A Novel Adaptive Framework for Data Complexity Analysis in Imbalanced Binary Classification

Debashis Roy
Department of Computer and System Sciences, Visva Bharati Santiniketan, WB, India

Anandarup Roy
Sarojini Naidu College for Women Dum Dum, Kol - 700028, India

Utpal Roy
Department of Computer and System Sciences, Visva Bharati Santiniketan, WB, India

ABSTRACT
Improving classification performance when the dataset is imbalanced—that is, when the negative (majority) class is stronger than the positive (minority) class—is one of the most important problems in machine learning. Several researchers alleviated this situation by developing various data-level and algorithm-level techniques. However, it is important to note that an imbalanced dataset is not the sole factor compromising classification performance. It's not just the imbalanced dataset that makes classification harder; things like overlap, local instance ambiguity, intrinsic structural complexity, and so on also make the classification more complicated. Very few researchers have focused on data complexity, especially along with imbalanced datasets. This paper proposes a novel adaptive framework that measures data complexities like instance overlap, multiresolution overlap, structural overlap, kNN-based complexity for minority instances, and more. This systematized adaptive measure selection framework sorts through the complexity of the data based on how imbalanced the datasets are and suggests preprocessing steps and the right models to make the classification task easier. The work includes a theoretical analysis, the lemma, and the corollary, as well as specific steps for putting the ideas into practice. This framework, which is aware of taxonomies and provides actionable insights that greatly improve the performance of imbalanced classification, makes it new and very useful for both researchers and practitioners.

General Terms
Pattern Recognition, Machine Learning, Imbalance learning, Algorithms

Keywords
Data Complexity, Imbalanced classification, Adaptive Measure Selection, Overlap, Theoretical Bounds

1. INTRODUCTION
Imbalanced learning is a tenacious problem in domains ranging from medical diagnostics to cybersecurity [1]. Classifiers are often misguided by a large number of majority classes, though the minority class is more significant due to rare events. An imbalanced dataset can further distort even classification results when data complexity is present [2]. Most of the time, researchers and practitioners don't pay attention to data complexity problems like feature overlap, ambiguous instance regions, and complex structural patterns. Instead, they focus on data classification and resampling methods [3]. Several experts says that most essential part in machine learning is data engineering, because the perfection of any model hinge on data such that nature of data, data complexity etc. Ho and Basu (2002) [4] proposed a data complexity measure at first, and many researchers have since published modified or extended versions of it [5, 6, 7]. Additionally, researchers proposed four specific data complexity measures such as CM, wCM, dwCM, BI3 for imbalanced learning [8, 9, 10]. Previous research has looked at different aspects of these problems, but there isn't yet a framework that brings together different types of data complexity and gives usable ways to deal with them. This study introduces a unified framework that combines traditional complexity measures, including Fisher's Discriminant Ratio and kNN-based metrics, with metrics that evaluate advance complexities like instance, multiresolution, and structural overlap, thereby providing a thorough assessment of data complexity in imbalanced classification. The complexity is measured according to degree of imbalance, such as highly imbalanced data, moderately imbalanced data, and high dimensional imbalanced data. The proposed adaptive measure selection algorithm suggests the most relevant metrics based on the characteristics of the imbalanced dataset. This gives us direct information for preprocessing and model choice. A theoretical framework is presented, complete with corollaries and lemmas, that has the best chance of getting rid of classification mistakes. The framework also includes detailed instructions on how to implement it, good ways to display data, and an overview of experiments using a synthetic controlled imbalanced dataset, which makes it a useful tool for both researchers and practitioners.

2. MOTIVATION AND RELATED WORK
In a number of studies, researchers have found that an uneven dataset by itself is not enough to confuse the classifiers. Instead, combining an uneven dataset with different complexities increases the rate of misclassification. Barella et al. (2022) described existing complexity measures with imbalanced dataset challenges. It also draws a clear picture of how data complexity changes before and after applying data-level methods to balance the imbalanced dataset [2]. Santos et al. (2023) show that overlap is a bigger problem when the dataset isn't balanced. They also show a new way to measure the complexity of overlap [11]. Vuttipittayamongkol et al. (2021) [12] highlighted the importance of addressing overlap rather than focusing on class balancing strategies. Similarly, Fatima et al. (2021) [13] proposed a feature selection methodology focused on minimizing the overlap in fraud detection. Komorniczak et al. (2022) [14] investigated how oversampling methods influence data complexity and classification results. It also emphasizes the importance of data complexity measures in the case of imbalanced learning. Lorena et al. (2021) reviewed the gauges of complexity in various domains, such as feature selection, classification selection, and data preprocessing [7]. This study combines adaptive data complexity analysis with imbalanced binary classification to show a full method that changes how it classifies data based on real-time indicators of data complexity. This makes the model easier to understand and improves

performance.

3. THEORETICAL BACKGROUND AND NOTATIONS USED

Let,

With N_0 majority class and N_1 minority samples $(N_0 >> N_1)$. The framework evaluates:

I. Feature Overlapping Measures (F1–F4).
II. Advanced Overlap Measures mentioned in paper [11]
- Instance Overlap (IO)
- Multiresolution Overlap (MRO)
- Structural Overlap (SO)

III. Neighborhood Measures (N1–N4, T1)
IV. Linear Separability Measures (L1–L3)
V. Imbalance-Specific Metrics (CM, wCM, dwCM, BI^3)
VI. Dataset Typology: Three primary scenarios are considered:
- Highly Imbalanced Datasets
- Moderately Imbalanced Datasets
- High-Dimensional Datasets with High/Moderate Imbalance.

4. DETAILS OF DATA COMPLEXITY MEASURES

4.1 Feature Overlapping Measures

Maximum Fisher's Discriminant Ratio for Feature j:

$$FDR_j = \frac{\sigma_{0,j}^2 + \sigma_{1,j}^2}{(\mu_{0,j} - \mu_{1,j})^2} \quad \Rightarrow \quad F1 = \max_{j=1,\ldots,d} FDR_j$$

FDR$_j$ measures how far apart the classes are for feature **j** by looking at the squared difference between the class means and the total variance. Higher values indicate better individual feature discrimination.

Maximum Fisher's Discriminant Ratio (F1) selects the feature with the best discrimination among all features. It provides a quick indication of whether any single feature can effectively distinguish between the classes.

- **Volume of Overlap Region (F2):**

$$F2 = \int_{\mathbb{R}^d} \min\{p_0(x), p_1(x)\}\, dx$$

F2 estimates the common volume where both class-conditional densities overlap. A higher F2 suggests greater inherent overlap in the feature space, which can limit classifier performance.

- **Feature Efficiency (F3):**

$$\eta_j = \frac{H(y)}{I(y; x_j)} \quad \Rightarrow \quad F3 = \max_j \eta_j$$

$I(y; xj)$ is the mutual information between the label and feature **j**, and **H(y)** is the entropy of the labels. **F3** reflects the efficiency of a single feature in conveying class information; higher values imply better discriminatory capability.

- **Collective Feature Efficiency (F4):**

$$F4 = \frac{H(y)}{I(y; x)}$$

In contrast to F3, **F4** assesses the joint discriminative power of all features together. It provides an overall measure of how informative the complete feature set is with respect to the class labels.

4.2 Advanced Overlap Measures

- **Instance Overlap (IO):**

$$IO = \frac{1}{N} \sum_{i=1}^{N} I\left(|P(y = 1 \mid x_i) - 0.5| < \epsilon\right)$$

Where is a small threshold.

IO quantifies the fraction of instances that are ambiguous, where the estimated posterior probability is near 0.5. Higher **IO** indicates a large number of borderline cases, making classification more challenging.

- **Multiresolution Overlap (MRO):** For clusters $\{C_r\}_{r=1}^{R}$ at different resolutions,

$$MRO = \frac{1}{R} \sum_{r=1}^{R} \left(\frac{1}{|C_r|} \sum_{c \in C_r} I\{c \text{ is mixed}\} \right)$$

By clustering the data at different resolutions (indexed by **r**), **MRO** captures how frequently clusters contain mixed classes. Values closer to **1** imply that at most scales, clusters are not pure, signifying severe overlap.

- **Structural Overlap (SO):** Given a KNN graph G with cross-class edges E_{cross} and total edges E_{total},

$$SO = \frac{E_{\text{total}}}{E_{\text{cross}}}$$

SO measures the fraction of edges in a **kNN graph** that connect instances from different classes. A high SO indicates that the classes are structurally intertwined, complicating the formation of clear decision boundaries.

4.3 Neighborhood Measures

- **Fraction of Points on the Class Boundary (N1):**

B is the set of points with at least one neighbor from the other class. N1 measures how many instances reside close to class boundary. A higher **N1** implies that a larger proportion of data is ambiguous, which can lead to misclassification.

- **Ratio of Average Intra-/Inter-Class NN Distances (N2):**

$$N2 = \frac{d_{\text{inter}}}{d_{\text{intra}}}$$

In the same class, **N2** compares the average distance to nearest neighbors against those in the opposite class. Under ratio, suggest well-separated clusters.

- **Leave-one-out NN error rate (N3):**

$$N3 = \frac{N}{\text{Number of misclassifications}}$$

N3 measure estimates the local classification error using a 1-nearest neighbor classifier in a leave-one-out setting. Elevated N3 values suggest that the local neighborhood structure is noisy and complex.

- **1-NN nonlinearity measure (N4):**

$$N4 = |E_{\text{orig}} - E_{\text{interp}}|$$

N4 Illustrates the disparity in error rates between the original dataset and its interpolated counterpart, indicating the nonlinearity of the decision boundary.

- **T1: Fraction of maximum covering spheres among minority points.**

$$T1 = \frac{N_1}{S_{\max}}$$

Here, **S_{max}** revels the maximum number of non-overlapping covering spheres able to cover the points of minority class. **T1** measures the minority class's distribution in the feature space; higher values suggest a more complicated or scattered minority structure.

4.4 Linear Separability Measures

- **L1: Minimized sum of error distances.**

$$L1 = \min_{w,b} \sum_{i=1}^{N} |f(x_i)|, \quad \text{with } f(x) = w^\top x + b$$

L1 minimizes the sum of absolute distances of misclassified points from the decision border to estimate the general misclassification margin. Lower values mean that, generally, instances are more likely to be properly classified.

- **L2: Training error of a linear classifier.**

The empirical error of a linear classifier is denoted by **L2**. Higher values imply that more complex models may be required, while lower values suggest that a linear model can adequately separate the classes.

- **L3: Difference between errors of linear and nonlinear classifiers.**

$$L3 = |E_{\text{nonlinear}} - E_{\text{linear}}|$$

L3 grasps the variations in error rates between a linear and nonlinear classifier (e.g., 1-NN classifier). A greater L3 suggests that the real decision boundary is complicated and non-linear since a nonlinear model greatly outperforms a linear model.

4.5 Imbalance-Specific Metrics

- **CM (Complexity Measure for Imbalanced Datasets) :** it is basically kNN-based complexity for minority instances.

For each minority instance **X_i (with y_i):**

$$\text{CM} = \frac{1}{N_1} \sum_{i:y_i=1} \left(\frac{\#\text{ majority neighbors of } \mathbf{x}_i}{k} \right).$$

CM identifies the local neighborhood composition for minority class instances. Large value of CM indicates most of minority samples are surrounded by majority instances, which suggests which suggests a severe imbalance at the local level.

- **wCM (Weighted Complexity Metric) :** It is using distance weighting to measure complexity.

$$wCM = \frac{1}{N_1} \sum_{i:y_i=1} \frac{\sum_{j=1}^{k} w(d_{ij}) I\{y_{ij} = 0\}}{\sum_{j=1}^{k} w(d_{ij})}$$

Although **wCM** is similar to **CM**, it encompasses distance weights to account for the proximity of neighbors. This revels a more nuanced perspective on the vulnerability of the minority class.

- **dwCM(Dual Weighted Complexity Metric):** It is integrating both distance and density weighting.

$$dwCM = \frac{1}{N_1} \sum_{i:y_i=1} \frac{\sum_{j=1}^{k} w(d_{ij}) \rho_j I\{y_{ij} = 0\}}{\sum_{j=1}^{k} w(d_{ij}) \rho_j}$$

where ρj is a local density estimate.

dwCM further refines **wCM** by including a local density factor, **ρj**. This dual weighting reports for both distance and density, offering a robust measure of minority risk.

- **BI^3 (Bayes Imbalance Impact Index):** it is based on the Bayes optimal posterior.

$$BI^3 = \mathbb{E}_{x \sim p(x|y=1)} \left[|P^*(x) - 0.5| \right]$$

with $P*(x)$ denoting the posterior from the Bayes optimal classifier.

BI^3 assesses how far the optimal posterior probability for minority samples deviates from 0.5. A lower BI3 means that even the Bayes optimal classifier has trouble confidently classifying minority cases, which shows that the imbalance has a big effect.

5. NOVEL ADAPTIVE FRAMEWORK AND DATASET TYPOLOGY

5.1 Theoretical Inference And Guidelines

To connect the gauge of data complexity and classification performance, Two key theoretical aspects are derived:

Corollary 1:

If

$$F2 = \int_{\mathbb{R}^d} \min\{p_0(x), p_1(x)\}\, dx \geq \gamma$$

Consequently, even a optimal classifier cannot fulfil an error rate lower than **γ**. This outcome indicates that significant feature overlap establishes a fundamental limit on accuracy.

Lemma 1:

Let **δ∈[0,1]** be a specified threshold representing a lower bound on the leave-one-out error rate of the 1-NN classifier. That is, if the observed leave-one-out error rate, denoted as **N3**, satisfies,

then the overall misclassification risk ***R*** is bounded below by

$$R \geq f(\delta),$$

where *f*:[0,1]→[0,1] is a monotonically increasing function. where in this formula, **δ** is not defined as **N3** itself but is a chosen threshold; if the gauged N3 surpasses this threshold, it suggests a higher lower bound on the overall risk ***R***. So, this lemma entrenches a lower bound on the error based on local neighborhood complexity.

Based on these results, the recommended guidelines prioritize:

- **Feature-based measures (F1–F4)** when substantial overlap is detected.
- **Instance overlap (IO)** if many samples lie in ambiguous regions.
- **Multiresolution (MRO) and structural (SO)** measures when clusters are mixed or the kNN graph shows significant cross-class connectivity.

- **Neighborhood (N1, N3)** and **linear separability (L2, L3)** measures to assess the complexity of the decision boundary.
- **Imbalance-specific metrics (CM, wCM, dwCM, BI3)** to evaluate minority class vulnerability and inform re-sampling or cost-sensitive strategies.

In certain cases, such as highly imbalanced datasets with extreme feature overlap and ambiguous instances, robust resampling and specialized classifiers may be necessary; for moderately imbalanced or high dimensional datasets, robust classification techniques may be combined with dimensionality reduction.

5.2 Adaptive Measure Selection Algorithm

The proposed adaptive algorithm comprises the following steps:

Algorithm: Adaptive Complexity Diagnosis and Recommendation

Input: Dataset **D**, number of neighbors ***K***, threshold **{γ,ϵ,δ}**, dataset type **T**.

Output: Recommended preprocessing and model strategy.

Steps:

I. Data Preprocessing:

- Apply normalization and dimensionality reduction if necessary.

II. Compute Complexity Measures:

- Compute:

III. Determine Dataset Typology:

- Categorize the dataset into one of the following:

  A. Highly Imbalanced

  B. Moderately Imbalanced

  C. High-Dimensional Imbalanced

IV. Adapt Based on Dataset Type T :

A. Highly Imbalanced Datasets:

- If $\boldsymbol{IO}$ or $\boldsymbol{F2}$ is high, recommend aggressive re-sampling or cost-sensitive learning.
- Emphasize imbalance-specific metrics:

B. Moderately Imbalanced Datasets:

- Balance feature engineering (**F-measures**) with neighborhood insights (**N-measures**).
- Recommend moderate re-sampling and model regularization.

C. High-Dimensional Imbalanced Datasets:

- Prioritize dimensionality reduction for improving **F1–F4** interpretability.
- Focus on structural measures **(SO)** and adaptive clustering for **MRO**.

1. Assess Classifier Complexity:

- Use $\boldsymbol{N1}, \boldsymbol{N3}, \boldsymbol{L1-L3}$ to evaluate classifier complexity.

2. Output Final Recommendation:

- Suggest preprocessing steps and the best classifier strategy.

5.3 Guidelines for Selecting Data Complexity Measures

Table 1. Exhibits a guideline to select complexity measures

Data Complexity	Suggested Complexity Measures	Rationale
High Feature Overlap	F1, F2, F3, F4	Measure how features separate classses singnificantly; high **F2** indicates extreme overlap.
Ambiguous Local Decision Boundaries	N1, N3, N4, T1	High **N1** and **N3** values indicate many points near the boundary, suggesting complex local structures.
Potential Linear Non-Separability	L1, L2, L3	Overhead **L1** and **L3** values infer that linear classifiers may fail; center on **L2** for baseline error measurement.
Ambiguous Instances	IO	A decent IO represents many instances reside in regions of uncertainty.
Multiscale Overlap Effects	MRO	Overlap that changes with resolution, marked by multiresolution clustering analysis.
Intertwined Structural Patterns	SO	Surpass SO indicates that classes are highly connected in the intrinsic data structure.
Imbalanced Datasets with Minority Vulnerability	CM, wCM, dwCM, BI3	Focus on the impact of imbalance; **BI³** indicates fundamental difficulty due to imbalance, while **wCM/dwCM** account for distance/density effects.
Mixed/Uncertain Conditions	A hybrid approach using multiple measures across categories	Establishes a comprehensive analysis when dataset characteristics are not clearly defined.

5.4 Adaptation Based on Dataset Typology

Table 2. Taxonomy of complexity measures that categorizes imbalanced datasets as follows:

Dataset Type	Characteristics	Key Complexity Focus	Suggested Strategy
Highly Imbalanced	$N_1 \ll N_0$; minority severely under represented	High IO, high **CM/wCM/dwCM, high BI³**; moderate to high F2	Vigorous re-sampling, cost-sensitive learning, and specialized minority-oriented feature engineering.
Moderately Imbalanced	***N*1** is small but not extreme; moderate class skew	Balanced F-measures and N-measures; moderate **IO** and **CM**	moderate treatment with with attention to local boundary measures.
High-Dimensional, Highly Imbalanced	Large d; extreme minority scarcity	Emphasis on F1–F4, SO, and MRO; potential curse of dimensionality effects	Dimensionality reduction, graph-based methods to capture SO, and specialized high-dimensional re-sampling techniques.
High-Dimensional, Moderately Imbalanced	Larged; moderate imbalance	Focus on feature efficiency (F3, F4) and structural measures (SO)	Dimensionality reduction, moderate imbalance corrections.

These suggestions ensure that the diagnostic procedure and ensuing modeling decisions are customized to the inherent challenges of the specific imbalanced dataset type with corresponding data complexities.

6. EXPERIMENT AND SIMULATION RESULTS

Although the proposed framework is primarily theoretical, the framework's practical applicability is demonstrated through simulated case studies. Synthetic data with various imbalanced dataset typologies was generated:

- **Highly Imbalanced**: Imbalance ratio of 0.05 with low-dimensional features.
- **Moderately Imbalanced**: Imbalance ratio of 0.3 with low-dimensional features.
- **High-Dimensional Highly Imbalanced**: 100 features with an imbalance ratio of 0.05.
- **High-Dimensional Moderately Imbalanced**: 100 features with an imbalance ratio of 0.3.

All complexity metrics were computed for each type and show the results in PCA plots, probability histograms, kNN graphs, MRO heatmap, and bar charts. The simulation studies show a correlation between higher complexity measures, such as high IO or SO, and increased classification error, which supports this adaptive recommendations for more advanced resampling, dimensionality reduction, or the selection of robust classifiers.

Scenario of Highly Imbalanced (HI) datasets:

Table 3. Complexity metrics computed for the highly imbalanced dataset

Measure	Value	Description
F1	0.3475	Maximum Fisher's Discriminant Ratio value identify moderate discrimination.
F2	0.0035	Overlap region; lower values indicate minimal overlap.
IO	0.0290	Instance Overlap value indicates Low instance ambiguity.
SO	0.0636	Structural Overlap denoting low structural overlap.
MRO	**0.9444**	Multiresolution Overlap value indicates close to 1 means very high mixed clusters.
N1	0.1440	Moderate boundary ambiguity.
N3	0.0540	Below 0.1 indicate that local classification is relatively robust.
L2	0.0530	Values near 0 indicate good separability for a linear model.
CM	**0.8333**	Complexity Measure for minority instances: Values above ~0.8 imply a severe imbalance in local neighborhood composition.
wCM	**0.8155**	High values (above ~0.8) confirm that minority points are predominantly influenced by majority instances.
BI^3	0.2037	Bayes Imbalance Impact Index: Values around 0.2 suggest moderate impact.

Fig. 1 demonstrates a diverse viewpoint of the complexity present in the highly imbalanced dataset. The PCA estimation reveals significant overlap, with minority instances implanted within minority clusters, stipulating poor class separability. The KNN-predicted probability distribution demonstrates the majority of samples are predicted with low confidence, clustering near 0.0, emphasizing severe instance-level ambiguity. High CM and wCM scores on the complexity bar chart verify significant local imbalance. The kNN graph shows more dense inter-class connection, reflecting structural entanglement. The MRO heatmap, with values near 1, indicates mixed clustering across multiple resolution levels. Furthermore, the L2 error surface shows that even with class separability increasing, training error decreases on a small scale, recommending the inadequacy of linear models. Finally, the ROC and PR curves exhibit modest AUC values, validating the dataset's intrinsic difficulty and complying with the necessity for cost-sensitive or structure-aware learning methods.

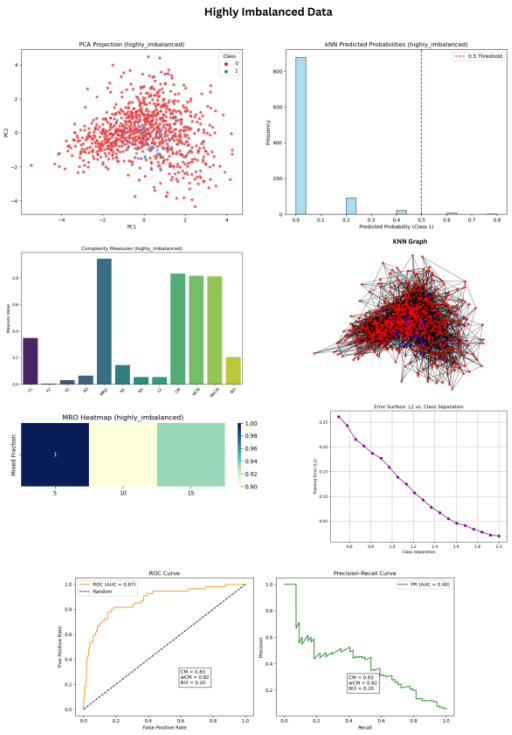

Fig. 1. Complexity analysis visualizations for the highly imbalanced dataset.

Framework-Driven Strategy Recommendations:

- Significant multiresolution overlap. Consider adaptive clustering-based re-sampling.
- High imbalance-specific metric (CM). Focus on minority-oriented re-sampling or cost-sensitive methods.

Case study of Moderately Imbalanced (MI) dataset:

Table 4. Complexity metric evaluated for moderately imbalanced dataset

Measure	Value	Description
F1	0.3929	Maximum Fisher's Discriminant Ratio indicates moderate discrimination (low < 0.2, moderate 0.2–0.5, high > 0.5).
F2	0.0039	Overlap volume surrogate; extremely low value indicates minimal overlap.
IO	0.1520	Instance Overlap; moderate ambiguity (low < 0.1, moderate 0.1–0.2, high > 0.2).
SO	0.1742	Structural Overlap; moderate intermingling (low < 0.1, moderate 0.1–0.3, high > 0.3).
MRO	**0.9556**	Multiresolution Overlap; very high value implies clusters are almost entirely mixed (values near 1 are worst).
N1	**0.4880**	Fraction of boundary points; nearly half the data lies on boundaries (lower is better; high > 0.4).
N3	0.1390	Leave-One-Out 1-NN error; moderate local classification error (low < 0.1, moderate 0.1–0.2).
L2	0.1520	Training error of a linear classifier; moderate error (lower values are preferable).
CM	**0.3535**	Complexity for minority samples; moderate vulnerability (low < 0.2, moderate 0.2–0.4, high > 0.4).
wCM	0.3423	Weighted complexity metric; similar to CM, indicating moderate risk.
dwCM	0.3392	Dual weighted complexity metric; also indicates moderate vulnerability.
BI^3	0.2874	Bayes Imbalance Impact Index; moderate impact (values near 0.5 would be worse).

Fig. 2 provides a circumstantial complexity analysis for a moderately imbalanced dataset. The PCA plots clearly show a more distinct inter-class separation than the highly imbalanced scenario, with reduced overlap. A wider predicted probability histogram indicates better confidence in classification and lower instance ambiguity. Reflecting moderate values across IO, SO, and CM, the complexity metric bar chart implies improved class boundary clarity. The kNN graph reveals minor cross-class connections, while the MRO heatmap reflects less frequent mixed clustering—both signifying improved structural purity. Correspondingly, the L2 error surface exhibits more consistent declines with increasing class separation, indicating better linear separability. ROC and PR curves demonstrate improved classifier performance with rising AUC values and affirm that moderate imbalance is more controllable with standard preprocessing and learning techniques.

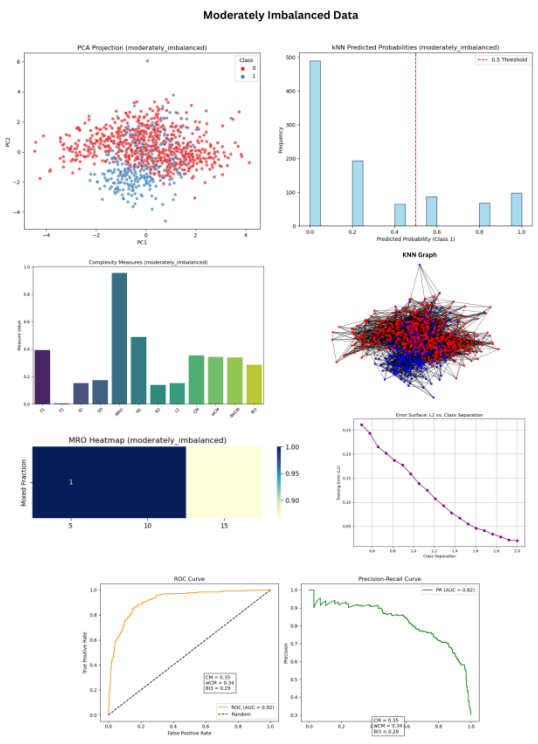

Fig. 2. Complexity analysis visualizations for the moderately imbalanced dataset

Framework-Driven Strategy Recommendations:

- Significant multiresolution overlap. Consider adaptive clustering-based re-sampling.

High-dimensional with Highly Imbalanced dataset:

Table 5. Complexity metrics computed for the high-dimensional, highly imbalanced dataset

Measure	Value	Description
F1	0.2113	Moderate discrimination; (Low: <0.2, Moderate: 0.2–0.5, High: >0.5)
F2	0.0021	Very low overlap; indicates minimal density overlap
IO	0.0250	Low instance ambiguity; (Low: <0.1, Moderate: 0.1–0.2, High: >0.2)
SO	0.0772	Low structural overlap; (Low: <0.1, Moderate: 0.1–0.3, High: >0.3)
MRO	**0.9333**	Extremely high mixed-cluster fraction; values near 1 denote severe mixing
N1	0.1800	Moderate boundary presence; (Low: <0.1, Moderate: 0.1–0.3, High: >0.3)
N3	0.0760	Low leave-one-out NN error; (Low: <0.1, Moderate: 0.1–0.2)
L2	0.0180	Very low linear classifier error; indicates excellent separability
CM	**0.9091**	Very high minority risk; (Low: <0.2, Moderate: 0.2–0.4, High: >0.4)
wCM	**0.9077**	Very high weighted minority risk; similar interpretation as CM
dwCM	**0.9076**	Very high dual weighted risk; aligns with CM and wCM
BI^3	0.2345	Moderate imbalance impact; (Low: <0.2, Moderate: 0.2–0.4, High: >0.4)

Fig. 3 visualizes the combined complexity, deriving both high dimensionality and extreme class imbalance. The PCA projection sounds noisy with major class overlap, indicating that high-dimensional space exacerbates feature confusion. Additionally, the predicted probability distribution, with an acute concentration around 0.5, reflects extreme classification uncertainty. The bar chart displays peak values of CM, SO, and MRO, signifying that minority instances are structurally saturated and boundary ambiguity is critical. The kNN graph exhibits complex inter-class connections, further confirming the presence of structural overlap. Consequently, the MRO heatmap near 1 across scales, magnifying persistent class mixing regardless of granularity. Notwithstanding more class separation, the L2 surface reveals little error decrease, suggesting that the decision boundaries are inherently non-linear and in high-dimensional space. Correspondingly, the ROC and PR curves are lacking, emphasizing the importance of dimensionality reduction, adaptive clustering, and ensemble learning techniques.

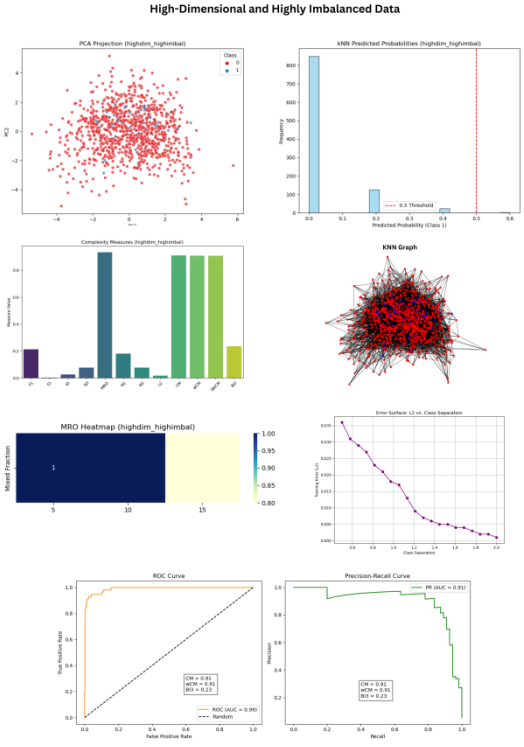

Fig. 3. Complexity analysis visualizations for the high-dimensional highly imbalanced dataset.

Framework-Driven Strategy Recommendations:

- Significant multiresolution overlap. Consider adaptive clustering-based re-sampling.
- High imbalance-specific metric (CM). Focus on minority-oriented re-sampling or cost-sensitive methods.

Scenario of High-dimensional with Moderately Imbalanced datasets:

Table 6. Complexity metrics computed for the high-dimensional, moderately imbalanced dataset.

Measure	Value	Description
F1	0.3218	Moderate discrimination.
F2	0.0032	Minimal overlap; (Very Low: near 0, Low: <0.005, Moderate: 0.005–0.01, High: >0.01).
IO	**0.2990**	High instance ambiguity.
SO	0.3108	High structural overlap.
MRO	**1.0000**	Extremely high mixed-cluster fraction; values near 1 denote severe mixing.
N1	**0.6970**	Very high boundary ambiguity.
N3	**0.2960**	High local error.
L2	0.1320	Moderate linear classifier error; (Low: <0.1, Moderate: 0.1–0.2, High: >0.2).
CM	0.6317	Moderate minority vulnerability.
wCM	0.6285	Moderate weighted risk; similar scale as CM.
dwCM	0.6283	Moderate dual-weighted risk; consistent with wCM.
BI^3	0.1693	Low–moderate imbalance impact.

Fig. 4 exhibits the complexity of a high-dimensional, moderately imbalanced dataset. The estimation of PCA reveals partial class separation, with residual noise still present but comparatively less severe than in the high-dimensional and high-imbalanced scenario. The distribution of probabilities in the histogram is more uniformly distributed, signifying less ambiguity at the instance level. The moderate CM and SO scores shown in the bar chart suggest that structural entanglement and minority class vulnerability are manageable. Correspondingly, the KNN graph exemplifies lower cross-class connectivity. The MRO heatmap demonstrates better cluster purity relative to more imbalanced settings. A significant decrease in the L2 error surface recommends improved potential for linear separation. Finally, ROC and PR curves with comparatively high AUC values affirm that, despite the high dimensionality imposing inherent complexity, the moderate imbalance setting permits efficient separation with appropriate preprocessing and model tuning.

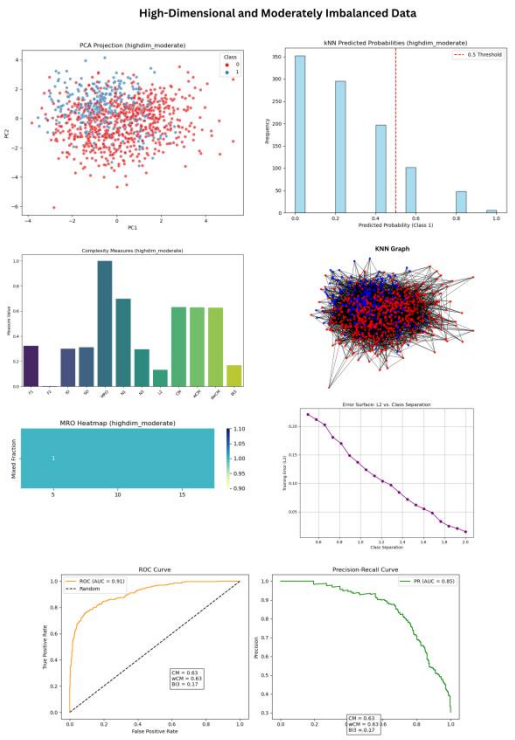

Fig. 4. Complexity analysis visualizations for the high-dimensional moderately imbalanced dataset.

Framework-Driven Strategy Recommendations:

- High structural overlap observed. Graph-based or structural regularization methods may help.
- Significant multiresolution overlap. Consider adaptive clustering-based re-sampling.
- High leave-one-out NN error (N3). Indicates local boundary complexity.
- High imbalance-specific metric (CM). Focus on minority-oriented re-sampling or cost-sensitive methods.

7. RESULTS AND DISCUSSION

To exhibit the efficacy of the proposed framework, a series of controlled simulations were used over four representative dataset scenarios: (i) Highly Imbalanced (HI), (ii) Moderately Imbalanced (MI), (iii) High-Dimensional Highly Imbalanced (HD-HI), and (iv) High-Dimensional Moderately Imbalanced (HD-MI). These datasets were synthetically constructed to imitate realistic imbalance ratios and high-dimensional settings. The proposed framework was systematically applied to each dataset to figure out data complexity measures, enabling comparative insights across the experimental conditions.Traditional metrics such as the Maximum Fisher's Discriminant Ratio (F1) and its derived overlap surrogate (F2) indicated that, even when individual features appeared moderately discriminative, the joint feature space often exhibited substantial overlap between classes. Moreover, novel measures like Instance Overlap (IO), Multiresolution Overlap (MRO), and Structural Overlap (SO) consistently highlighted that a high proportion of instances lie in ambiguous regions and that the underlying structure (as revealed by kNN graphs) is highly intermingled. Neighborhood metrics (e.g., N1 and N3) further supported the observation that many data points reside close to the decision boundaries, leading to elevated leave-one-out errors. Furthermore, linear separability measures (L2)

signified those basic classifiers scuffling on these imbalanced datasets, and imbalance-oriented metrics (CM, WCM, BI3) underlined the vulnerability of the minority class. These observations highlight the need for adaptive techniques shaped to the specific characteristics of the dataset, ranging from advanced resampling and cost-sensitive processes to advanced feature engineering and dimensionality reduction.

7.1 Complexity Metrics Evaluation

Each dataset was evaluated using the full set of complexity measures described in the proposed framework. Fig. 5 visualizes the comparison of key complexity metrics over the four dataset scenarios (HI, MI, HD-HI, and HD-MI). This grouped bar chart makes it easy to read how CM, MRO, IO, and N1 metrics differ by dataset types. As an example, it points out the very high CM and MRO in HI and HD-HI cases and the significantly higher IO and SO in HD-MI. This diagram adds to Tables 3 to 6 to provide an overview of the progress of structural and local complexity, as well as imbalance complexity, across various properties of the input datasets.

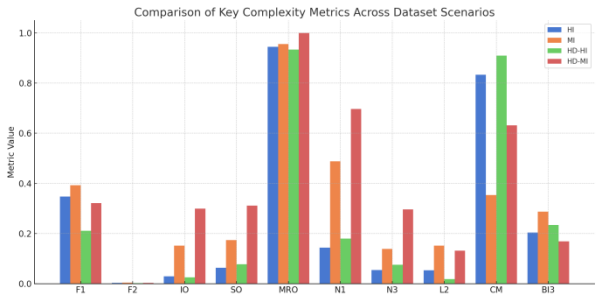

Fig. 5. Grouped bar chart comparing key complexity metrics across four dataset scenarios.

These results clearly show a strong variation of data complexity among different dataset typologies.

Highly Imbalanced Dataset (HI):
In the Highly Imbalanced (HI) dataset, while feature overlap (F2) is very low, complexity arises from the extremely high class imbalance (CM = 0.8333) and substantial multiresolution overlap (MRO = 0.9444). These indicate that most minority class instances are structurally dominated and appear in mixed clusters, making learning difficult.

Recommendation (Framework Output): Apply ensemble learning with SMOTE variants or cost-sensitive learning. To address class impurity and neighborhood ambiguity, consider graph-based structural modeling.

Moderately Imbalanced Dataset (MI):
Although the Moderately Imbalanced (MI) dataset exhibits moderate overlap and structural complexity On the basis of IO and SO values. The relatively high N1 (0.4880) and N3 (0.1390) values indicate that nearly half of the instances lie near decision boundaries. Therefore the imbalance is lower than HI, boundary complexity is relatively high.

Recommendation (Framework Output): Use neighborhood-sensitive classifiers in combination with moderate resampling. Integrate margin-based regularization to enhance the boundary clarity and generalization ability of the classifier.

High-Dimensional Highly Imbalanced Dataset (HD-HI):
For High-Dimensional and Highly Imbalanced (HD-HI) dataset, the framework enclosed the HD-HI-persuade challenges. The extremely high CM (0.9091), MRO (0.9333) and a very low F1 (0.2113) under the fact of less feature overlapping indicate high complexity even though the feature overlap is minimum.

Recommendation (Framework Output): Employ resampling followed by dimensionality reduction (such as PCA or t-SNE). Incorporate ensemble models designed for high-dimensional imbalance or structural learning models such as graph-based classifiers.

High-Dimensional Moderately Imbalanced Dataset (HD-MI):
On the other hand, dataset High-Dimensional Moderated Imbalance (HD-MI) had the maximum values of IO (0.2990), SO (0.3108), and N1 (0.6970) which indicate that a moderate imbalance with high dimensionality shows the highest boundary.

Recommendation (Framework Output): Apply complexity-aware feature selection along with dimensionality reduction. Make use of regularized ensemble classifiers that manage structural overlap and prioritize margin preservation.

7.2 Summary of Observations

The following key observations emerge from the analysis:

- The Multiresolution Overlap (MRO) was consistently high across datasets, suggesting that class impurity is scale-enabled and must be accounted for within imbalanced learning.
- In datasets HI and HD-HI, Imbalance-Specific Metrics (CM, wCM, BI3) outperformed others, revealing the weakness of minority classes.
- The Local Neighborhood Complexity (N1, N3) was more pronounced in MI and HD-MI datasets indicating that boundary ambiguity seems to play a greater role in MI and HD-MI.
- High-Dimensional Datasets provided more noise and overlap of structure, and the need for reducing dimensions.
- The recommendations of the framework — ranging from aggressive re-sampling to graph-based structural learning-match with the complexity profile found in each case.

This shows validate the utility of the proposed framework in diagnosing sources of complexity, which if triggered, the appropriate preprocessing or model selection strategy can be guided to follow for imbalanced learning.

8. CONCLUSION AND FUTURE WORK

This study proposes an adaptive framework for analyzing data complexity in imbalanced binary classification problems. Derived indepth from conventional as well as advanced complexity measures — specifically feature overlap (F1–F4), instance ambiguity (IO), multiresolution overlap (MRO), structural overlap (SO) and minority vulnerability (CM, wCM, BI3) — the framework offers a multi-dimensional way to characterize classification difficulty. The framework recognizes the specific complexity features associated with specific dataset scenarios (i.e. high degree imbalanced or moderate level imbalanced or high dimension etc.) and thereby accordingly adapts its recommendations to guide suitable preprocessing techniques and model selection according to the data.

Theoretical contributions such as Lemma 1 and Corollary 1

constitute a formal relationship between local complexity measures and classification error bounds, contributing a solid mathematical foundation. The adaptive measure selection algorithm provides additional robustness to the applicability of this framework, assuring that model development need not be one-size-fits-all but rather driven by immediate complexity diagnostics. Controlled simulation results confirm that datasets with comparable imbalance ratios may actually possess diverse underlying complexities, and thus different learning strategies are required to achieve optimally performant models — a nuance often missed in conventional imbalance treatment approaches.

Future Scope:
Even though this study only looked at binary classification on datasets that were artificially controlled, there are many other areas that could be studied in the future:

Real-world validation: Applying the framework to real-world datasets from different domains, such as healthcare, fraud detection, cybersecurity, and text classification, will compute its effectiveness and robustness in practical scenarios.

Multi-Class Extension: Add multi-class imbalanced learning to the framework so that problems can happen between more than two minority and majority classes.

Integration into AutoML Pipelines: Incorporating complexity computation and a recommendation engine into automated machine learning pipelines (AutoML) may afford dynamic model selection and preprocessing during training.

Dynamic Learning Systems: Evolving real-time adaptive learning systems capable of modifying sampling strategies and classifier parameters on-the-fly based on evolving complexity metrics during data stream learning.

Graph-based Complexity Modeling: Looking into how graph neural networks and topological data analysis can be used to make structural complexity modeling better than typical kNN-based methods.

9. REFERENCES

[1] Chen, W., Yang, K., Yu, Z. et al. A survey on imbalanced learning: latest research, applications and future directions. Artif Intell Rev 57, 137 (2024). https://doi.org/10.1007/s10462-024-10759-6.

[2] Victor H. Barella, Luís P.F. Garcia, Marcilio C.P. de Souto, Ana C. Lorena, André C.P.L.F. de Carvalho, Assessing the data complexity of imbalanced datasets, Information Sciences, Volume 553, 2021, Pages 83-109, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2020.12.006.

[3] Xiaohui Wan, Zheng Zheng, Fangyun Qin, and Xuhui Lu. 2024. Data Complexity: A New Perspective for Analyzing the Difficulty of Defect Prediction Tasks. ACM Trans. Softw. Eng. Methodol. 33, 6, Article 141 (July 2024), 45 pages. https://doi.org/10.1145/3649596.

[4] Tin Kam Ho and M. Basu, "Complexity measures of supervised classification problems," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 289-300, March 2002, doi: 10.1109/34.990132.

[5] Ho, T.K., Basu, M., Law, M.H.C. (2006). Measures of Geometrical Complexity in Classification Problems. In: Basu, M., Ho, T.K. (eds) Data Complexity in Pattern Recognition. Advanced Information and Knowledge Processing. Springer, London. https://doi.org/10.1007/978-1-84628-172-3_1.

[6] Lorena, A.C., de Souto, M.C.P. (2015). On Measuring the Complexity of Classification Problems. In: Arik, S., Huang, T., Lai, W., Liu, Q. (eds) Neural Information Processing. ICONIP 2015. Lecture Notes in Computer Science(), vol 9489. Springer, Cham. https://doi.org/10.1007/978-3-319-26532-2_18.

[7] Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. 2019. How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. ACM Comput. Surv. 52, 5, Article 107 (September 2020), 34 pages. https://doi.org/10.1145/3347711.

[8] Nafees Anwar, Geoff Jones, and Siva Ganesh. 2014. Measurement of data complexity for classification problems with unbalanced data. Stat. Anal. Data Min. 7, 3 (June 2014), 194–211.

[9] Singh, Deepika, et al. "Weighted k -nearest neighbor based data complexity metrics for imbalanced datasets." Statistical Analysis and Data Mining, vol. 13, no. 4, Jun. 2020, pp. 394-404. https://doi.org/10.1002/sam.11463.

[10] Y. Lu, Y. -M. Cheung and Y. Y. Tang, "Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 9, pp. 3525-3539, Sept. 2020, doi: 10.1109/TNNLS.2019.2944962.

[11] Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, João Santos, A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research, Information Fusion, Volume 89, 2023, Pages 228-253, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.08.017.

[12] Pattaramon Vuttipittayamongkol, Eyad Elyan, Andrei Petrovski, On the class overlap problem in imbalanced data classification, Knowledge-Based Systems, Volume 212, 2021, 12., 106631, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.106631.

[13] 13.E. B. Fatima, B. Omar, E. M. Abdelmajid, F. Rustam, A. Mehmood and G. S. Choi, "Minimizing the Overlapping Degree to Improve Class-Imbalanced Learning Under Sparse Feature Selection: Application to Fraud Detection," in IEEE Access, vol. 9, pp. 28101-28110, 2021, doi: 10.1109/ACCESS.2021.3056285.

[14] 14.Komorniczak, Joanna et al. "Data complexity and classification accuracy correlation in oversampling algorithms." Learning with Imbalanced Domains: Theory and Applications (2022), Proceeding of Machine Learning Research.