Li T., Xue Y.Y., Wang G.T., LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM, *Knowledge-Based Systems*, 2020, 196, 105845.

[74] Limanto S., Buliali J.L., Saikhu A., GLoW SMOTE-D: Oversampling Technique to Improve Prediction Model Performance of Students Failure in Courses, *IEEE Access*, 2024.

[75] Ling C. X., Huang J., Zhang H., AUC: a statistically consistent and more discriminating measure than accuracy, *Ijcai*, 2003, 3, 519–524.

[76] Liu F., Qian Q., Cost-sensitive variational autoencoding classifier for imbalanced data classification, *Algorithms*, 2022, 15(5), 139.

[77] Liu Y., Yu X., Huang J.X., An Aijun, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets, *Information Processing & Management*, 2011, 47(4), 617–631.

[78] Liu Y., Wang H., Fei Y., Liu Y., Shen L., Zhuang Z., Zhang X., Research on the prediction of green plum acidity based on improved XGBoost, *Sensors*, 2021, 21(3), 930.

[79] Lommers K., Harzli O. E., Kim J., Confronting machine learning with financial research, *arXiv preprint arXiv:2103.00366*, 2021.

[80] Malhotra R., Jain J., Predicting defects in imbalanced data using resampling methods: an empirical investigation, *PeerJ Computer Science*, 2022, 8, e573.

[81] Ma G., Wang Y., Can the Chinese domestic bond and stock markets facilitate a globalising renminbi, *Economic and Political Studies*, 2020, vol. 8, no. 3, pp. 291–311, Taylor & Francis.

[82] Mellor A., Boukir S., Haywood A., and Jones S., Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, 2015, pp. 155–168, Elsevier.

[83] More A.S., Rana D.P., Performance enrichment through parameter tuning of random forest classification for imbalanced data applications, *Materials Today: Proceedings*, 2022, vol. 56, pp. 3585–3593, Elsevier.

[84] Mukherjee M., Khushi M., SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features, *Applied System Innovation*, 2021, 4(1), 18.

[85] Niu A., Cai B., Cai S., Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm, *Complexity*, 2020, vol. 2020, no. 1, p. 8563030, Wiley Online Library.

[86] Ozdemir B., Evolution of risk management from risk compliance to strategic risk management: From Basel I to Basel II, III and IFRS 9, *Journal of Risk Management in Financial Institutions*, vol. 11, no. 1, 2018, pp. 76–85, Henry Stewart Publications.

[87] Pes B., Learning from high-dimensional and class-imbalanced datasets using random forests, *Information*, 2021, 12(8), 286.

[88] Paing M.P., Pintavirooj C., Tungjitkusolmun S., Choomchuay S., Hamamoto K., Comparison of sampling methods for imbalanced data classification in random forest, *2018 11th Biomedical Engineering International Conference (BMEiCON)*, 2018, 1–5.

[89] Prakash A., Thangaraj J., Roy S., Srivastav S., Mishra J. K., Model-aware XGBoost method towards optimum performance of flexible distributed Raman amplifier, *IEEE Photonics Journal*, 2023, 15(4), 1–10.

[90] Qiu W., Credit risk prediction in an imbalanced social lending environment based on XGBoost, *2019 5th International Conference on Big Data and Information Analytics (BigDIA)*, 2019, pp. 150–156, IEEE.

[91] Qolomany B., Al-Fuqaha A., Gupta A., Benhaddou D., Alwajidi S., Qadir J., Fong A. C., Leveraging machine learning and big data for smart buildings: A comprehensive survey, *IEEE Access*, 2019, 7, 90316–90356.

[92] Brennan, P. "A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection." *Institute of Technology Blanchardstown, Dublin, Ireland*, 2012.

[93] Rao C.R., Karl Pearson chi-square test the dawn of statistical inference, *Goodness-of-fit tests and model validity*, 2002, 9–24, Springer.

[94] Ratnaningsih T., SMOTE-MRS: A Novel SMOTE-Multiresolution Sampling Technique for Imbalanced Distribution to Improve Prediction of Anemia, *(Journal Name Pending)*, 2024.

[95] Rawat S.S., Mishra A.K., Review of Methods for Handling Class-Imbalanced in Classification Problems, *arXiv preprint arXiv:2211.05456*, 2022.

[96] Rijsbergen V., Information retrieval; Butterworth, 1978, *J. Librariansh.*, 1979, 11, 237.

[97] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P., An assessment of the effectiveness of a random forest classifier for land-cover classification, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2012, 67, 93–104.

[98] Rodríguez-Galiano V.F., Abarca-Hernández F., Ghimire B., Chica-Olmo M., Atkinson P. M., Jeganathan C., Incorporating spatial variability measures in land-cover classification using Random Forest, *Procedia Environmental Sciences*, vol. 3, 2011, pp. 44–49, Elsevier.

[99] Rout N., Mishra D., Mallick M.K., Handling imbalanced data: a survey, in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016*, 2018, 431–443, Springer.

[100] Schapire R.E., Freund Y., Boosting: Foundations and algorithms, *Kybernetes*, 2013, 42(1), 164–166.

[101] Sarker I. H., Machine learning: Algorithms, real-world applications and research directions, *SN Computer Science*, 2021, 2(3), 160.

[102] Shi S., Li J., Zhu D., Yang F., Xu Y., A hybrid imbalanced classification model based on data density, *Information Sciences*, 2023, 624, 50–67.

[103] Singh A., Ranjan R. K., Tiwari A., Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms, *Journal of Experimental & Theoretical Artificial Intelligence*, 2022, 34(4), 571–598.

[104] Singla P., Domingos P., Discriminative training of Markov logic networks, in *AAAI*, 2005, 868–873.

[105] Sibindi R., Mwangi R. W., Waititu A. G., A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices, *Engineering Reports*, 2023, 5(4), e12599.

[106] Soltanzadeh P., Hashemzadeh M., RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem, *Information Sciences*, 542, 92–111, 2021.

[107] Srivastava J., Sharan A., SMOTEEN Hybrid Sampling Based Improved Phishing Website Detection, *Authorea Preprints*, 2023.

[108] Sun Z., Song Q., Zhu X., Sun H., Xu B., Zhou Y., A novel ensemble method for classifying imbalanced data, *Pattern Recognition*, 2015, 48(5), 1623–1637.

[109] Sun Z., Ying W., Zhang W., Gong S., Undersampling method based on minority class density for imbalanced data, *Expert Systems with Applications*, 2024, 249, 123328.

[110] Szeghalmy S., Fazekas A., A comparative study on noise filtering of imbalanced data sets, *Knowledge-Based Systems*, 2024, 301, 112236.

[111] Tao X., Guo X., Zheng Y., Zhang X., Chen Z., Self-adaptive oversampling method based on the complexity of minority data in imbalanced datasets classification, *Knowledge-Based Systems*, 2023, 277, 110795.

[112] Taye M.M., Understanding of machine learning with deep learning: architectures, workflow, applications and future directions, *Computers*, 2023, 12(5), 91.

[113] Ting K.M., An instance-weighting method to induce cost-sensitive trees, *IEEE Transactions on Knowledge and Data Engineering*, 2002, vol. 14, no. 3, pp. 659–665, IEEE.

[114] Tomašević N., Mladenić D., Class imbalance and the curse of minority hubs, *Knowledge-Based Systems*, 2013, 53, 157–172.

[115] Tomek I., Two modifications of CNN, *IEEE Transactions on Systems, Man, and Cybernetics*, 1976, 6, 769–776.

[116] Uddin Md G., Nash S., Diganta M. T. M., Rahman A., Olbert A. I., Robust machine learning algorithms for predicting coastal water quality index, *Journal of Environmental Management*, 2022, 321, 115923.

[117] Verikas A., Gelzinis A., Bacauskiene M., Mining data with random forests: A survey and results of new tests, *Pattern Recognition*, 2011, 44(2), 330–349.

[118] Wang A.X., Chukova S.S., Nguyen B.P., Synthetic minority oversampling using edited displacement-based k-nearest neighbors, *Applied Soft Computing*, 2023, 148, 110895.

[119] Wang K., Wan J., Li G., Sun H., A hybrid algorithm-level ensemble model for imbalanced credit default prediction in the energy industry, *Energies*, 2022, 15(14), 5206.

[120] Wang C., Deng C., Wang S., Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost, *Pattern Recognition Letters*, 2020, 136, 190–197.

[121] Wilson D.L., Asymptotic properties of nearest neighbor rules using edited data, *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, 3, 408–421.

[122] Yan J., Tang B., He H., Detection of false data attacks in smart grid with supervised learning, *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, 1395–1402.

[123] Yanenkova I., Nehoda Y., Drobyazko S., Zavhorodnii A., Berezovska L., Modeling of bank credit risk management using the cost risk model, *Journal of Risk and Financial Management*, 2021, 14(5), 211.

[124] Yin S., Dey D.K., Valdez E.A., Gan G., Vadiveloo J., Skewed link regression models for imbalanced binary response with applications to life insurance, *arXiv preprint arXiv:2007.15172*, 2020.

[125] Xu T., Comparative Analysis of Machine Learning Algorithms for Consumer Credit Risk Assessment, *Transactions on Computer Science and Intelligent Systems Research*, 2024, vol. 4, pp. 60–67.

[126] Zhang C., Tan K.C., Li H., Hong G.S., A cost-sensitive deep belief network for imbalanced classification, *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 30(1), 109–122.

[127] Zhang C., Wang G., Zhou Y., Yao L., Jiang Z.L., Liao Q., Wang X., Feature selection for high dimensional imbalanced class data based on F-measure optimization, in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2017, 278–283, IEEE.

[128] Zhang W., He Y., Wang L., Liu S., Meng X., Landslide susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing, *Geological Journal*, 2023, 58(6), 2372–2387.

[129] Zhang W., Wu C., Zhong H., Li Y., Wang L., Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization, *Geoscience Frontiers*, 2021, 12(1), 469–477.

[130] Zhao Z., Cui T., Ding S., Li J., Bellotti A.G., Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction, *Mathematics*, 2024, vol. 12, no. 5, p. 701, MDPI.

[131] Zheng M., Wang F., Hu X., Miao Y., Cao H., Tang M., A method for analyzing the performance impact of imbalanced binary data on machine learning models, *Axioms*, 2022, 11(11), 607.

[132] Zhu L., Qiu D., Ergu D., Ying C., Liu K., A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science*, 2019, 162, 503–513.

[133] Zhu T., Lin Y., Liu Y., Synthetic minority oversampling technique for multiclass imbalance problems, *Pattern Recognition*, 2017, 72, 327–340.

[134] Zou Y., Gao C., Gao H., Business failure prediction based on a cost-sensitive extreme gradient boosting machine, *IEEE Access*, 2022, vol. 10, pp. 42623–42639, IEEE.

Received 31.08.2024, Accepted 13.02.2025