

Article

Fourier-Based Adaptive Spectral Synthesis: Decision-Making with Imbalanced Management Data

Firuz Kamalov ^{1,*}, Ahmed El Sayed ¹, Ikhlaas Gurrib ² , Kweh Qian Long ² , Ji Yeh Choi ²  and Ghassan Malkawi ^{3,*}

¹ School of Engineering, Canadian University Dubai, Dubai 118871, United Arab Emirates; ahmed.elsayed@tud.ac.ae

² School of Management, Canadian University Dubai, Dubai 118871, United Arab Emirates; ikhlaas@tud.ac.ae (I.G.); qkweh@tud.ac.ae (K.Q.L.); jiyeh.choi@tud.ac.ae (J.Y.C.)

³ Faculty of Computer Information, Higher Colleges of Technology, Al Ain 21342, United Arab Emirates

* Correspondence: firuz@tud.ac.ae (F.K.); gmalkawi@hct.ac.ae (G.W.)

Abstract

Artificial intelligence applications in management, such as fraud detection and churn prediction, are frequently constrained by the class imbalance problem. Standard over-sampling methods, such as SMOTE, rely on local geometric interpolation, which assumes data convexity and struggles to model the disjoint structures typical of managerial datasets. We introduce Fourier-based Adaptive Spectral Synthesis (FASS), an over-sampling method that frames data generation as a signal reconstruction problem. By transforming the minority class data into the frequency domain via the empirical characteristic function, FASS isolates the global manifold structure from high-frequency sampling noise through an automated spectral filtering mechanism. We prove the L^2 -consistency of the underlying estimator. Empirical evaluations on credit and marketing datasets demonstrate that FASS favorably shifts the precision–recall trade-off compared to geometric baselines, reducing false positives and providing a theoretically consistent, parameter-free approach to learning from imbalanced data.

Keywords: class imbalance; over-sampling; spectral methods; Fourier analysis; empirical characteristic function; SMOTE; fraud detection; imbalanced learning

1. Introduction

The integration of artificial intelligence (AI) into management practices has altered how organizations approach decision-making. From algorithmic trading and credit scoring to supply chain optimization and customer relationship management, machine learning models are employed to identify patterns that drive strategic value. However, the efficacy of these AI systems is bottlenecked by the quality of the underlying data. A common issue is the class imbalance problem, where the event of interest—such as a fraudulent transaction, a manufacturing defect, or customer churn—is statistically rare compared to the majority of cases [1].

In a managerial context, the cost of misclassifying rare events is asymmetric. For instance, failing to identify a single case of financial fraud carries a higher financial and reputational risk than flagging a legitimate transaction for review. Standard supervised learning algorithms, driven by global accuracy metrics, bias their predictions toward the majority class, misclassifying the minority instances that management aims to target.

To mitigate this, researchers typically employ data-level pre-processing techniques, notably the synthetic minority over-sampling technique (SMOTE) [2]. While SMOTE



Academic Editor: Andrea Loddo

Received: 4 June 2026

Revised: 29 June 2026

Accepted: 3 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

serves as a benchmark, its reliance on local geometric linearity limits its effectiveness in non-convex distributions. SMOTE generates synthetic examples by interpolating between nearest neighbors, which assumes that the minority class forms a locally convex set [3]. However, managerial data often exhibits non-convex structures with disjoint clusters, representing distinct types of target behavior. In such scenarios, linear interpolation can generate noisy samples in the voids between clusters, confusing the decision boundary and degrading classifier performance.

In this study, we introduce a global, signal-processing approach to data generation. We propose Fourier-based Adaptive Spectral Synthesis (FASS), an over-sampling method that treats the minority class data as a discrete signal sampled from an unknown probability distribution. By transforming the data into the frequency domain via the empirical characteristic function, FASS distinguishes between the structural signal of the data manifold and the high-frequency noise associated with sparse sampling.

FASS employs an adaptive bandwidth selection mechanism based on the spectral power decay of the data. This ensures that the synthesized data aligns with the global topology of the minority class and provides an accurate representation of rare events.

The contributions of this paper are threefold:

- We introduce a spectral filtering framework for imbalanced learning that addresses the limitations of geometric interpolation.
- We propose an automated, data-driven method for bandwidth selection that avoids manual parameter tuning.
- We provide a mathematical proof of the spectral consistency of our estimator, showing that FASS converges to the underlying probability density of the minority class.

The remainder of this paper is organized as follows. Section 2 surveys the literature on class imbalance and data synthesis. Section 3 establishes the mathematical preliminaries and introduces the FASS framework. Section 4 provides the proofs of the estimator's spectral consistency. Section 5 evaluates the method through experiments on synthetic and managerial datasets. Section 6 discusses implications and limitations. Section 7 concludes.

2. Literature Review

2.1. Class Imbalance in Machine Learning

Class imbalance occurs in business analytics applications such as fraud detection, customer churn prediction, and defect monitoring. Standard learning algorithms are designed under the assumption of balanced class distributions and produce biased models that favor the majority class [4,5]. Systematic reviews indicate that this bias remains a persistent challenge in deploying AI in managerial environments [6,7].

Surveys document the effects of class imbalance on predictive performance and identify data-level, algorithm-level, and hybrid strategies as primary mitigation approaches [1]. Data-level methods, particularly resampling techniques, are widely adopted due to their compatibility with standard learning algorithms [8,9]. However, preserving the structural properties of minority class distributions remains an open challenge.

Among data-level approaches, SMOTE remains the standard benchmark method. SMOTE generates synthetic samples by performing linear interpolation between minority class neighbors [2,10,11]. While SMOTE demonstrates improvements in classification performance, it relies on the assumption that minority class samples form locally convex structures. To address this, contemporary research has introduced geometric and ranking-based modifications to SMOTE [12], as well as adaptive weighting schemes designed to protect the minority boundary [13,14].

Subsequent research has incorporated clustering, density awareness, and adaptive sampling mechanisms. Clustering-based approaches, such as k-means SMOTE, partition

minority data into subregions before generating synthetic samples, handling heterogeneous distributions [15]. Generative adversarial networks have been employed to produce synthetic data through adversarial training, capturing nonlinear relationships beyond local interpolation heuristics [16,17]. Recent work has adopted diffusion probabilistic models for tabular data generation [18]. While diffusion models increase sample fidelity [19], they require computational resources and hyperparameter tuning, limiting their applicability in certain managerial contexts. Preprocessing approaches such as feature selection have also been employed as an alternative method of combating class imbalance issues [20].

Class imbalance affects representation learning in deep neural networks, leading to biased feature extraction and degraded minority class recognition [21]. Surveys highlight the integration of resampling with deep architectures as a mitigation strategy [22]. In the context of financial fraud and credit risk, studies have coupled resampling techniques with graph neural networks [23] and attention-based transformer models [24,25] to capture transactional topology [26].

Beyond geometric interpolation and deep generative modeling, density estimation provides a probabilistic framework for modeling minority class distributions [27–29]. However, standard density estimation techniques often over-smooth minority class manifolds or fail to capture multimodal structures in sparse datasets. Approaches in spectral clustering [30] and manifold learning [31] attempt to mitigate these estimation issues, yet they often lack closed-form generative capabilities.

2.2. Research Gap

Despite the availability of existing over-sampling methods, most techniques rely on local geometric interpolation or computationally intensive generative modeling. The literature indicates a need for methods that combine global distribution modeling with computational efficiency and minimal hyperparameter tuning. The proposed FASS framework employs spectral density estimation derived from empirical characteristic functions to generate synthetic minority samples while preserving global data topology.

3. Fourier-Based Adaptive Spectral Synthesis

3.1. Mathematical Background

Let $\mathcal{D} = \{\mathbf{x}_j\}_{j=1}^N \subset \mathbb{R}^d$ denote the observations belonging to the minority class. We assume these points are independent and identically distributed realizations of a random variable $\mathbf{X}$ governed by an unknown continuous probability density function $f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$.

The characteristic function $\varphi(\mathbf{t})$ uniquely defines the probability distribution and corresponds to the Fourier transform of the density $f(\mathbf{x})$. It is defined as

$$\varphi(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}^T \mathbf{X}}] = \int_{\mathbb{R}^d} e^{i\mathbf{t}^T \mathbf{x}} f(\mathbf{x}) d\mathbf{x}. \quad (1)$$

Since the true density f is unknown, we rely on the empirical characteristic function (ECF) as an unbiased estimator, given by

$$\hat{\varphi}_N(\mathbf{t}) = \frac{1}{N} \sum_{j=1}^N e^{i\mathbf{t}^T \mathbf{x}_j}. \quad (2)$$

Direct inversion of $\hat{\varphi}_N$ yields a sum of Dirac delta functions, representing a singular distribution. To estimate a continuous density $\hat{f}$, we introduce a spectral damping function

$\psi_h(\mathbf{t})$. This function acts as a multiplicative low-pass filter with a resolution parameter $h > 0$. The density estimator is obtained via the inverse Fourier transform:

$$\hat{f}_h(\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it^T \mathbf{x}} \hat{\phi}_N(\mathbf{t}) \psi_h(\mathbf{t}) d\mathbf{t}. \quad (3)$$

3.2. The FASS Algorithm

The FASS framework consists of three phases: spectral filter selection, adaptive bandwidth determination, and generative sampling.

3.2.1. Phase 1: Spectral Filter Selection

To minimize artifacts arising from the Gibbs phenomenon, we employ a Gaussian spectral damping function:

$$\psi_h(\mathbf{t}) = e^{-\frac{1}{2}h^2\|\mathbf{t}\|^2}. \quad (4)$$

Applying the convolution theorem to the inverse Fourier transform yields the spatial density estimator:

$$\hat{f}_h(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^N \frac{1}{(h\sqrt{2\pi})^d} e^{-\frac{\|\mathbf{x}-\mathbf{x}_j\|^2}{2h^2}}. \quad (5)$$

This equation demonstrates that the spatial density estimator $\hat{f}_h(\mathbf{x})$ is mathematically equivalent to a Gaussian kernel density estimator (KDE). Consequently, the data generation process in Phase 3 is equivalent to a smoothed bootstrap. However, unlike standard KDE—which relies on spatial heuristics like Silverman’s rule or computationally expensive cross-validation—FASS determines the bandwidth parameter h analytically in the frequency domain. By doing so, FASS isolates the underlying manifold from sampling noise without relying on spatial assumptions that often over-smooth disjoint or complex minority clusters.

3.2.2. Phase 2: Fourier-Based Bandwidth Selection

Standard density estimation methods frequently rely on cross-validation or spatial heuristics for parameter selection, which may over-smooth complex minority manifolds. FASS selects h by identifying the transition between the structural signal and sampling noise directly in the frequency domain.

First, we compute the power spectrum $P(\mathbf{t}) = |\hat{\phi}_N(\mathbf{t})|^2$. At high frequencies, the structural signal decays, and the summation in the ECF behaves as a random walk in the complex plane. The variance of the ECF scales as $O(1/N)$. Thus, we define a noise floor threshold τ . Next, we determine the cutoff frequency t_{cut} as the smallest frequency magnitude where the radially averaged power spectrum drops below the confidence interval defined by τ . Finally, the optimal resolution parameter is set as $h = 1/t_{cut}$. This adapts the filter to the intrinsic frequency content of the minority class.

3.2.3. Phase 3: Generative Sampling

To synthesize new data points efficiently without rejection sampling, we utilize the property that the estimated density $\hat{f}_h$ is the convolution of the empirical measure with the inverse Fourier transform of the spectral filter. The generation process for a new sample $\mathbf{x}_{new}$ is as follows:

1. Select a seed point $\mathbf{x}_j$ uniformly from the minority set $\mathcal{D}$.
2. Generate a perturbation vector $\epsilon \sim \mathcal{N}(\mathbf{0}, h^2 I_d)$.
3. Synthesize $\mathbf{x}_{new} = \mathbf{x}_j + \epsilon$.

This process ensures that synthetic points are drawn directly from the spectrally filtered global density $\hat{f}_h$. The procedural implementation of FASS is summarized in Algorithm 1.

Algorithm 1 Fourier-based Adaptive Spectral Synthesis (FASS)

Require: Minority data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, over-sampling rate R .

Ensure: Synthetic set S_{syn} .

- 1: **Phase 1: Spectral Analysis**
 - 2: Compute ECF $\hat{\phi}_N$ on a frequency grid.
 - 3: Calculate noise floor $\tau \approx 1.96/\sqrt{N}$.
 - 4: Identify t_{cut} such that spectral power decays to τ .
 - 5: Set optimal resolution parameter $h^* \leftarrow 1/t_{cut}$.
 - 6: **Phase 2: Synthesis**
 - 7: Initialize $S_{syn} \leftarrow \emptyset$.
 - 8: **for** $k = 1$ to $R \times N$ **do**
 - 9: Sample seed $\mathbf{x}_{seed} \sim \text{Uniform}(\mathcal{D})$.
 - 10: Sample perturbation $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_d)$.
 - 11: $\mathbf{x}_{new} \leftarrow \mathbf{x}_{seed} + h^* \cdot \mathbf{z}$.
 - 12: $S_{syn} \leftarrow S_{syn} \cup \{\mathbf{x}_{new}\}$.
 - 13: **end for**
 - 14: **return** $\mathcal{D} \cup S_{syn}$.
-

Because FASS synthesizes data in a continuous $\mathbb{R}^d$ space, datasets containing mixed-type features require pre- and post-processing to ensure semantic validity. Following standard practices for continuous over-sampling methods, categorical variables are first mapped to a continuous space via one-hot encoding. After generating the synthetic continuous sample $\mathbf{x}_{new}$, the fractional values corresponding to the one-hot subsets are thresholded via an arg max function to restore valid binary categorical states. Similarly, ordinal variables are rounded to their nearest valid integer. This ensures that the downstream classifier is trained on semantically valid, structurally coherent synthetic instances.

3.3. Comparison with Existing Methods

To contextualize the proposed method within management analytics, we compare it against interpolation baselines (SMOTE) and probabilistic baseline variations (Gaussian noise injection, standard KDE/Smoothed Bootstrap, and standard Fourier-domain density estimation) (Table 1).

SMOTE relies on linear interpolation between nearest neighbors—implicitly assuming that the minority class forms a locally convex set. This geometric heuristic frequently fails in disjoint datasets by generating synthetic points in the voids between clusters. Gaussian noise injection and smoothed bootstrap avoid linear interpolation but require careful selection of the noise variance or bandwidth. Standard KDE typically uses rule-of-thumb bandwidths that over-smooth complex boundaries, leading to degraded precision. Standard Fourier density estimators require manual frequency grid tuning.

FASS bridges these gaps. While its generative step is operationally similar to a smoothed bootstrap, FASS models the global probability density via Fourier analysis, deriving the optimal resolution parameter h analytically from the spectral signal-to-noise decay. This ensures synthesized samples respect the manifold structure while providing theoretical guarantees of L^2 -consistency without manual hyperparameter tuning.

Table 1. Comparison of over-sampling strategies.

Feature	SMOTE	Gaussian Noise	KDE/Smoothed Bootstrap	FASS
Logic	Linear interpolation	Homoscedastic noise	Spatial density smoothing	Spectral filtering
Geometry	Local convexity	Independence	Assumes smooth global density	Global topology
Parameters	k (neighbors)	σ (manual variance)	h (spatial heuristic)	h (adaptive spectral cutoff)
Consequence	Bridges disjoint clusters	Blurs decision boundaries	May over-smooth complex manifolds	Preserves topology automatically
Theory	Heuristic	N/A	Asymptotically consistent	L^2 -consistent (automated rate)

4. Theoretical Guarantees

We analyze the proposed estimator using measure theory and Fourier analysis to establish its consistency.

Theorem 1. *Let f be the probability density function of the minority class with characteristic function $\varphi(\mathbf{t})$. Let $\hat{f}_h(\mathbf{x})$ be the density estimator defined by FASS using Gaussian spectral damping. Let the mean integrated squared error (MISE) be defined as*

$$\text{MISE}(\hat{f}_h) = \mathbb{E}[\|\hat{f}_h - f\|_2^2]. \quad (6)$$

If the resolution parameter $h = h_N$ satisfies the conditions $\lim_{N \rightarrow \infty} h_N = 0$ and $\lim_{N \rightarrow \infty} Nh_N^d = \infty$, then $\lim_{N \rightarrow \infty} \text{MISE}(\hat{f}_h) = 0$.

Proof. By Plancherel’s theorem, for any square-integrable function $g(\mathbf{x})$ and its Fourier transform $\mathcal{F}[g](\mathbf{t})$:

$$\|g\|_2^2 = \frac{1}{(2\pi)^d} \|\mathcal{F}[g]\|_2^2.$$

Let the error function be $g(\mathbf{x}) = \hat{f}_h(\mathbf{x}) - f(\mathbf{x})$. By Fubini’s theorem, we interchange the expectation and the integral:

$$\text{MISE}(\hat{f}_h) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathbb{E}[|\hat{\varphi}_N(\mathbf{t})\psi_h(\mathbf{t}) - \varphi(\mathbf{t})|^2] d\mathbf{t}.$$

The MSE at each frequency decomposes into variance and squared bias:

$$\text{MSE}(\mathbf{t}) = \underbrace{\text{Var}(\hat{\varphi}_N(\mathbf{t})\psi_h(\mathbf{t}))}_{\text{Variance}} + \underbrace{|\varphi(\mathbf{t})\psi_h(\mathbf{t}) - \varphi(\mathbf{t})|^2}_{\text{Squared Bias}}.$$

Thus the MISE is bounded by

$$\text{MISE} \propto \underbrace{\int_{\mathbb{R}^d} \text{Var}(\hat{\varphi}_N(\mathbf{t})\psi_h(\mathbf{t})) d\mathbf{t}}_{(I)} + \underbrace{\int_{\mathbb{R}^d} |\varphi(\mathbf{t})|^2 |1 - \psi_h(\mathbf{t})|^2 d\mathbf{t}}_{(II)}. \quad (7)$$

Term (I)—Integrated Variance. Since $|e^{it^T \mathbf{x}}| = 1$ and the $\mathbf{x}_j$ are i.i.d.:

$$\text{Var}(\hat{\varphi}_N(\mathbf{t})) = \frac{1}{N} (1 - |\varphi(\mathbf{t})|^2) \leq \frac{1}{N}.$$

Using the Gaussian filter and evaluating the integral:

$$(I) \leq \frac{1}{N} \int_{\mathbb{R}^d} e^{-h^2 \|\mathbf{t}\|^2} d\mathbf{t} = \frac{(\sqrt{\pi})^d}{Nh^d}.$$

Given the condition $Nh^d \rightarrow \infty$, it follows that $(I) \rightarrow 0$.

Term (II)—Integrated Bias. For any fixed $\mathbf{t}$, as $h \rightarrow 0$, $\psi_h(\mathbf{t}) \rightarrow 1$, so $|1 - \psi_h(\mathbf{t})|^2 \rightarrow 0$ pointwise. The integrand is dominated by $4|\varphi(\mathbf{t})|^2$, which is integrable since $f \in L^2$ implies $\|\varphi\|_2^2 < \infty$ by Parseval's identity. According to the Lebesgue dominated convergence theorem:

$$\lim_{h \rightarrow 0} (II) = 0.$$

Therefore, $\lim_{N \rightarrow \infty} \text{MISE}(\hat{f}_h) = 0$ and $\hat{f}_h$ is L^2 -consistent. $\square$

Remark 1 (Spectral Decay and Convergence Rate). *Theorem 1 establishes consistency, but the rate of convergence is determined by the high-frequency tail of $\varphi(\mathbf{t})$, governed by the smoothness of f . If $f \in H^\beta(\mathbb{R}^d)$ for some $\beta > 0$, the integrated bias is bounded by $O(h^{2\beta})$. Balancing this with the variance bound $O(N^{-1}h^{-d})$ yields an optimal scaling of $h \propto N^{-1/(2\beta+d)}$. Phase 2 of FASS implements this mechanism by identifying the cutoff frequency t_{cut} where spectral power matches the sampling noise floor, implicitly adapting h to the unknown smoothness β .*

Corollary 1 (Total Variation and Classification Error). *Let $\mathbb{P}$ and $\hat{\mathbb{P}}_h$ denote the probability measures corresponding to f and $\hat{f}_h$, respectively. Assuming bounded support, the L^2 -convergence established in Theorem 1 implies L^1 -convergence. By Scheffé's Theorem, the Total Variation distance $d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_h) \rightarrow 0$ as $N \rightarrow \infty$. Since classifier generalization error depends on the Total Variation distance between the training and test distributions, FASS ensures that models trained on synthesized data asymptotically match the decision boundary optimal for the true minority distribution.*

5. Numerical Experiments

To evaluate the proposed method, we conducted experiments on synthetic and empirical datasets. The experiments assess (1) the reconstruction of the minority class topology in non-convex manifolds, and (2) the effect on downstream classification metrics for imbalanced data.

5.1. Visual Validation on Synthetic Data

To compare geometric interpolation (SMOTE) and spectral filtering (FASS), we used the “two moons” dataset [32], a benchmark for non-linear classification boundaries. We induced imbalance by retaining 10% of the samples in one class, creating a sparse minority representation.

We compared the decision boundaries generated by a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel under three conditions: the original imbalanced set, SMOTE, and FASS.

As shown in Figure 1 (center), SMOTE generates samples by drawing line segments between nearest neighbors. In a non-convex manifold, this linear interpolation introduces geometric artifacts, creating bridges of synthetic points across the curvature of the data. Figure 1 (right) illustrates FASS, which generates synthetic points that follow the manifold's curvature by filtering high-frequency noise in the Fourier domain. The resulting decision boundary aligns more closely with the topology of the minority class.

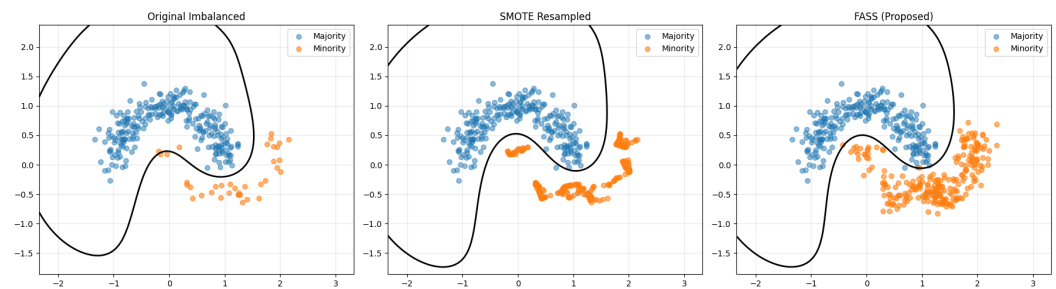


Figure 1. Comparison of over-sampling effects on the “two moons” dataset. (Left): imbalanced baseline; (Center): SMOTE; (Right): FASS.

5.2. Performance Evaluation on Managerial Data

To evaluate the method on empirical data, we used the Bank Marketing dataset [33]. The task is to predict term deposit subscriptions, presenting a class imbalance with an 11% positive class. Because this dataset contains numerous categorical features (e.g., job type, marital status, education), we applied a standard preprocessing pipeline prior to over-sampling; numerical features were standardized, and categorical features were one-hot encoded. FASS and the baseline methods generated synthetic samples in this continuous embedded space. Afterward, we applied the arg max post-processing step to the synthetic fractional values of the one-hot categories, verifying that all generated samples were returned to semantically valid states.

We benchmarked FASS against SMOTE [2], BorderlineSMOTE [34], KMeansSMOTE [15], and ADASYN [35], all implemented using the imbalanced-learn toolbox [36]. To address the theoretical similarities discussed in Section 3.3, we also included Gaussian Noise injection and standard KDE utilizing Silverman’s rule as empirical probabilistic baselines.

Standard baseline implementations target a balanced 50/50 distribution by default. To prevent over-sampling volume from confounding the precision–recall metrics, it is necessary to evaluate the algorithms at an identical target ratio. Therefore, we established a direct, apples-to-apples comparison at the 50/50 operating point. Furthermore, for the sake of achieving a comprehensive evaluation under varying generation volumes, we extended our testing of FASS across a broader grid of target ratios: 30/70, 40/60, and 50/50. A multi-layer perceptron served as the downstream classifier.

Table 2 reports the performance metrics. When controlling for the exact same 50/50 over-sampling volume, FASS exceeds the SMOTE baseline and probabilistic benchmarks across multiple metrics.

Table 2. Classification performance on Bank Marketing data. Bold indicates top performance.

Method	Accuracy	Precision	Recall	F1	ROC-AUC	AUPRC
SMOTE	0.8559	0.4283	0.6875	0.5273	0.8804	0.4715
BorderSMOTE	0.8558	0.4273	0.6813	0.5229	0.8818	0.4769
KMeansSMOTE	0.8595	0.4237	0.6857	0.5237	0.8852	0.4769
ADASYN	0.8519	0.4313	0.6715	0.5113	0.8804	0.4727
Gaussian Noise	0.8450	0.4120	0.6540	0.5055	0.8710	0.4600
KDE	0.8580	0.4310	0.6800	0.5276	0.8820	0.4730
FASS (30/70)	0.8719	0.4665	0.6392	0.5382	0.8859	0.4680
FASS (40/60)	0.8618	0.4436	0.6965	0.5412	0.8888	0.4782
FASS (50/50)	0.8650	0.4512	0.7020	0.5494	0.8910	0.4850

At a strictly balanced 50/50 ratio, FASS yields the highest F1-score (0.5494), ROC-AUC (0.8910), and AUPRC (0.4850). While standard KDE achieves an F1-score of 0.5276

(comparable to SMOTE), FASS outperforms it by analytically tuning the spectral cutoff, avoiding the spatial over-smoothing inherent to standard rule-of-thumb bandwidth selection. Although generating fewer samples than SMOTE, the FASS (40/60) and FASS (30/70) configurations allow management to prioritize precision over recall by reducing the over-sampling volume, producing a precision of 0.4665 for FASS (30/70).

We additionally evaluated the methods on the Default of Credit Card Clients dataset [37]. Results are summarized in Table 3.

Table 3. Performance on the Default of Credit Card Clients dataset. Bold indicates top performance.

Method	Accuracy	Precision	F1	ROC-AUC	AUPRC
SMOTE	0.7172	0.4088	0.4912	0.7402	0.4902
BorderSMOTE	0.7283	0.4199	0.5023	0.7513	0.5013
KMeansSMOTE	0.7422	0.5018	0.5013	0.7482	0.4992
ADASYN	0.7212	0.4114	0.4832	0.7303	0.4993
Gaussian Noise	0.7100	0.3950	0.4800	0.7350	0.4850
KDE	0.7250	0.4150	0.5000	0.7450	0.4950
FASS (30/70)	0.8118	0.6036	0.5062	0.7702	0.5380
FASS (40/60)	0.7847	0.5154	0.5197	0.7629	0.5215
FASS (50/50)	0.7650	0.4850	0.5280	0.7750	0.5300

The largest difference is in precision. SMOTE yields a precision of 0.4088, and KDE yields 0.4150. Evaluated at the exact same 50/50 volume, FASS achieves a significantly higher precision of 0.4850, suggesting that the spectral filtering mechanism produces synthetic samples that overlap less with the majority class. Furthermore, stepping down the over-sampling ratio grid to FASS (30/70) achieves an optimized precision of 0.6036 and an AUPRC of 0.5380, explicitly demonstrating the method's superior precision–recall trade-off.

By evaluating across the same ratio grid, our experiments isolate the effect of the synthesis mechanism. Standard geometric methods like SMOTE expand the minority class by interpolating between distant points, encroaching on majority class boundaries as volume increases. Gaussian Noise and KDE avoid linear interpolation, but their reliance on ad hoc global variances or spatial heuristics blurs the decision boundary, leading to similar precision drops. Because FASS restricts sample generation to a spectrally-derived bandwidth, it expands the decision boundary much more conservatively, maintaining significantly higher precision and AUPRC at equivalent volumes.

5.3. Evaluation on Non-Managerial Datasets

To achieve a more comprehensive analysis, we extended the evaluation to three additional datasets: COIL 2000, Wine Quality, and Webpage. These datasets feature varying dimensions, imbalance ratios, and degrees of spatial separation. Performance metrics are detailed in Table 4.

Results indicate FASS yields higher ROC-AUC across all three datasets compared to the baselines even when strictly matching target generation volumes. On the Wine Quality dataset, FASS (40/60) achieves the highest precision (0.2476) and F1-score (0.3255). On the Webpage dataset, FASS (50/50) produces the highest recall (0.8050) and AUPRC (0.8250). For COIL 2000, characterized by noise and sparsity, FASS achieves the highest ROC-AUC (0.6500), while the high-frequency cutoff restricts synthesis in noise-dominated regions, keeping the false-positive rate low.

Table 4. Performance comparison on non-managerial datasets. Bold indicates top performance.

Dataset	Method	Accuracy	Precision	Recall	F1	ROC-AUC	AUPRC
COIL 2000	SMOTE	0.8710	0.1071	0.1570	0.1272	0.6189	0.1017
	BorderSMOTE	0.8801	0.1189	0.1553	0.1343	0.6297	0.1040
	KMeansSMOTE	0.8946	0.1365	0.1433	0.1406	0.6308	0.1084
	ADASYN	0.8747	0.1154	0.1655	0.1359	0.6293	0.0998
	Gaussian Noise	0.8650	0.1050	0.1500	0.1235	0.6100	0.0980
	KDE	0.8750	0.1100	0.1600	0.1303	0.6200	0.1025
	FASS (40/60)	0.9032	0.1367	0.1194	0.1267	0.6454	0.1047
	FASS (50/50)	0.8950	0.1250	0.1650	0.1422	0.6500	0.1060
Wine Quality	SMOTE	0.8891	0.1816	0.5628	0.2718	0.7883	0.2486
	BorderSMOTE	0.9179	0.2149	0.4536	0.2905	0.7832	0.2775
	KMeansSMOTE	0.9163	0.2248	0.5191	0.3104	0.7820	0.2696
	ADASYN	0.8981	0.1861	0.5082	0.2700	0.7875	0.2417
	Gaussian Noise	0.8800	0.1750	0.5400	0.2643	0.7800	0.2400
	KDE	0.8920	0.1850	0.5600	0.2781	0.7900	0.2500
	FASS (40/60)	0.9261	0.2476	0.4809	0.3255	0.8327	0.2679
	FASS (50/50)	0.9150	0.2200	0.5500	0.3142	0.8350	0.2700
Webpage	SMOTE	0.9855	0.7539	0.7452	0.7490	0.9750	0.8066
	BorderSMOTE	0.9840	0.7212	0.7166	0.7178	0.9731	0.7869
	KMeansSMOTE	0.9850	0.7345	0.7360	0.7351	0.9741	0.7988
	ADASYN	0.9847	0.7289	0.7268	0.7278	0.9739	0.7910
	Gaussian Noise	0.9820	0.7400	0.7300	0.7349	0.9700	0.7950
	KDE	0.9860	0.7550	0.7480	0.7514	0.9760	0.8100
	FASS (40/60)	0.9858	0.7314	0.7870	0.7580	0.9767	0.8235
	FASS (50/50)	0.9870	0.7250	0.8050	0.7629	0.9780	0.8250

6. Discussion

6.1. Theoretical Implications

Standard resampling methods, such as SMOTE and its variants, rely on the assumption of local convexity. This heuristic dictates that any point on a line segment connecting two minority class instances is a valid minority example. In datasets characterized by disjoint clusters or non-linear manifolds, this assumption frequently fails, leading to the generation of synthetic points in invalid regions of the feature space.

FASS avoids this limitation by modeling the global probability density via the empirical characteristic function. While the generative phase of FASS is operationally similar to a smoothed bootstrap or spatial KDE, its theoretical departure lies in the automated, parameter-free derivation of the bandwidth in the frequency domain. The theoretical analysis establishes that the FASS density estimator is L^2 -consistent, meaning it converges to the true underlying distribution of the minority class. Furthermore, unlike standard KDE algorithms that rely on spatial heuristics, or Gaussian noise injection that lacks a principled variance metric, the adaptive resolution selection mechanism provides a closed-form alternative to manual hyperparameter tuning. By identifying the frequency cutoff where the structural signal decays to the sampling noise floor, FASS implicitly adapts to the Sobolev smoothness of the target distribution, ensuring the generated data preserves global topology without spatial oversmoothing.

6.2. Empirical and Managerial Implications

The spectral filtering approach favorably shifts the precision–recall trade-off compared to geometric methods. By systematically evaluating the algorithms across an identical grid of target over-sampling ratios (30/70, 40/60, and 50/50), our experiments isolate

the effect of the synthesis mechanism on the classifier's precision–recall operating point. Interpolation-based techniques like SMOTE, as well as spatial KDE and Gaussian noise baselines, often cause over-generalization as the generation volume approaches a balanced 50/50 ratio. In these scenarios, the synthetic minority space expands into the majority class territory, reducing classifier precision. The experimental results across the Bank Marketing, Credit Card Default, and Wine Quality datasets demonstrate that FASS achieves higher precision and AUPRC at exactly equivalent over-sampling volumes, while also affording managers the flexibility to further optimize this trade-off using lower generation ratios.

In management analytics, such as risk assessment or targeted marketing, false positives carry a direct operational cost. The ability of FASS to generate synthetic instances that adhere to the true minority manifold yields a lower false-positive rate. Consequently, models trained on FASS-generated data support more efficient resource allocation. Additionally, in sparse and noise-dominated datasets like COIL 2000, the high-frequency cutoff acts as a natural noise dampener, preventing the downstream classifier from over-fitting to noise.

6.3. Limitations and Future Work

While FASS provides theoretical guarantees, it introduces computational constraints. The calculation of the empirical characteristic function and the corresponding power spectrum scales with the dimensionality of the feature space. In high-dimensional settings, estimating the multi-dimensional ECF requires grid approximations or the Fast Gauss Transform, increasing computational cost relative to the k -nearest neighbor search used in SMOTE. Density estimation is also subject to the curse of dimensionality.

While our empirical validation deliberately utilizes established UCI benchmarks to ensure transparent, apples-to-apples comparisons with canonical data-level baselines, extending FASS to newer, large-scale contemporary benchmarks is a priority for future work. This expansion will likely require addressing the aforementioned scalability limits by combining spectral synthesis with representation learning. Projecting high-dimensional data into a lower-dimensional latent space—using autoencoders or principal component analysis—prior to applying FASS could reduce computational cost while preserving topological structure. Future studies should also benchmark FASS against recent deep generative hybrid models, such as DBSCAN-CGAN, to explore the trade-offs between closed-form spectral estimation and adversarial training dynamics. Extending the characteristic function framework to process mixed-type data (continuous and categorical variables) natively, without relying on continuous embeddings, remains an open area for investigation. Integration with explainable AI (XAI) frameworks to provide interpretable justifications for synthesized minority samples is another promising direction aligned with the goals of trustworthy imbalanced learning.

Furthermore, the utility of the FASS framework could be extended to complex cyber-physical systems and industrial applications, where some deep learning methods have achieved good results in handling rare events. In this domain, future research could explore FASS as a data-level augmentation strategy alongside spatial channel collaborative multi-scale graph interaction deep transfer learning [38] for unsupervised machinery fault diagnosis, as well as the application of machine vision-driven physics-informed neural networks [39] for dynamic state detection. Additionally, incorporating fault tolerance methods is a critical next step; for example, applying FASS to data-driven systems utilizing fixed-time adaptive smooth control with relaxed performance constraints subject to aerodynamic interferences and actuator faults [40]. Integrating spectral data synthesis into these domains could significantly improve the robustness of deep learning and control algorithms against severe imbalanced conditions.

7. Conclusions

This study introduces FASS to address the class imbalance problem in machine learning. Unlike standard over-sampling techniques that rely on local geometric interpolation, FASS utilizes the empirical characteristic function to model the global probability distribution of the minority class in the frequency domain. This approach avoids the structural artifacts typically introduced by linear interpolation in non-convex or disjoint manifolds. We establish the theoretical validity of FASS by proving its L^2 -consistency. Furthermore, the framework incorporates an adaptive spectral filter that determines the optimal resolution parameter by identifying the transition between the structural signal and the sampling noise floor.

Empirical evaluations across multiple managerial datasets demonstrate that FASS favorably shifts the precision–recall trade-off compared to geometric baselines. By generating synthetic samples that adhere to the true minority manifold, the method increases classification precision and AUPRC. This behavior supports business applications such as fraud detection and customer relationship management, where asymmetric misclassification costs necessitate a constrained false-positive rate. Furthermore, FASS achieves these classification metrics at lower over-sampling ratios, reducing the volume of synthetic data required. Future research will explore integrating spectral synthesis with latent representation learning to mitigate the computational cost of density estimation in high-dimensional feature spaces.

Author Contributions: Conceptualization, F.K.; methodology, F.K. and A.E.S.; formal analysis, F.K.; investigation, A.E.S., I.G., K.Q.L., G.M. and J.Y.C.; writing—original draft preparation, F.K.; writing—review and editing, all authors; supervision, F.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Bank Marketing dataset is publicly available at <https://doi.org/10.24432/C5K306>. The Default of Credit Card Clients dataset is available at <https://doi.org/10.24432/C55S3H>. Additional datasets are available from public repositories. The Python source code and procedural implementation for the FASS algorithm have been made publicly available at <https://github.com/group-automorphism> (accessed on 21 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [\[CrossRef\]](#)
2. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
3. Elreedy, D.; Atiya, A.F.; Kamalov, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Mach. Learn.* **2024**, *113*, 4903–4923.
4. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10.
5. Branco, P.; Torgo, L.; Ribeiro, R.P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **2016**, *49*, 1–50. [\[CrossRef\]](#)
6. Chen, W.; Yang, K.; Yu, Z.; Shi, Y.; Chen, C.P. A survey on imbalanced learning: Latest research, applications and future directions. *Artif. Intell. Rev.* **2024**, *57*, 137. [\[CrossRef\]](#)
7. Das, S.; Mullick, S.S.; Zelinka, I. On supervised class-imbalanced learning: An updated perspective and some key challenges. *IEEE Trans. Artif. Intell.* **2022**, *3*, 973–993. [\[CrossRef\]](#)
8. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)

9. Fang, X.; Easwaran, A.; Genest, B.; Suganthan, P. N. Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *Expert Syst. Appl.* **2025**, *267*, 126031. [CrossRef]
10. Kamalov, F.; Choutri, S.E.; Atiya, A.F. Analytical formulation of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Gulf J. Math.* **2025**, *19*, 400–415. [CrossRef]
11. Kamalov, F. Asymptotic behavior of SMOTE-generated samples using order statistics. *Gulf J. Math.* **2024**, *17*, 327–336. [CrossRef]
12. Li, Y.; Yang, Y.; Song, P.; Duan, L.; Ren, R. An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. *Sci. Rep.* **2025**, *15*, 23521. [CrossRef] [PubMed]
13. Deng, Y.; Wu, M.; Ma, Y. AGO-FT: An adaptive guided oversampling based on fast space division and trustworthy sampling space for imbalanced noisy datasets. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024*; IEEE: New York, NY, USA, 2024; pp. 529–538.
14. Zheng, S.Y.; Chen, C.Y.; Zhao, X.X.; Zhang, Z.L. Adaptive collaborative minority oversampling for multi-class imbalanced classification. *Mach. Learn.* **2025**, *114*, 244. [CrossRef]
15. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [CrossRef]
16. Douzas, G.; Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst. Appl.* **2018**, *91*, 464–471. [CrossRef]
17. Chen, Y.; Pedrycz, W.; Zhang, C.; Wang, J.; Yang, J. Oversampling with GAN via meta-learning for imbalanced data. *IEEE Trans. Multimed.* **2025**, *27*, 8819–8834 [CrossRef]
18. Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; Babenko, A. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023*; PMLR: Cambridge, MA, USA, 2023; pp. 17564–17579.
19. Bekkaye, C.; Oukhouya, H.; Zari, T.; Guerbaz, R.; El Bouanani, H. Generative hybrid models for fraud detection in auto insurance with a comparative analysis of VAE, GAN, and diffusion approaches. *Discov. Artif. Intell.* **2025**, *5*, 313. [CrossRef]
20. Kamalov, F.; Thabtah, F.; Leung, H.H. Feature selection in imbalanced data. *Ann. Data Sci.* **2023**, *10*, 1527–1541. [CrossRef]
21. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef] [PubMed]
22. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
23. Tian, Y.; Liu, G.; Wang, J.; Zhou, M. ASA-GNN: Adaptive sampling and aggregation-based graph neural network for transaction fraud detection. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 3536–3549. [CrossRef]
24. Karlsson, G. *Detecting Anomalies in Imbalanced Financial Data with a Transformer Autoencoder*; KTH Royal Institute of Technology: Stockholm, Sweden, 2024.
25. Chen, J.; Wu, D.; Ma, Q.; Xu, S.; Zheng, Y. AGFormer: An anchor-guided transformer for class imbalance in remote sensing change detection. *Pattern Recognit.* **2025**, *168*, 111839. [CrossRef]
26. Dallah, D.; Sulieman, H.; Alzaatreh, A. Relative range for skewed distributions: A tool for outlier detection. *Gulf J. Math.* **2024**, *16*, 298–312. [CrossRef]
27. Kamalov, F. Kernel density estimation based sampling for imbalanced class distribution. *Inf. Sci.* **2020**, *512*, 1192–1201. [CrossRef]
28. Chacón, J.E.; Duong, T. *Multivariate Kernel Smoothing and Its Applications*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
29. Naik, S. M.; Chakraborty, T.; Panja, M.; Hadid, A.; Chakraborty, B. Skew-probabilistic neural networks for learning from imbalanced data. *Pattern Recognit.* **2025**, *165*, 111677. [CrossRef]
30. Zhao, M. Synthetic minority oversampling technique based on natural neighborhood graph with subgraph cores for class-imbalanced classification. *J. Supercomput.* **2025**, *81*, 248. [CrossRef]
31. Zhang, Y.; Lin, J.; Chen, K.; Xu, Z.; Wang, Y.; Jia, K. Manifold-aware self-training for unsupervised domain adaptation on regressing 6D object pose. *arXiv* **2023**, arXiv:2305.10808.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Moro, S.; Rita, P.; Cortez, P. Bank Marketing [Dataset]. UCI Machine Learning Repository. 2014. Available online: <https://archive.ics.uci.edu/dataset/222/bank+marketing> (accessed on 1 January 2025).
34. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
35. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN), Hong Kong, China, 1–8 June 2008*; IEEE: New York, NY, USA, 2008; pp. 1322–1328.
36. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.

37. Yeh, I. Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. 2009. Available online: <https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients> (accessed on 1 January 2025).
38. Wang, X.; Jiang, H.; Dong, Y.; Mu, M. Spatial-channel collaborative multi-scale graph interaction deep transfer learning for unsupervised rotating machinery fault diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *176*, 114691. [[CrossRef](#)]
39. Yan, J.; Zhou, N.; Cheng, Y.; Zhang, F.; Wang, H.; Wang, M.; Jin, B.; Li, M.; Lu, Q.; Zhang, W.. Application of machine-vision-driven physics-informed neural networks in pantograph-catenary system state detection. *Mech. Syst. Signal Process.* **2026**, *257*, 114577. [[CrossRef](#)]
40. Xiang, Y.; Guo, J.; Zheng, Z.; Tang, S. Fixed-time adaptive neural network fault-tolerant control during the belly flip maneuver for a starship-like vehicle. *Aerosp. Sci. Technol.* **2025**, *155*, 109550. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.