

# Self-paced Ensemble and Big Data Identification: A Classification of Substantial Imbalance Computational Analysis

**Shahzadi Bano**

Zhengzhou University

**Weimei Zhi**

[iwzmzhi@zzu.edu.cn](mailto:iwzmzhi@zzu.edu.cn)

Zhengzhou University

**Baozhi Qiu**

Zhengzhou University

**Muhammad Raza**

Xi'an Technological University

**Nabila Sehito**

Zhengzhou University

**Mian Muhammad Kamal**

Southeast University

**Ghadah Aldehim**

Princess Nourah Bint Abdulrahman University

**Nuha Alruwais**

King Saud University



---

## Research Article

**Keywords:** self-paced, big data ensemble, classification, computational, simulation, substantial imbalance

**Posted Date:** September 5th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3310321/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

**Version of Record:** A version of this preprint was published at The Journal of Supercomputing on December 18th, 2023. See the published version at <https://doi.org/10.1007/s11227-023-05828-6>.

# Abstract

**Background:** The concept of self-paced learning in the context of ensemble learning involves the idea of allowing each individual member, or base learner, within an ensemble to learn at its own pace. Ensemble learning refers to a machine learning technique that combines multiple learning models, known as base learners, to improve predictive accuracy and overall performance.

**Motivation:** The research focuses on self-paced ensemble and big data classifications, with considerable data imbalance as a mediating factor. This idea is a brand-new domain with a lot of untapped potential. For example, the growth of information technology has resulted in the spread of massive data in our daily lives. Many real-world applications often create imbalanced datasets for critical classification tasks. For example, to anticipate click-through rates, online advertising companies may produce many datasets, such as user viewing or interactions with advertisements

**Research object :** This research focuses on the challenges associated with learning classifiers from large-scale, highly imbalanced datasets prevalent in many real-world applications. Traditional algorithms learning often need better performance and high computational efficiency when dealing with imbalanced data. Factors such as class imbalance, noise, and class overlap make it demanding to learn effective classifiers.

**Methods:** The self-paced ensemble method addresses the challenges of high imbalance ratios, class overlap, and noise presence in large-scale imbalanced classification problems. By incorporating the knowledge of these challenges into our learning framework, we establish the concept of classification hardness distribution

**Conclusion:** This research concludes that the self-paced ensemble is a revolutionary learning paradigm for massive imbalance categorization, capable of improving the performance of existing learning algorithms on imbalanced data and providing better results for future applications.

## 1. Introduction

The research focuses on self-paced ensemble and big data classifications, with considerable data imbalance as a mediating factor. This idea is a brand-new domain with a lot of untapped potential. For example, the growth of information technology has resulted in the spread of massive data in our daily lives. Many real-world applications often create imbalanced datasets for critical classification tasks. For example, to anticipate click-through rates, online advertising companies may produce many datasets, such as user viewing or interactions with advertisements [1–3].

Massive data categorization and its integration with the information technology industry are evolving rapidly and employing algorithms. Predictions are two of the many ways and strategies possible for credit fraud detection (CFD) in the current problematic situation. Dal Pozzolo, Boracchi [4] discussed that one of the greatest testbeds for computational intelligence algorithms may be identifying fraud in credit card transactions. Since consumers' behaviors vary over time and fraudsters alter their tactics, this issue

includes a variety of pertinent difficulties, including idea drift, class imbalance, and verification delay. According to many writers, it is based on a dataset containing many genuine credits card transactions, of which only a tiny fraction is fraudulent. Medical diagnosis, record linkage, and network intrusion detection, among other operations, have similar circumstances and activities [1–3]. Real-world datasets are also likely to have extra complexity, like noise and missing data. The severely unbalanced, large-scale, and noisy data make the downstream classification tasks more difficult, which is one of the difficulties to overcome in this paper.

In today's world, unbalanced datasets and conventional categorization approaches for data processing are no longer helpful. Traditional classification approaches do not perform well in modelling and predicting imbalanced datasets due to outdated algorithm restrictions like C4.5, SVM, or Neural Networks, which perform poorly [1, 5, 6]. When the dataset is extensive and noisy, the problem becomes much more challenging. Because of the large number of majority instances, the minority class is often neglected due to the erroneous assumption that positive and negative samples are distributed equally. Minority classes, on the other hand, typically have beliefs that cover a more comprehensive range of topics than majority classes [1, 7, 8]. From the preceding critical scientific discussions, it is clear that researchers need to investigate self-paced ensemble and big data identification with some significant imbalance classifications, which is one of the most persistent problems in the real world for various large databases and their organizations. On the other hand, Algorithm-level methods focus on altering current learners to lessen their prejudice towards majority groups. It necessitates a thorough understanding of the updated learning algorithm and accurate pinpointing of the causes behind its inability to mine skewed distributions. Cost-sensitive learning is the most often used algorithm-level technique. Reducing bias toward the majority class raises minority relevance throughout the learning process by assigning considerable costs to minority cases and lower costs to mainstream occurrences.

Many scientific research have been suggested to address this problem, which may be divided into three groups. Data-level approaches change the distribution of instances to make them more balanced and eliminate tough cases. Due to their distance-based architecture, they may be inapplicable on datasets having categorical characteristics or missing values [9, 10]. The application of large-scale data have high computational cost (e.g., SMOTE, ADASYN) [11, 12]. Algorithm-level solutions modify current algorithms learning directly to reduce the bias toward majority items. Organizations do need support from domain specialists ahead of time (e.g., setting a cost matrix in cost-sensitive learning) [13, 14]. As a result, such algorithms fail when working with batch-trained classifiers like neural networks since the class distribution on the training data needs to be balanced.

An ensemble classifier is constructed by integrating one of the above processes with an ensemble learning algorithm. SMOTE Bagging, for example, has a substantial training cost and low applicability on practical tasks [15].

Easy Ensemble, Balance Cascade [16], and other models might lead to underfitting or overfitting when the dataset is noisy. A wide range of existing approaches cannot manage the highly unbalanced, large-scale, and noisy classification job that is a prevalent challenge in real-world applications for the reasons

mentioned above and more. Such tasks are challenging because current approaches need to recognize the problems inherent in imbalanced learning. Not only does a class imbalance negatively impact classification performance, but so do other issues, such as the existence of noise samples [17] and an overlapping underlying distribution across classes [18, 19]. The high imbalance ratio might amplify these effects even more. Furthermore, the sensitivity of different models to these variables varies. One must consider all the characteristics mentioned above for more precise categorization.

To address these issues, classification tribulation is fundamental—hardness measures how difficult a classifier is for accurately categorizing a sample. As a result, the job’s difficulty is included in the distribution of categorization hardness. For example, noises are more likely to have higher hardness values, and the fraction or proportion of high-hardness samples represented the extent of class overlap. As a further benefit, the hardness distribution is inherently adaptable to various models, and the hardness distribution may guide the re-sampling approach to improving performance.

In this research, the researchers introduce a novel learning framework based on categorization difficulty termed Self-paced Ensemble (abbreviated as SPE). Instead of simply balancing the positive and negative data or using instance weights, we consider the dataset’s distribution of classification difficulty and iteratively choose the most informative majority of data samples in line with the distribution. A self-paced mechanism controls the under-sampling process. This self-paced procedure enables our system to focus on the more complex data samples while retaining awareness of the straightforward sample distribution to prevent overfitting. Figure 1 depicts the pipeline for the self-paced ensemble.

The following are the main contributions of this research study. For example, we show why traditional imbalance learning techniques fail when applied to a massively imbalanced classification and categorization function. The research study does extensive statistical analysis and visualization trials that benefit other categorization systems comparable to ours. Secondly, we introduced the SPE as a learning framework for massively imbalanced data categorization. Similarly, this SPE is exceptionally computationally efficient and may improve the performance of any canonical classifier (such as C4.5, SVM, GBDT, and Neural Network) on real-world severely unbalanced jobs. SPE is more precise, quick, resilient, and adaptable than traditional approaches. Lastly, the idea of categorization or classification complexity is introduced in this particular research. The learning process of our suggested framework SPE is automatically tuned in a model-specific manner by considering the distribution of classification difficulty across the dataset. In contrast to current approaches, our learning framework does not call for pre-defined distance measures, which are often unavailable in real-world situations but uses the previous data set to know how the third party controlled the imbalanced data.

## **2. Problem Symbols, Definitions and Methods**

The research defines the issue of class imbalance. Then, we proposed some essential symbols and definitions and demonstrated the standard assessment used in imbalanced cases. Class imbalance: a dataset is considered imbalanced if it contains far too few samples of a specific type. When it comes to real-world applications, such as medical, fraud detection, or click-through-rate prediction, there is an

imbalance between the healthy and sick, normal and fraud, clicked and ignored etc. Because of their accuracy-oriented nature, canonical learning techniques tend to favor the majority when applied to a dataset with a skewed distribution. In real applications, class imbalance typically coexists with other problematic factors like large data sets, noise, and missing values. Even though this subject has gathered a lot of attention, existing techniques still do not meet expectations.

An example of a symbol is a practical application where the ratio condition is dominant. The minority class had less samples than the majority class, and this was taken into account while performing ratio imbalance classification. The minority class is always referred to as an additive (+) continuous class, whereas the dominant class is referred to as negative (-) continuous class. All training samples are referred to as F in this context (x, y). Afterwards, the minority class set Q and the majority class set N are established. The Eq. (1) has explained the proposed relationship between two variables.

$$Q = \{(x, y) | y = 1\}, N = \{(x, y) | y = 0\},$$

1

In the case of (highly) imbalanced conditions, we have  $|N| \gg |Q|$ . Accordingly, the imbalance data need ratio in the form of Imbalance Ratio (IR), which is defined as the number of cases from the majority group divided by those from the minority group in each dataset. The below Eq. (2) defines the actual calculation of the IR procedure.

$$\text{ImbalanceRatio (IR)} = \frac{n_{majority}}{n_{minority}} = \frac{N}{Q},$$

2

The empirical study uses alternative assessment criteria, such as the number of accurate/false positive/negative predictions does not adequately represent the model's performance. The study developed modelling based on the structural equation model (SEM) to determine each class's actual forecast. The goodness of model fit is the primary criterion to evaluate the matrix identified adequately in the ratio scenario. The research used classification hardness distribution, self-paced ensemble (SPE), and self-paced under-sampling (hardness harmonized and self-paced factor) to validate the visualized model one and two for the imbalanced data class. The study described dependent, independent, and mediating variables—data collection, analysis, estimating methodologies, multi-dimensional evaluation, bootstrapping, and validation. Lastly, the intervening models were measured through SEM (AMOS), and the simulation model was tested using Python software. The study was interested in the prediction model, which predicts the outcomes using a model-based SEM. Covariance-based SEM is preferred if a researcher wants to confirm a model's theory or data fitness. Covariance-based SEM usually needs specific requirements (assumptions), such as a normal distribution of data, reflective indicators, or minimum numbers of cases [20]. All research includes basic philosophical assumptions about objective reality, and acceptable research methodologies are chosen for knowledge acquisition and machine learning. The study design was a quantitative and statistical simulation based on G-mean and median. The study starts with

theoretical development and then moves to hypothesis formation. On the other hand, Sekaran [21] described that good research starts from theoretical development and moves toward hypotheses formulation to formally concluded with the help of key findings and results.

### 3. Data Using and Analysis

The dataset was taken from the Pakistani banks' records, which have already fallen into credit fraud. Similarly, data analysis began with a data screen, filtering, missing values, frequency distribution, outliers' detection, and normality testing. Again, the fundamental assumptions of unbiased outcomes dictate that data must be normally distributed before analysis. The imbalance data would be checked for normality testing and then used for simulation with the help of phyton.

### 4. Data Analysis

The study used confirmatory measurement factor analysis for the hardness of harmonization; this research experimentally analyzes all the indicators and develops an equation to describe the correlation coefficient among class imbalance, imbalance ratio, credit fraud, self-paced ensemble, factors, and hardness harmonization. According to the measurement model, the classes' efficiency, applicability, and sensitivity are improving, and the model may be used to evaluate the structural equations simulation, as shown in Fig. 1. A second model (the fit model) was used to assess the actual forecasting of the self-paced ensemble, class imbalance, credit fraud, imbalance, self-paced factor, hardness, and harmonization, which have a significant correlation coefficient (see Table 1 second line). The researchers updated the original model and added iterations for each construct with covariate pathways to reach their intended findings. Furthermore, the model measurement achieves statistically significant results.

Path analysis's primary objective is to identify any causal connections between the data hardness and the abovementioned factors. One of the more sophisticated methods for establishing whether or not there is a cause-and-effect relationship between a group of variables is thoroughly analyzed by SEM [22].

Particularly noteworthy was how the article distinguished a casual  $R^2$  link between the original model and model fit. The logic of path analysis is to create a diagram showing the genuine causal flow or the actual direction of cause-and-effect for the future prediction that is associated with arrows, covariates, and exhibits. The path analysis beauty estimates the continuous association between the quantified connection from the direct path to indirect causal effects and forecasts a good model for future problems resolving. Consequently, the route diagram visually represents the theoretical justification of cause-and-effect connections between several variables leading to numerical outputs (ratio and percentages). According to Agresti and Finlay [23], the primary characteristic of path analysis is to create direct and indirect causal effects among the outcome predictors. In the process of obtaining scientific information, the utilization of indirect effects is quite helpful. When a variable impacts an endogenous indicator and its effects on other variables or indicators, this is known as an indirect effect. In the subjective model, it is referred to as an indirect impact and is also known as an intervening indication. On the other hand, SEM was developed, and class imbalance to assess the intervening effect of fostering adoption and correlation

coefficient determination, and the values of Fig. 1 showed that self-paced ensemble, credit fraud, self-paced factor and hardness harmonization are the indication of the data in the real-world activities. The first model's display and model fit are displayed in Table 1. The model fitness was measured with Eq. 3.

## 4.1. Figures and Tables

Table 1

Fit Indices Prediction Models of Self-Paced Ensemble, Class Imbalance, Credit Fraud, Imbalance, Self-Paced Factor, Hardness (n = 80402).

Model	$\chi^2_{df}$	$\chi^2/df$	GFI	CFI	NNFI	RMSEA	SRMR
Initial Model	9.367	1.578	0.9	0.88	0.875	0.071	0.081
Model Fit	2.177	1.743	0.913	0.932	0.89	0.081	0.091
$\Delta\chi^2$	6.624						

Note: N = 80402, All the changes in chi square values are computed relative to model,  $\chi^2 > .05$ , GFI = Goodness of fit index, CFI = comparative fit index, NNFI (TLI) = Non-normed fit index, RMSEA = root mean square error of approximation, SRMR = Standardized root mean square,  $\Delta\chi^2$  = chi square change.

$$C(\alpha, \alpha) = [N - r] \left[ \sum_{g=1}^G \frac{(N)^g f(\mu^g, \Sigma_{g, x^{(g)}, S^{(g)}})}{N} \right] = [N - r] F(\alpha, \alpha)$$

$$fkl(\mu^g \Sigma^{(g)} x^{(g)} S^{(g)}) = \log \left[ \sum_{g=1}^G \right] + tr(S^{(g)} \Sigma^{(g-1)} + (x^{(g)} - \mu^g) \Sigma^{(g-1)} (x^{(g)} - \mu^g)).$$

$$c = (N^1 - 1)F^{(1)} = (N - 1)F.$$

$$C = \sum_{g=1}^{(G)} N^{(g)} F^{(g)} = FN. \quad (3)$$

$$(D1) CMIN Initial Model = 9.367$$

$$CMIN Model Fit = 2.624$$

$$\Delta\chi^2 = 9.367 - 2.177 = 6.777$$

$$D2 fml(\mu^g \Sigma^{(g)} x^{(g)} S^{(g)}) = fkl(\mu^g \Sigma^{(g)} x^{(g)} S^{(g)}) - fkl(\mu^g \Sigma^{(g)} x^{(g)} S^{(g)})$$

$$= \log \left[ \sum_{g=1}^G \right] + tr(S^{(g)} \Sigma^{(g-1)} + (x^{(g)} - \mu^{(g)}) \Sigma^{(g-1)} (x^{(g)} - \mu^{(g)})).$$

$$CMIN Initial Model = \chi^2/df = 1.578$$

$$CMIN Model fit = \chi^2/df = 1.743$$

The fit indices indicated that hardness hominization and self-paced ensemble are those variables that are absolutely fit, and the models are proved to be significant on each iteration. Similarly, the chi-square test of absolute model fit is sensitive to under-sample size and number of parameters. The equation was mathematically measured with absolute and relative fit (see Eq. 4).



$$GFI = 1 - \frac{\hat{F}}{\hat{F}_b}$$

$$f \left( \Sigma(g), s^{(g)} \right) = \frac{1}{2} \text{tr} \left[ K^{(g-1)} \left( x^{(g)} - \Sigma(g-1) \right) \right] 2.$$

Model fit value of GFI = .913

$$CFI = 1 - \frac{\max(\hat{C} - d, 0)}{\max(\hat{C}_b - d_b, 0)} = 1 - \frac{NCP}{NC P_b}$$

$$RNI = 1 - \frac{\hat{C} - d}{\hat{C}_b - d_b}$$

Model fit value of CFI = .932

$$TLI = 1 - \frac{\frac{\hat{C}_b}{d_b} - \frac{\hat{C}}{d}}{\frac{\hat{C}_b}{d_b} - 1}$$

Model fit value of TLI = .810

$$SRMR = \sqrt{\sum_{g=1}^G \left\{ \sum_{i=1}^{pR} \sum_{j=1}^{j \leq i} \left( \frac{\sigma^{(gij)} - \sigma^{(gij)}}{\sigma^{(gij)}} \right)^2 \right\} / \sum_{g=1}^G p * (g)}.$$

Model fit value of SRMR = .091

$$\text{Population RMSEA} = \sqrt{\frac{F}{d}}$$

$$\text{Estimated RMSEA} = \sqrt{\frac{F}{d}}$$

$$LO 90 = \sqrt{\frac{\delta L / n}{d}}$$

$$HI 90 = \sqrt{\frac{\delta U / n}{d}}$$

$$RMSEA = .081$$

Hu and Bentler [24] suggested cutoff point for  $\chi^2/df$ , which is between 1 and 3. Likewise, RMSEA and SRMR values should be less than .08 and CFI, TLI or NNFI and GFI values greater than .9 and if it is less than from  $.9 \leq .8$  then this value is significant. The measurement model values are depicted in Table 1. however, the Tomás, Meliá [25] reported that the covariance correlation coefficient between study variable should legitimately has a good variance. The criteria of modification indices measure should at least 4.0 for error covariance [26]. In nutshell, the study proved that covariance “chi-square Chang” was greater than 4.0 and the study results was 6.624 thorough in modification process. Figure 3 also defined the path coefficient correlation, and it was significant because P-values were less than ( $p < .05$ ). In this regard, the correlation coefficient was depicting, and the inter-correlation direction was significant. The result revealed that self-paced ensemble, class imbalance, credit fraud, imbalance, self-paced factor and hardness harmonization have a coefficient correlation with each other, which could control the data credit fraud in the real world.

Moreover, the findings of this experimental investigation on synthetic actual unbalanced datasets are briefly explained in this section. The study proved that suggested approach’s applicability for combining several base classifier types and then measured through python also. The study also provides some

graphs to explain further how our suggested technique differs from the previous for the imbalance learning method. We used a variety of criteria to analyze the experiment's findings and show how effective our suggested framework is.

On the other hand, the study trained the model with Python software. Similarly, the classification hardness distribution model was introduced. In this part, the study discusses the notion of "classification difficulty" before we present our approach. The study scientifically measure the advantages of including hardness distribution in a system for imbalance learning. To better comprehend the connection between hardness, imbalance ratio, class overlapping, and model capacity, we also give an understandable representation in Fig. 2.

Based on prior findings, we set out to create an under-sampling method that minimizes the impact of noise and trivial samples while increasing the significance of borderline samples in the manner we anticipated. For example, "hardness harmonization" is the process of dividing many samples into "k" bins based on their hardness values, where "k" is a hyperparameter. Each receptacle denotes a specific hardness grade. Most occurrences are then under-sampled into a balanced dataset while maintaining the same overall hardness contribution across all bins. This technique is referred to as harmonize for gradient-based optimization.

The study found that most occurrences are under-sampled into a balanced dataset while maintaining the same overall hardness contribution across all plots. This technique, which harmonizes the gradient contribution in batch training of neural networks, is known as "harmonize" in the gradient-based optimization suggested by [27]. In this practical situation, research use a similar approach to balance the difficulty in the first iteration.

The hardness harmonization is only used in some versions. The fundamental cause is that when the ensemble classifier increasingly fits the training set, the number of trivial samples increases throughout training. As a result, just balancing the hardness contribution still results in many pointless samples (Fig. 2). Since these samples are less informative, they significantly slow down learning during subsequent rounds. Instead, the research uses self-paced harmonizing under-sampling by introducing the concerning self-pace factor. For the model, the research study begins by balancing the hardness contributions of each bin and then progressively reduces the sample probability of those bins with a high population. A self-paced component controls the lowering level. When it increases, we pay more attention to the more challenging samples rather than the straightforward hardness contribution harmonizing. The study methodology primarily concentrates on those necessary borderline samples in the first iterations; therefore, outliers and noise impact the study model's capacity to generalize less. The research methodology maintains a decent percentage of high-confidence samples as the skeleton in the latter rounds when it is pretty giant, thereby preventing our framework from overfitting.

Table 2  
Performance Generalization for KNN and Ada Boost

Model	Hyper	RandUnder	Clean	Smote	Easy10	Cascade10	SPE10
KNN	K_neighbors = 5	0.291 ± 0.002	0.392 ± 0.000	0.260 ± 0.003	0.422 ± 0.003	0.412 ± 0.004	0.489 ± 0.005
AdaBoost	n_estimator = 10	0.232 ± 0.017	0.358 ± 0.000	0.298 ± 0.005	0.478 ± 0.021	0.389 ± 0.012	0.580 ± 0.007

## 5. Synthetic Dataset

Empirical study initially demonstrates the experimental findings on the fabricated dataset to provide more insights into our system. To test our strategy, next, construct a 3D checkerboard dataset. There are Gaussian components in the dataset. All Gaussian components share the covariance matrix. The study chose to use 900 samples of the data for the minorities and 9000 samples for the majority number. Independent samples from the same original distribution were taken for the training, validation, and test the sets. The actual dataset illustration was shown in Fig. 3.

The results in Table 2 revealed that ten independent runs were performed based on mean as well as standard deviation to reduce randomness based on the checkerboard. The researcher applied the criteria of hyper-parameters for each base classifier and observed the reduced randomness. Nevertheless, this particular study focuses on the two simulation canonical classifiers such as “K-Nearest Neighbors” (KNN) and “Adaptive Boosting” (AdaBoost). These two classifiers are the best to measure the effectiveness and applicability of distinct imbalanced learning data. As a result, the overall measure brings good performance and effectiveness in the different imbalanced learning methods.

## 6. Synthetic Dataset Results with Generalized Performance

It is imperative to run the results lists on the checkerboard performance. Notably, the current study provides the mean and standard deviation of six (6) separate runs to eliminate randomness. The hyper-parameters used for each base classifier are also included. Figure 4 reveals the three distinct iterations based on (0.05, 0.10, 0.15), such as SPE regularly outperforms competing techniques on the checkerboard dataset. Likewise, distance-based resampling produced subpar results when working with specific classifiers; for example, the cascade curve line is excellent on the 0.05 level compared to 0.10 and 0.15. The study's primary factor leads to the validity of such resample strategies and to recognize differences in model capability. Comparatively, it was revealed from the results that the Cascade ensemble performs better and more robustly, and the suggested ensemble framework SPE has created a hardness distribution over the experimental data in Fig. 4.

The study evaluated each approach using different base models to see how the number of base models affects the performance of ensemble methods on the real-world dataset in Fig. 5. Over-sampling processes

need more data and resources to train each base model, so this comparison needs to be revised. All ensemble approaches were applied to the credit fraud dataset with the help of logistic regression, geometric mean, and skewness for assessment, considering the computational cost generated by the vast number of synthetic samples on the large-scale highly imbalanced dataset with the help of 80402, 80450 and 80410 samples see in Table 4. We also list the total number of data samples used for training each method's ensemble's base models. The experimental outcomes for our suggested technique and five ensemble methods are shown in Table 4.

Table 3  
Statistical Results of the Imbalance Ratio.

Dataset	Attributes	Sample	Data Feature	ImbalanceRatio	Model
Credit Fraud	63	80402	Numerical & Linear	823.32:1	AdaBoost <sub>10</sub>
Record Linkage	53	80450	Numerical & Linear	981.52:1	AdaBoost <sub>10</sub>
Payment Simulation	49	80410	Numerical & Linear	801.12:1	AdaBoost <sub>10</sub>

Table 4  
.Generalized performance on three world datasets.

Dataset	Model	RandUnder	Clean	SMOTE	Easy	Cascade	SPE10	Sample
Credit Fraud	Simulation AdaBoost	0.061 ± 0.004	0.682 ± 0.000	0.413 ± 0.000	0.159 ± 0.013	0.599 ± 0.013	0.841 ± 0.019	80402
		0.121 ± 0.008	0.911 ± 0.000	0.688 ± 0.001	0.360 ± 0.020	0.881 ± 0.034	0.844 ± 0.016	80402
		0.358 ± 0.010	0.718 ± 0.000	0.689 ± 0.001	0.399 ± 0.027	0.882 ± 0.002	0.863 ± 0.005	80402
Record Linkage	Simulation AdaBoost	0.110 ± 0.022	0.050 ± 0.000	0.712 ± 0.070	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	80450
Payment Simulation	Simulation AdaBoost	0.312 ± 0.040	0.313 ± 0.030	0.743 ± 0.050	0.681 ± 0.049	0.801 ± 0.005	0.100 ± 0.001	80410

The Materials and Methods should be described with sufficient details to allow others to replicate and build on the published results. Please note that the publication of your manuscript implies that you must make all materials. Moreover, SPE employs a relatively similar number of training data as the other three under-sampling-based ensemble strategies but surpasses them considerably on all assessment criteria. When compared to two ensemble approaches that rely on oversampling, SPE exhibits comparable performance while using much less training data. Methods based on oversampling should be significantly more sampled. Under conditions of significant imbalance ratio, they produce many synthetic samples in

Table 3. This increases the training set's size and needs much more processing power to train each base model. A more extensive data set and a higher imbalance ratio might worsen the problem. The study ran more thorough tests on the credit fraud and payment simulation datasets, as seen in Figure .6. Furthermore, in Fig. 6, we used different methods with different colors, and each color represents SPE methods. The results demonstrated that robustness on the two real-world data has different hardens levels: various "k" and "H" selections. Remember that "k" impacts how accurate our estimate of the hardness distribution is; thus, choosing a low value for "k" could result in poor performance. Likewise, the study revealed that SPE only employs a small amount of data during training. Still, it provides the desired result and is superior to over-sampling-based techniques. Additionally, SPE performs consistently throughout numerous independent runs on both workloads and controls the credit fraud ratio. For instance, other methods lack stability and have more variability as compared to SPE.

## 7. Discussion

Much of the research presented in this work can apply to any imbalanced data parameters, and different researchers argued that a fundamental issue has overwhelmed machine learning with imbalanced data categorization [28, 29]. The point of imbalance learning was addressed in the current research. Dong and Qian [30] recently published a systematic review of the approaches and practical applications in imbalance learning. Most proposed works [31, 32] used distance-based techniques to obtain resampled data for training canonical classifiers. Based on them, numerous works [33, 34] combine resampling with ensemble learning. The present study showed that imbalanced learning tactics have to be very successful for artificial intelligence. Methods based on distance have several drawbacks. First, spread on a real-world dataset is challenging to define, especially when it has categorical or continuous features or missing values. Second, computing the distances between each sample can be extremely expensive when applied to large-scale datasets. Due to their model-agnostic designs, distance-based methods, despite being successfully used for resampling, do not ensure good performance for various classifiers; as an alternative to resampling the entire dataset, some other techniques attempt to give different weights to samples. Ding, Chen [35] performed the need and assistance of subject matter experts for the risk failing when using batch training techniques, for instance, the neural approach. Because prior experiments [36, 37] have shown that setting arbitrary costs without domain knowledge prevents them from performing to their fullest potential of credit card fraud, this study recommends methods in this paper that have both employee perception and dataset use.

Several studies in other fields borrow the concept of choosing "informative" samples but concentrate on entirely different issues (e.g., active learning [38, 39] and self-paced learning [40]). The study illustrates one of many possible applications of a self-paced learning algorithm that tries to present the training data in a meaningful order that promotes learning. At the same time, an active learner interacts with the user to obtain the labels of new data points. However, the study carries out the sampling without taking the overall data distribution into account, their process of fine-tuning is easily upset when the training set is unbalanced. In contrast, SPE balances the dataset using an under-sampling plus ensemble strategy that works with any canonical classifier. Instead of randomly choosing "informative" data samples, SPE

performs adaptive and robust under-sampling by considering the dynamic hardness distribution across the entire dataset. As a result, traditional distance-based resampling techniques may inadequately work with classifiers because they fail to account for the variation in model capacity. They are computationally inefficient, especially for large datasets, because it takes additional work to calculate the distances between samples. Additionally, since real-world datasets may include categorical features and missing values, it is frequently challenging to establish a precise distance metric in practice. Most ensemble-based methods still suffer from the above issues because they incorporate distance-based resampling into their pipelines. Compared to previous works, SPE is simpler to apply and more computationally efficient because it does not call for any pre-established distance metrics or computations. SPE is adaptable to various models and robust to noises and missing values because it self-paced harmonizes the hardness distribution concerning the given classifier.

## Conclusion

The outcome of various experiments led to the conclusion that the issue of highly imbalanced, massive, and noisy data categorization often occurs in real-world applications, which has become a problem for highly imbalanced data classification. In summary, this technique has demonstrated high-quality results for the imbalanced data cases; we have shown that conventional machine learning or imbalance learning algorithms have poor computing efficiency and provide unacceptable outcomes for such real-world data. This research proposes the self-paced ensemble, which is one of the revolutionary learning paradigms, especially for the high imbalance categorization or classification, and this study proved that the self-paced ensemble has an innovative learning framework and improves the information. This experimental analysis leads to valuable conclusions; most importantly, the current study introduces the challenge of high imbalance ratio, class overlap, and noise presence, which are vital for the vast imbalance classification and for solving with different statistical and algorithmic techniques. The study incorporates the knowledge of these challenges into the learning framework. The study established the idea of hardness harmonization asymmetrical distribution, which the structural modelling technique proved. We performed in-depth experimental tasks on a range of challenging real-world data activities and tested different sample sizes and methods for bringing improvement. Our mixed method framework performs better, has a more extensive range of applications, and is more computationally efficient than previous approaches, which other scientists do. In conclusion, this study demonstrated that introducing an advanced paradigm could be a predictor for including task challenges in classifying imbalanced datasets and producing hardness harmonization data to control credit fraud. The contribution of this paper has several real-world applications for imbalanced datasets, and it is a straightforward technique to make hardness to manage credit fraud in the practical world.

## Declarations

Data Availability Statement: The corresponding author can provide the dataset used for the experiments in this study upon reasonable request.

Funding: This research received no external funding.

Competing interests: The authors declare no conflict of interests

Consent for publication: Not Applicable.

Declarations Ethics approval and consent to participate: Not Applicable.

## References

1. Liu, Z., et al., *Towards Inter-class and Intra-class Imbalance in Class-imbalanced Learning*. arXiv preprint arXiv:2111.12791, 2021: p. 1-14.
2. Ding, R., et al., *Semi-supervised Optimal Transport with Self-paced Ensemble for Cross-hospital Sepsis Early Detection*. arXiv preprint arXiv:2106.10352, 2021: p. 1-14.
3. Ristea, N.-C. and R.T. Ionescu, *Self-paced ensemble learning for speech and audio classification*. arXiv preprint arXiv:2103.11988, 2021. **v1**: p. 1-5.
4. Dal Pozzolo, A., et al., *Credit card fraud detection: a realistic modeling and a novel learning strategy*. IEEE Transactions on Neural Networks and Learning Systems, 2018. **29**(8): p. 3784-3797.
5. Quinlan, J.R., *Induction of Decision Trees*. Machine Learning, 1986. **1**(1): p. 81-106.
6. Cortes, C. and V. Vapnik, *Support-vector Networks*. Machine Learning, 1995. **20**(3): p. 273-297.
7. He, H. and E.A. Garcia, *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 2009. **21**(9): p. 1263-1284.
8. Chen, S. and H. He, *Nonstationary stream data learning with imbalanced class distribution*. Imbalanced Learning: Foundations, Algorithms, and Applications. 2013. 151-186.
9. Tomek, I., *Two Modifications of CNN*. IEEE Trans. Systems, Man and Cybernetics,, 1976. **6**(11): p. 769–772.
10. Mani, I. and I. Zhang. *kNN approach to unbalanced data distributions: a case study involving information extraction*. in *Proceedings of workshop on learning from imbalanced datasets*. 2003. ICML.
11. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 2002. **16**: p. 321-357.
12. He, H., et al. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. 2008. IEEE.
13. Elkan, C. *The foundations of cost-sensitive learning*. in *International joint conference on artificial intelligence*. 2001. Lawrence Erlbaum Associates Ltd.
14. Liu, X.-Y. and Z.-H. Zhou. *The influence of class imbalance on cost-sensitive learning: An empirical study*. in *Sixth International Conference on Data Mining (ICDM'06)*. 2006. IEEE.

15. Wang, S. and X. Yao. *Diversity analysis on imbalanced data sets by using ensemble models*. in *2009 IEEE symposium on computational intelligence and data mining*. 2009. IEEE.
16. Liu, X.-Y., J. Wu, and Z.-H. Zhou, *Exploratory undersampling for class-imbalance learning*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008. **39**(2): p. 539-550.
17. Napierała, K., J. Stefanowski, and S. Wilk. *Learning from imbalanced data in presence of noisy and borderline examples*. in *International conference on rough sets and current trends in computing*. 2010. Springer.
18. García, V., J. Sánchez, and R. Mollineda. *An empirical study of the behavior of classifiers on imbalanced and overlapped data sets*. in *Iberoamerican congress on pattern recognition*. 2007. Springer.
19. Prati, R.C., G.E. Batista, and M.C. Monard. *Learning with class skews and small disjuncts*. in *Brazilian Symposium on Artificial Intelligence*. 2004. Springer.
20. Hair, J.F., M. Gabriel, and V. Patel, *AMOS covariance-based structural equation modeling (CB-SEM): Guidelines on its application as a marketing research tool*. Brazilian Journal of Marketing, 2014. **13**(2).
21. Sekaran, U., *Research Method for Business: A Skill Approach*; John Willey and Sons. Inc. New York, 2006.
22. Hair, J.F., M. Gabriel, and V. Patel, *AMOS covariance-based structural equation modeling (CB-SEM): Guidelines on its application as a marketing research tool*. Brazilian Journal of Marketing, 2014. **13**(2): p. 1-12.
23. Agresti, A. and B. Finlay, *Statistical models for the social sciences*. Upper Saddle River, NJ: Prentice-Hall. Revascularization Procedures after Coronary Angiography." *Journal of the American Medical Association*, 1997. **269**: p. 2642-46.
24. Hu, L.t. and P.M. Bentler, *Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives*. *Structural equation modeling: a multidisciplinary journal*, 1999. **6**(1): p. 1-55.
25. Tomás, J.M., J.L. Meliá, and A. Oliver, *A cross-validation of a structural equation model of accidents: organizational and psychological variables as predictors of work safety*. *Work & Stress*, 1999. **13**(1): p. 49-58.
26. Byrne, B.M., *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. 2016: Routledge.
27. Li, B., Y. Liu, and X. Wang. *Gradient harmonized single-stage detector*. in *Proceedings of the AAAI conference on artificial intelligence*. 2019.
28. Czarnowski, I., *Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams*. *Journal of Computational Science*, 2022. **61**: p. 101614.
29. Zhai, J., J. Qi, and S. Zhang, *Imbalanced data classification based on diverse sample generation and classifier fusion*. *International Journal of Machine Learning and Cybernetics*, 2022. **13**(3): p. 735-750.



30. Dong, J. and Q. Qian, *A Density-Based Random Forest for Imbalanced Data Classification*. Future Internet, 2022. **14**(3): p. 90.
31. Dai, W., et al., *Deep learning approach for defective spot welds classification using small and class-imbalanced datasets*. Neurocomputing, 2022. **477**: p. 46-60.
32. Wang, Z., et al., *Geometric imbalanced deep learning with feature scaling and boundary sample mining*. Pattern Recognition, 2022. **126**: p. 108564.
33. Kimura, T., *Customer Churn Prediction With Hybrid Resampling And Ensemble Learning* Journal of Management Information & Decision Sciences, 2022. **25**(1): p. 1-23.
34. Shi, H., et al., *Resampling algorithms based on sample concatenation for imbalance learning*. Knowledge-Based Systems, 2022. **245**: p. 108592.
35. Ding, H., et al., *Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection*. Future Generation Computer Systems, 2022. **131**: p. 240-254.
36. Singh, A., R.K. Ranjan, and A. Tiwari, *Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms*. Journal of Experimental & Theoretical Artificial Intelligence, 2022. **34**(4): p. 571-598.
37. Liu, F. and Q. Qian, *Cost-Sensitive Variational Autoencoding Classifier for Imbalanced Data Classification*. Algorithms, 2022. **15**(5): p. 139.
38. Wan, L., C. Dong, and X. Pei, *Self-paced learning-based multi-graphs semi-supervised learning*. Multimedia Tools and Applications, 2022. **81**(5): p. 7025-7046.
39. Bengar, J.Z., et al. *Class-Balanced Active Learning for Image Classification*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.
40. Liu, B., et al., *A new self-paced learning method for privilege-based positive and unlabeled learning*. Information Sciences, 2022. **609**: p. 996-1009.

## Figures

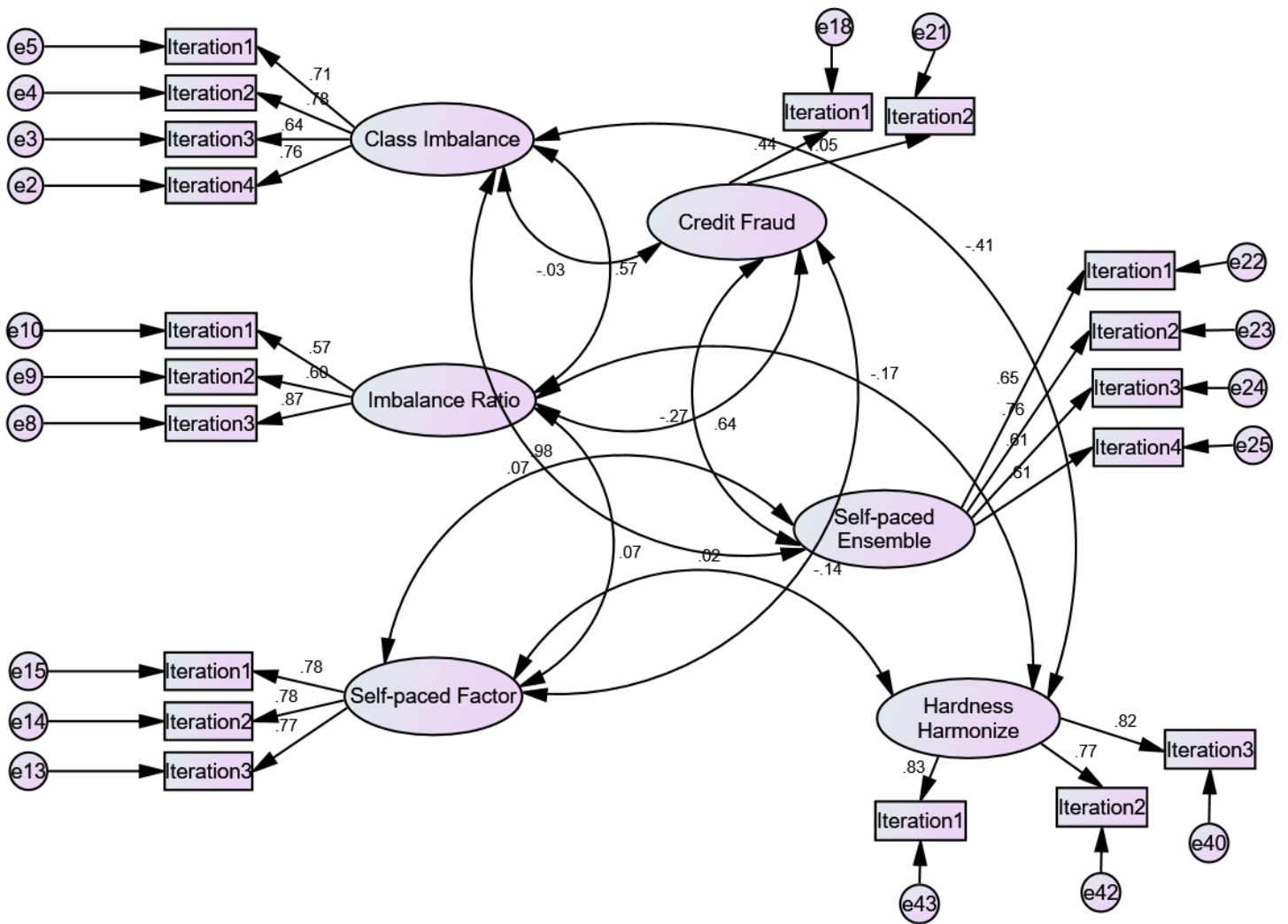


Figure 1

Measurement of the Self-paced ensemble and Hardness of Harmonization (n=80402).

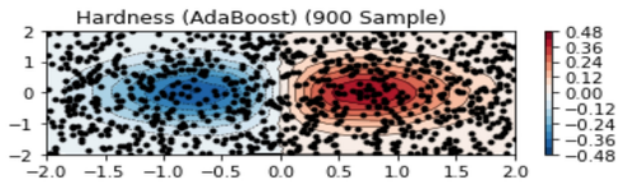
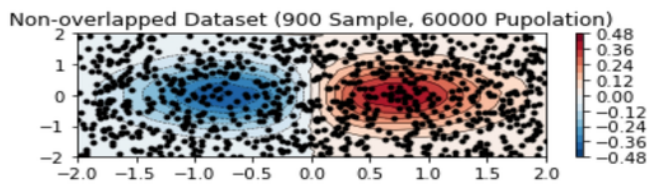
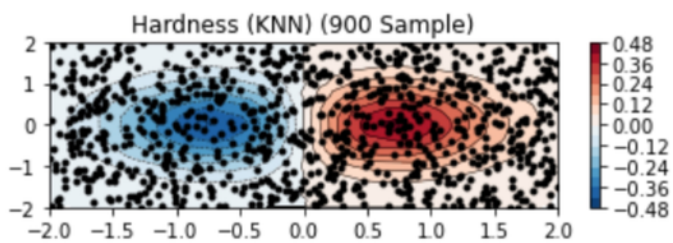
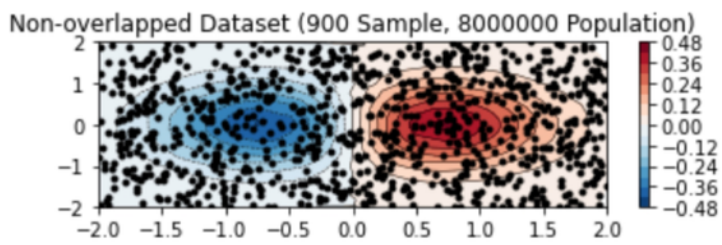
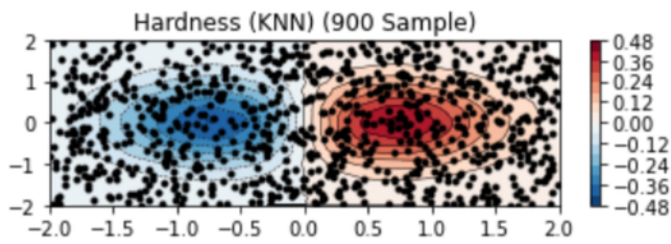
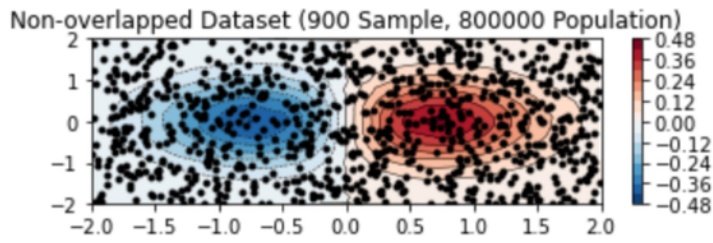


Figure 2

Performance Generalization for KNN and Ada Boost

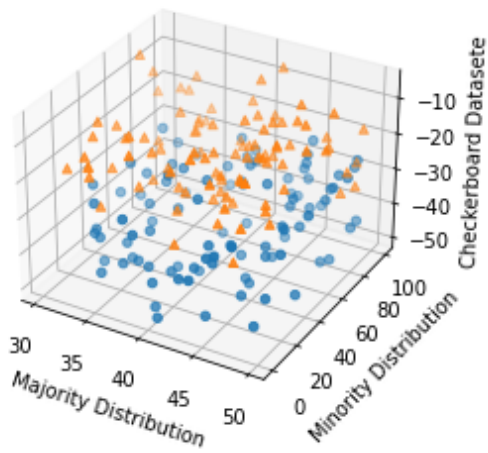


Figure 3

Original distribution of training data set.

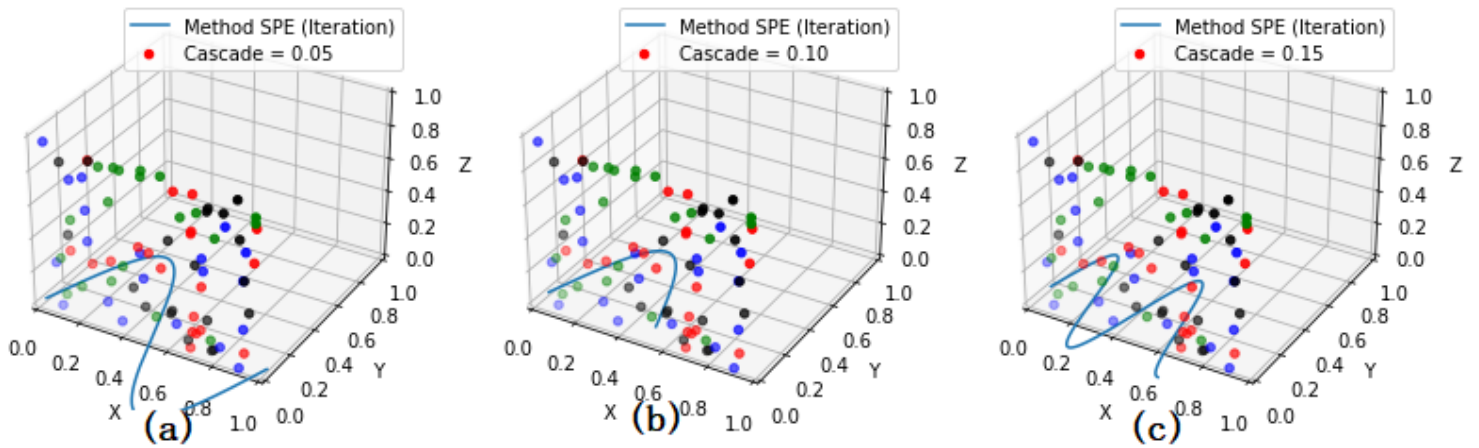


Figure 4

Overlapping training curve.

Generalized Performance on Real World Datasets

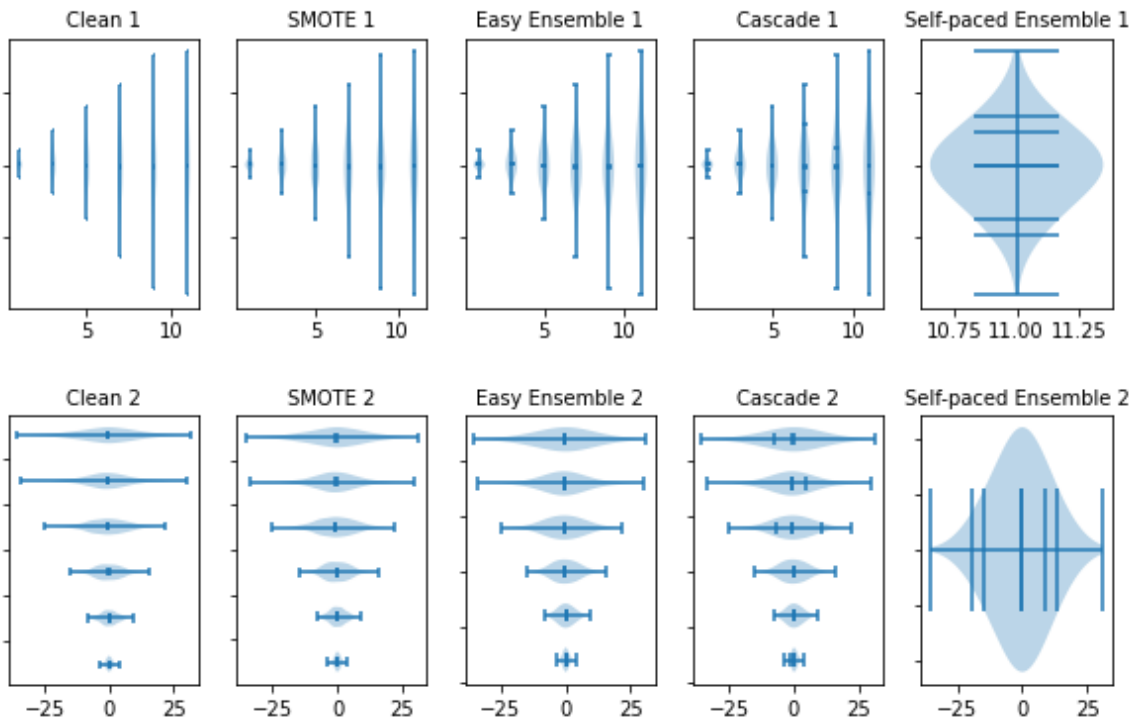


Figure 5

Real world dataset.

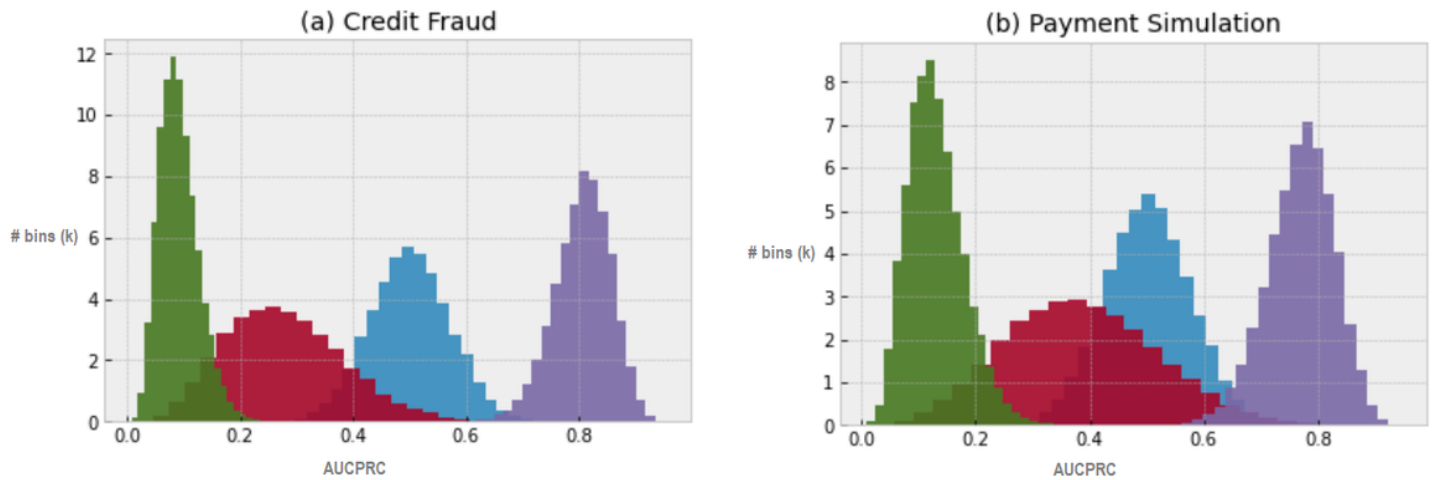


Figure 6

Base Classifier.