



VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
INSTITUTE OF COMPUTER SCIENCE
DEPARTMENT OF COMPUTATIONAL AND DATA MODELING

Computer Modeling Final Master Thesis

Handling Class Imbalance in k-Nearest Neighbors

Klasių Disbalanso sprendimas k Artimiausių Kaimynų Klasifikatoriui

Done by:

Jovita Matvejevaitė

Supervisor:

prof. Gintautas Dzemyda

Vilnius
2026

Contents

Abstract	4
Santrauka	5
Introduction	6
1 Literature Review	7
1.1 Cost-Sensitive kNN Classification	7
1.2 Neighborhood Distributions for Imbalanced Classification	8
1.3 Optimal Binary Classification Beyond Accuracy	9
1.4 Prior Probability Balancing in kNN Classification	10
1.5 Summary	11
2 Methodology	12
2.1 kNN	12
2.2 SMOTE	12
2.3 STIR-kNN	13
3 Experimental Methodology	16
3.1 Datasets	16
3.1.1 Numerical Dataset	16
3.1.2 Heterogeneous Dataset	17
3.1.3 Time-dependent Historical Dataset	18
3.1.4 Summary	19
3.2 Data Preprocessing	20
3.3 Evaluation Metrics	20
3.4 Experimental Setup	21
4 Results and Analysis	22
4.1 Experimental Results	22
4.1.1 Electrical Grid Stability	23
4.1.2 Census Income	26
4.1.3 Bank Marketing	29
4.2 Ablation Study: Contribution of STIR-kNN Components	32
4.2.1 Electrical Grid Stability	33
4.2.2 Census Income	34
4.2.3 Bank Marketing	35
4.3 Summary	37
Conclusions and Recommendations	39
Future Works	41
Declaration of Artificial Intelligence	42
References	43

A	Appendix	45
A.1	STIR-kNN	45

Abstract

This work introduces Stochastic Inference-Time Rebalancing kNN (STIR-kNN), an inference-time modification of the k-nearest neighbor classifier for imbalanced binary classification. The proposed method aims to improve minority-class recall while limiting the corresponding loss in majority-class precision, without modifying the training data. STIR-kNN combines adaptive neighborhood expansion with a stochastic vote-adjustment mechanism that selectively rebalances local class influence when minority-class evidence is detected. The approach is evaluated against standard kNN and kNN trained on SMOTE-augmented data across datasets with varying imbalance ratios, including numerical, heterogeneous, and time-dependent historical data. Experimental results demonstrate that STIR-kNN consistently enhances minority-class recall while maintaining competitive overall performance, highlighting the effectiveness of inference-time rebalancing for imbalanced kNN classification. These findings suggest that inference-time neighborhood rebalancing is an effective and flexible strategy for addressing class imbalance in kNN classifiers. STIR-kNN demonstrates that substantial gains in minority detection can be achieved without re-training or synthetic data generation, making it well-suited for applications where data integrity and deployment simplicity are critical.

Santrauka

Klasių Disbalanso sprendimas k Artimiausių Kaimynų Klasifikatoriui

Šiame darbe pristatomas stochastinis perbalansavimas išvadų darymo metu k-artimiausių kaimynų klasifikatoriui (STIR-kNN). Šis metodas skirtas binarinės klasifikacijos rezultatams koreguoti išvadų darymo etape. Siūlomu metodu siekiama, nekeičiant mokymo duomenų, padidinti mažumos klasės atpažinimą, kartu ribojant atitinkamą daugumos klasės precizijos praradimą. Eksperimentams naudoti duomenų rinkiniai su skirtingais disbalanso koeficientais, apimantys skaitmeninius, heterogeninius ir nuo laiko priklausomus istorinius duomenis. Rezultatai rodo, kad STIR-kNN nuosekliai didina mažumos klasės atpažinimą, kartu išlaikydamas konkurencingą bendrąjį našumą. Šios išvados pabrėžia, kad kaimynystės perbalansavimas išvadų darymo etape yra efektyvi ir lanksti strategija sprendžiant klasių disbalanso problemas kNN klasifikatoriuose. STIR-kNN įrodo, kad žymų mažumos klasės aptikimo pagerėjimą galima pasiekti be pakartotinio modelio mokymo ar sintetinių duomenų generavimo. Dėl šios priežasties metodas yra itin tinkamas taikymams, kuriuose kritiškai svarbus duomenų vientisumas.

Introduction

The k-Nearest Neighbors (kNN) algorithm is a well-established, distance-based method frequently employed for classification tasks, because of its simplicity and minimal requirements for setup. The kNN can also be used for regression tasks but in this work, we will solely focus on the classification problem. In this setting, kNN assigns a class label to an instance based on majority voting among its k nearest neighbors, as determined by a distance metric. Despite being the popular choice, particularly in scenarios involving small to medium-sized datasets with clearly separable classes, it has some flaws, one of its limitations is high dimensional data. The performance of kNN deteriorates significantly in high-dimensional spaces due to a phenomenon known as the curse of dimensionality [13]. The term "curse of dimensionality" was first introduced by Richard E. Bellman in 1957 [2], referring to the exponential increase in volume associated with adding extra dimensions to Euclidean space [17].

Furthermore, kNN is highly sensitive to the class imbalance [21][19]. Because the algorithm relies on majority voting, it implicitly assumes equal class priors. Under severe class imbalance, this assumption results in majority-class dominance and poor recall for minority classes.

To mitigate these effects, researchers have introduced various strategies that modify the kNN classifier or its training data. One such approach is Weighted kNN (WkNN), which assigns each neighbor a contribution proportional to its distance from the query instance, thereby giving greater influence to closer neighbors [6]. Despite this adjustment, weighted variants remain highly sensitive to noise, the choice of distance metrics, and hyperparameter tuning, and they do not consistently outperform standard kNN across diverse datasets [9].

Another prominent class of methods relies on training-time data augmentation, most notably the Synthetic Minority Over-sampling Technique (SMOTE) [4]. SMOTE addresses imbalance by generating synthetic minority instances to achieve class parity in the training set. However, this strategy may introduce noisy or unrealistic samples, potentially leading to over-fitting and degraded generalization performance [1].

Taken together, these approaches highlight the limitations of training-time modification strategies, which either rely on static distance weighting or alter the data distribution itself, without addressing how imbalance manifests during the inference process.

In this research, we introduce a novel framework that shifts the corrective intervention for class imbalance from the training phase to the inference phase — Stochastic Inference-Time Rebalancing kNN (STIR-kNN). Our approach constitutes an inference-time modification of the standard kNN classifier, specifically designed for imbalanced binary classification. The primary objective is to substantially enhance minority-class recall while minimizing the corresponding degradation in majority-class precision.

The proposed model targets application domains characterized by asymmetric misclassification costs, where failing to detect a minority-class instance (a false negative) entails significantly higher operational or safety risk than producing a false positive [7, 11]. The motivation for this design aligns with prior work on cost-sensitive classification, including cost-sensitive variants of kNN [20], which explicitly model asymmetric misclassification costs to penalize false negatives more heavily than false positives. Such approaches have demonstrated that uniform neighbor voting is suboptimal when misclassification costs are unequal. However, existing cost-sensitive kNN methods typically embed cost information directly into the decision rule or training process, resulting in deterministic reweighting schemes that remain static across inference contexts.

Beyond cost-sensitive formulations, prior work has demonstrated that the local neighborhood

structure of k-nearest neighbor classifiers is itself strongly affected by class imbalance. Kriminger et al. [14] analyze class-conditional nearest neighbor distance distributions and demonstrate that fixed-size neighborhoods often fail to capture sparse minority-class evidence, even when relevant samples exist nearby in feature space. Their findings indicate that raw neighbor counts and globally fixed values of k can lead to decision boundaries that favor majority classes due to differences in class density rather than true class separability. Crucially, this limitation arises not from the absence of minority-class samples, but from the restriction imposed by a globally fixed neighborhood size, suggesting that the neighborhood itself must be allowed to expand adaptively in order to reveal sparse minority-class evidence.

Motivated by these observations, STIR-kNN introduces an elastic neighborhood mechanism that relaxes the use of a globally fixed neighborhood size. Rather than committing to a predetermined value of k , the method employs an adaptive neighborhood expansion (k_{elastic}) that activates only when minority-class evidence is detected. This expansion scales sub-linearly with the global imbalance ratio, allowing the classifier to explore a broader local context when necessary while preserving locality in majority-dominated regions.

In contrast to oversampling-based approaches such as SMOTE, STIR-kNN preserves the original training data and the standard kNN decision rule, introducing instead a stochastic post-processing mechanism applied exclusively at inference time. When corrective intervention is triggered within the elastic neighborhood, local class influence is rebalanced via stochastic vote adjustment rather than deterministic cost weighting or synthetic data generation.

This formulation aligns with recent theoretical frameworks established by Singh and Khim [18], which demonstrate that stochastic classifiers may be required to optimize population-level utilities, such as recall, beyond global accuracy. Furthermore, research into prior-probability balancing [8] suggests that modifying local class influence is more effective than relying on raw neighbor counts. Our framework integrates these insights through a stochastic label-injection mechanism that scales dynamically with the global imbalance ratio via a sub-linear logarithmic function. This provides a robust, inference-time correction that bypasses the computational overhead of retraining and avoids the potential pitfalls of synthetic sample generation.

1 Literature Review

1.1 Cost-Sensitive kNN Classification

Cost-sensitive learning addresses classification problems in which different types of misclassification errors incur unequal costs, a situation that frequently arises in imbalanced data applications such as medical diagnosis and risk assessment. Rather than minimizing classification error, cost-sensitive methods aim to minimize the expected misclassification cost. Zhang extends the k-nearest neighbor classifier to explicitly incorporate misclassification costs into the decision process [20].

The cost-sensitive framework assumes the availability of a cost matrix, where each entry $C(i, j)$ represents the cost of predicting a sample as class i when its true class is j . Given a query instance x , the classification decision is based on minimizing the expected cost rather than maximizing posterior probability. This principle is formalized through the conditional risk, defined as:

$$L(x, i) = \sum_j P(c_j | x) C(i, j), \quad (1)$$

where $P(c_j | x)$ denotes the estimated posterior probability of class c_j given x . The predicted class is then selected as the one that minimizes the conditional risk.

Zhang proposes two cost-sensitive kNN variants that differ in how cost information is integrated. In the Direct cost-sensitive kNN approach, posterior class probabilities are estimated from the class distribution of the k nearest neighbors, and the final prediction is obtained by directly minimizing the expected misclassification cost using the cost matrix. This approach preserves the standard neighborhood structure of kNN while modifying the decision criterion to reflect asymmetric costs.

The Distance-based cost-sensitive kNN method incorporates misclassification costs into the neighbor selection and voting process by weighting neighbors according to both their distance and the associated cost of misclassification. This allows the classifier to favor neighbors that contribute less to the expected cost, thereby adapting the local neighborhood influence based on application-specific cost considerations.

The effectiveness of cost-sensitive kNN methods has been evaluated empirically on a range of real-world datasets. Performance is assessed primarily using misclassification cost, alongside the area under the ROC curve (AUC), under varying cost ratios that reflect asymmetric penalties for false negatives and false positives. Experimental results reported by Zhang et al.[20] show that cost-sensitive kNN variants outperform a cost-sensitive C4.5 decision tree baseline [16], particularly as the cost of misclassifying minority-class instances increases. These findings indicate that explicitly modeling asymmetric misclassification costs at the decision stage enables kNN classifiers to better align predictions with application-specific risk profiles, especially in highly imbalanced settings.

Importantly, these cost-sensitive kNN approaches incorporate misclassification costs during inference, influencing both distance weighting and the final decision rule, while avoiding any resampling or modification of the training data.

1.2 Neighborhood Distributions for Imbalanced Classification

Kriminger et al. investigate the effects of class imbalance on local neighborhood structure in nearest neighbor classification and propose a novel method, termed Class Conditional Nearest Neighbor Distribution (CCNND), that exploits class-specific distance distributions to improve minority-class detection [14]. Rather than modifying the training data through resampling or incorporating explicit misclassification costs, their approach relies on the geometric properties of the data within each class to guide the classification decision.

The core motivation of CCNND is that, in imbalanced datasets, minority-class samples are typically more sparsely distributed in feature space than majority-class samples. As a result, nearest neighbor distances within the minority class tend to be larger and more variable. Traditional kNN classifiers compare distances across classes directly, which leads to decision boundaries that favor majority classes. Kriminger and Príncipe argue that nearest neighbor distances should instead be interpreted relative to the typical distance distributions observed within each class.

Formally, let the training set be composed of L classes, denoted by $\mathcal{C} = \{C_i\}_{i=1}^L$, where each class C_i contains N_i samples in \mathbb{R}^m . For a given query sample x , let $\mathbf{d}_i(x)$ denote the vector of distances from x to its k nearest neighbors belonging to class C_i . During training, the algorithm computes $\mathbf{d}_i(x)$ for all samples $x \in C_i$ and constructs an empirical cumulative distribution function (CDF) of nearest neighbor distances for each class.

At inference time, for a test sample x_{test} , the corresponding distance vector $\mathbf{d}_i(x_{\text{test}})$ is computed

for each class. The algorithm then evaluates the probability that a point with nearest neighbor distances greater than those of x_{test} would be observed within class C_i . This probability is estimated empirically as:

$$p_i(x_{\text{test}}) = \frac{|S_i|}{|C_i|}, \quad (2)$$

where

$$S_i := \{x_j \in C_i \mid \mathbf{d}_i(x_{\text{test}}) < \mathbf{d}_i(x_j)\}. \quad (3)$$

The test sample is assigned the label corresponding to the class with the highest probability:

$$\hat{y}(x_{\text{test}}) = \arg \max_{i \in \{1, \dots, L\}} p_i(x_{\text{test}}). \quad (4)$$

By comparing nearest neighbor distances relative to class-conditional distance distributions rather than directly across classes, CCNND adjusts classification decisions to account for differences in class density and sparsity. This formulation allows the method to adapt automatically to varying degrees of imbalance and extends naturally to multiclass settings.

The authors evaluate CCNND on several imbalanced datasets. Performance is assessed using sensitivity (i.e., recall of the minority class), specificity (true negative rate) and the G-mean metric. Experimental results demonstrate that CCNND achieves competitive or superior G-mean and sensitivity compared to state-of-the-art sampling-based and cost-sensitive methods, such as SMOTE with different costs (SDC), while significantly outperforming standard kNN and support vector machine classifiers [14]. Notably, CCNND maintains high true positive rates for minority classes without requiring explicit cost tuning or synthetic data generation.

In addition, the method is evaluated on a real-world, multi-class online streaming dataset from the oil and gas industry. In this setting, CCNND outperforms both standard kNN and neighbor-weighted kNN (NWkNN) in terms of overall error rate and minority-class detection, highlighting its suitability for dynamic environments where class imbalance and class prevalence may change over time. These results underscore the importance of local neighborhood structure and class-conditional distance distributions in addressing imbalanced nearest neighbor classification.

1.3 Optimal Binary Classification Beyond Accuracy

Most theoretical results in binary classification characterize performance exclusively in terms of accuracy. However, accuracy is often an inadequate metric in imbalanced classification settings, as it may obscure poor performance on the minority class. Singh and Khim address this limitation by developing a general theoretical framework for optimal binary classification under performance measures derived from the confusion matrix, such as precision, recall, and F -scores [18].

The authors generalize the classical Bayes-optimal classifier, which is known to maximize accuracy, to a broader class of performance measures referred to as confusion matrix measures. These measures capture the trade-offs between different types of classification errors and better reflect practical evaluation criteria used in imbalanced learning. A key insight of this work is that, under general performance measures, an optimal classifier may not be deterministic. Instead, stochastic classifiers, those that randomize predictions for certain inputs, are sometimes necessary to achieve optimal performance.

An important contribution of the research is the removal of restrictive distributional assumptions that appear in prior theoretical analyses, such as absolute continuity of class probability

distributions. By allowing stochastic decision rules, the authors show that optimal classifiers can be characterized without relying on such assumptions. This result provides a more realistic theoretical foundation for imbalanced classification, particularly in settings where class probabilities may concentrate on specific regions of the feature space.

Building on this framework, Singh and Khim analyze the behavior of the k -nearest neighbor classifier under alternative performance measures. Rather than focusing on misclassification risk, their analysis connects the performance of k NN-based classifiers to the accuracy of regression function estimation under uniform loss. The authors further introduce the concept of *Uniform Class Imbalance*, a specific imbalance regime in which the minority class remains rare across the entire feature space. Under this regime, they show that the optimal choice of the neighborhood size parameter k differs substantially from that in balanced classification problems.

Overall, this work provides a theoretical justification for evaluating and tuning classifiers, including k NN, using performance measures beyond accuracy. By linking optimal classification decisions to both the choice of evaluation metric and the structure of class imbalance, it offers a rigorous context for understanding classifier behavior in imbalanced binary classification tasks.

1.4 Prior Probability Balancing in k NN Classification

Gøttcke and Zimek propose an internal modification of the k NN classifier that addresses class imbalance by reformulating the standard majority-vote decision rule within a probabilistic framework, referred to as k -Nearest Neighbors with Balanced Prior Probabilities (k NN-BPP) [8]. The proposed method operates directly at the decision stage of the classifier and does not require any modification of the training data or the use of resampling techniques.

The authors reinterpret k NN classification using Bayes' theorem. For a class c_i , the posterior probability given a query point x is defined as:

$$P(c_i | x) = \frac{P(x | c_i) P(c_i)}{\sum_{j=1}^m P(x | c_j) P(c_j)}, \quad (5)$$

where m denotes the number of classes, $P(x | c_i)$ is the class-conditional likelihood, and $P(c_i)$ represents the prior probability of class c_i . In the conventional k NN setting, these prior probabilities are implicitly estimated from the training data and therefore reflect the global class distribution. As a result, classification decisions are biased toward majority classes in imbalanced datasets.

To mitigate this effect, the authors propose enforcing uniform prior probabilities across all classes:

$$P(c_i | x) = \frac{P(x | c_i) \frac{1}{m}}{\sum_{j=1}^m P(x | c_j) \frac{1}{m}}, \quad (6)$$

thereby removing the influence of global class frequencies from the decision process. This probabilistic reformulation ensures that all classes are treated equally from a prior perspective and alters the behavior of the classifier in local neighborhoods.

By balancing prior probabilities, the classifier places greater emphasis on how well a class is represented locally relative to its overall presence in the dataset. This shift increases sensitivity to minority classes and promotes improved recall in imbalanced classification settings. The authors further show that this behavior cannot, in general, be replicated by static class-weighted k NN approaches, as such methods do not adapt to local neighborhood composition.

The authors evaluate the proposed kNN-BPP classifier on a broad collection of real-world and synthetic datasets with numerical attributes, exhibiting class imbalance degree from mild to extreme. Performance is assessed using recall-oriented measures, in particular macro-averaged recall and the geometric mean (G-mean), reflecting the emphasis on minority-class sensitivity. The proposed method is compared against the standard kNN classifier, several deterministic decision-level kNN variants designed for imbalanced data (including class-weighted and cost-sensitive formulations), as well as data-level imbalance handling techniques such as random undersampling (RUS) and SMOTE.

Across a wide range of neighborhood sizes, kNN-BPP demonstrates consistent performance gains over the standard kNN classifier and other deterministic, decision-level kNN variants. Moreover, its performance is shown to be comparable to, and in some cases surpass, that of oversampling-based approaches such as SMOTE, while preserving the computational complexity of conventional kNN. Additional experiments on synthetic datasets with multimodal minority-class distributions indicate that kNN-BPP is less susceptible to the performance degradation observed for oversampling methods, underscoring the robustness of prior-probability balancing when synthetic sample generation may introduce spurious artifacts [8].

1.5 Summary

The reviewed literature presents several complementary perspectives on addressing class imbalance in k -nearest neighbor classification and related binary classification tasks. A common theme across these works is the recognition that standard accuracy-based decision rules are often insufficient in imbalanced settings, motivating alternative objectives and decision mechanisms.

Cost-sensitive approaches provide a practical framework for handling imbalanced classification problems by explicitly accounting for unequal misclassification costs. By minimizing expected cost rather than classification error, cost-sensitive kNN methods adapt the decision process to application-specific priorities while preserving the instance-based structure of the classifier. These approaches are particularly relevant in domains where different types of errors have clearly defined consequences.

In addition to cost-based formulations, prior work has shown that class imbalance fundamentally affects the local neighborhood structure exploited by kNN classifiers. Analyses of class-conditional nearest neighbor distributions demonstrate that fixed-size neighborhoods often fail to capture sparse minority-class evidence, even when such evidence exists in close proximity within the feature space. These findings highlight a limitation of globally fixed neighborhood sizes and underscore the importance of local neighborhood composition in imbalanced classification.

Beyond these algorithmic formulations, recent theoretical work has emphasized the limitations of accuracy as a performance measure in imbalanced classification. Generalizations of the Bayes-optimal classifier to performance measures derived from the confusion matrix offer a principled foundation for evaluating classifiers under alternative criteria such as recall, precision, and F -scores. This theoretical perspective also highlights that optimal classification decisions may require stochastic decision rules and depend on the nature and severity of class imbalance.

Building on these theoretical insights, kNN-specific methods have been proposed that address class imbalance directly within the probabilistic decision process. By modifying prior class probabilities, these approaches reduce the influence of global class distributions and increase sensitivity to minority classes without altering the training data or introducing additional model complexity. Such methods demonstrate how imbalance-aware decision rules can be integrated into kNN while

maintaining its computational simplicity.

Collectively, these studies provide a structured foundation for investigating kNN-based classification under class imbalance. They highlight that effective imbalance handling requires more than global accuracy optimization, instead calling for decision mechanisms that explicitly account for asymmetric misclassification costs, the structure of local neighborhoods, and evaluation objectives that emphasize minority-class performance. At the same time, the reviewed approaches demonstrate that such adaptations can be achieved while preserving the inherent interpretability, locality, and computational efficiency that characterize nearest neighbor methods.

2 Methodology

2.1 kNN

The k-Nearest Neighbors algorithm (kNN) was first introduced by Cover and Hart in 1967: "The nearest neighbor decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points" [5]. They theoretically justified its performance by demonstrating that the nearest neighbor rule error rate is bounded by twice the Bayes error rate, the lowest possible classification error achievable under ideal conditions.

kNN is a non-parametric and instance-based learning algorithm, meaning it makes no assumptions about the underlying data distribution and stores all training instances for use during prediction. It is also considered a lazy learning method because model building is deferred until a query is made, making prediction computationally expensive, especially on large datasets. Despite its simplicity, kNN is often effective for many classification tasks due to its intuitive approach and simple implementation.

To measure distance between points to find nearest neighbor, Euclidean distance was used:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad x_i, y_i \in \mathbb{R}^n \quad (7)$$

Here, n represents the number of features, and x_i and y_i denote the values of the i -th feature of two points \mathbf{x} and \mathbf{y} in \mathbb{R}^n space.

After identifying the k nearest neighbors of a query point \mathbf{x} using the chosen distance metric, the k-Nearest Neighbors algorithm assigns a class label based on majority voting. Let k_c denote the number of neighbors belonging to class c . The predicted class \hat{c} is then given by

$$\hat{c} = \arg \max_c k_c. \quad (8)$$

This voting scheme can be interpreted as a maximum a posteriori (MAP) decision rule, where the posterior probability $P(c | \mathbf{x})$ is estimated locally by the fraction of neighbors belonging to class c , and class prior probabilities are implicitly estimated from the empirical class frequencies in the training data [3].

In this work, the kNN algorithm was implemented using the `KNeighborsClassifier` class from the `scikit-learn` Python library.

2.2 SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is widely regarded as a foundational data-level method for addressing class imbalance and is commonly used as a benchmark for eval-

uating new imbalance-correction frameworks [4]. Unlike naive over-sampling approaches that replicate existing minority instances, thereby increasing the risk of overfitting, SMOTE generates synthetic minority samples by interpolating between existing minority observations in the feature space.

For a given minority instance $x_i \in \mathbb{R}^n$, the algorithm first identifies its k -nearest neighbors belonging to the minority class. A synthetic sample is then created by randomly selecting one such neighbor x_{z_i} and computing a new point along the line segment connecting x_i and x_{z_i} according to:

$$x_{\text{new}} = x_i + \delta \cdot (x_{z_i} - x_i), \quad (9)$$

where $\delta \in [0, 1]$ is a random scalar drawn from a uniform distribution.

This procedure effectively expands the minority-class decision region by populating the feature space between existing minority samples, rather than duplicating them. While this strategy often improves classifier performance in moderately imbalanced settings, SMOTE may introduce synthetic samples in low-density or overlapping regions of the feature space, particularly when the minority class exhibits complex or multimodal distributions. As a result, its effectiveness can degrade in scenarios where local neighborhood structure is critical for reliable classification [11].

In this work, SMOTE was implemented using the `SMOTE` class from the `over_sampling` module of the `imblearn` Python library.

2.3 STIR-kNN

The Stochastic Inference-Time Rebalancing kNN (STIR-kNN) operates as a probabilistic wrapper. The framework implements a gated logic, divided into the following phases:

Phase 1: Intervention Gating Let n_{maj} and n_{min} denote the numbers of majority and minority class samples in the training set, respectively, and let y_{maj} and y_{min} represent the corresponding class labels. The global imbalance ratio is defined as:

$$IR = \frac{n_{\text{maj}}}{n_{\text{min}}}. \quad (10)$$

For a test instance x whose baseline k -nearest neighbors (kNN) classifier predicts the majority class, an inference-time intervention is activated stochastically with probability:

$$p_{\text{int}} = 1 - \frac{1}{IR}. \quad (11)$$

The intervention is explicitly conditioned on the kNN prediction \hat{y} , such that

$$P(\text{intervention}) = \begin{cases} p_{\text{int}}, & \text{if } \hat{y} = y_{\text{maj}}, \\ 0, & \text{if } \hat{y} = y_{\text{min}}. \end{cases} \quad (12)$$

This asymmetric intervention strategy reflects the assumed cost structure of the problem, in which false negatives are substantially more costly than false positives, and therefore limits corrective adjustments to majority-class predictions.

The proposed gating mechanism ensures that the frequency of intervention adapts smoothly to the severity of class imbalance. In balanced settings ($IR \rightarrow 1$), $p_{\text{int}} \rightarrow 0$, recovering the standard kNN decision rule. As the imbalance ratio increases, the intervention probability approaches unity, progressively increasing the likelihood of local rebalancing in regions dominated by the majority class.

Phase 2: Elastic Evidence Discovery To ensure that corrections are grounded in local data, we search for minority-class evidence within an expanded neighborhood. We introduce an adaptive search radius, k_{elastic} , which scales sub-linearly with the class imbalance to overcome the locality limitations of a fixed k :

$$k_{\text{elastic}} = \max\left(k + 1, \lfloor k \cdot \sqrt{IR} \rfloor\right). \quad (13)$$

The square-root scaling facilitates a sub-linear expansion of the search volume, effectively compensating for the decreased spatial density of minority class samples as IR increases. This strategy preserves the local manifold structure, ensuring the neighborhood is wide enough to recover sparse minority signals but narrow enough to avoid diluting feature relevance, which can occur with overly large neighborhoods.

The intervention proceeds to the injection phase only if minority-class evidence is detected. Formally, let $\mathcal{L}(k_{\text{elastic}})$ denote the set of labels within the expanded neighborhood. The process continues only if

$$y_{\min} \in \mathcal{L}(k_{\text{elastic}}). \quad (14)$$

Otherwise, the original prediction $\hat{y} = y_{\text{maj}}$ is retained, preventing corrections in regions of the feature space that lack empirical minority support.

Phase 3: Stochastic Label Injection and Voting Once minority-class evidence is confirmed within the neighborhood, the STIR-kNN framework performs a local rebalancing that adjusts the label distribution to reflect the theoretical importance of the minority class relative to the global imbalance ratio. This begins with the calculation of a continuous target influence, T , defined as a logarithmic function of the imbalance ratio:

$$T = \log_2(IR), \quad (15)$$

which ensures that the required minority representation grows at a diminishing rate. A doubling of the global imbalance increases the target minority count by only one. This sub-linear growth prevents overly aggressive rebalancing and preserves majority-class precision even under extreme imbalance.

Since the target influence T is generally non-integer, the desired number of minority-class labels to be introduced into the neighborhood must be mapped to a discrete quantity. This is achieved via stochastic quantization, which preserves the expected value of T while avoiding hard thresholds.

Let $f = T - \lfloor T \rfloor$ denote its fractional part. The target number of minority labels, N_{target} , is then sampled as:

$$N_{\text{target}} = \begin{cases} \lfloor T \rfloor + 1 & \text{with probability } f, \\ \lfloor T \rfloor & \text{with probability } 1 - f. \end{cases} \quad (16)$$

This probabilistic discretization ensures a smooth transition in model behavior as the imbalance ratio varies, preventing abrupt changes caused by fixed thresholds.

Local rebalancing is performed only if the current number of minority labels in the neighborhood is less than N_{target} . In such cases, a subset of majority-class labels is randomly selected and reassigned to the minority class until the target is reached. By modifying existing labels rather

than introducing new neighbors, the neighborhood size k remains fixed, thereby preserving local density and maintaining the geometric structure of the feature space.

Once minority-class evidence has been confirmed and went through a process of the local neighborhood adjustment through stochastic label injection, the final corrected prediction for an instance x , denoted \hat{y}_{corr} , is obtained by randomly selecting a label from the adjusted neighborhood $\mathcal{L}_{\text{adjusted}}$:

$$\hat{y}_{\text{corr}} \sim \text{Discrete}(\mathcal{L}_{\text{adjusted}}). \quad (17)$$

In practice, this corresponds to randomly picking a label from the list $\mathcal{L}_{\text{adjusted}}$, so that labels appearing more frequently have a proportionally higher probability of being chosen. This stochastic selection mechanism preserves the empirical distribution of labels within the neighborhood while accounting for the rebalancing effect of injected minority labels. Consequently, minority-class instances are assigned a statistically meaningful chance of selection, directly enhancing recall in imbalanced regions of the feature space.

To formalize the inference-time correction procedure, we present Algorithm 1, which summarizes the stochastic neighborhood rebalancing steps for kNN predictions. The algorithm highlights the key components of our method, including probabilistic intervention, adaptive neighborhood expansion, and stochastic vote adjustment, while abstracting low-level implementation details for readability.

For full reproducibility, the complete Python implementation of the method is provided in Appendix A.1. This includes all index-level operations, random sampling, and neighborhood manipulations, allowing other researchers to reproduce our experiments exactly.

Algorithm 1. Stochastic Inference-Time Rebalancing kNN (Simplified)

Require: Trained kNN, test set X_{test} , training labels y_{train} , minority label y_{min} , k

Ensure: Corrected predictions \hat{y}_{corr}

- 1: Compute imbalance ratio IR and intervention probability p_{int}
 - 2: Compute adaptive neighborhood k_{elastic}
 - 3: **for** each X_{test} instance x_i **do**
 - 4: **if** $\hat{y}_{\text{kNN}}[i]$ is majority **and** $\text{Random}() < p_{\text{int}}$ **then**
 - 5: $N \leftarrow$ labels of k nearest neighbors
 - 6: **if** minority exists in N or expanded neighborhood of size k_{elastic} **then**
 - 7: Compute target minority influence $T = \log_2(IR)$
 - 8: Adjust N by flipping majority labels stochastically to meet T
 - 9: $\hat{y}_{\text{corr}}[i] \leftarrow$ Random choice from adjusted N
 - 10: **end if**
 - 11: **end if**
 - 12: **end for**
 - 13: **return** \hat{y}_{corr}
-

3 Experimental Methodology

3.1 Datasets

To evaluate the robustness of the STIR-kNN framework across diverse data structures, three benchmark datasets were sourced from the UC Irvine Machine Learning Repository [12]. These datasets were selected to represent varying feature types, including purely numerical, heterogeneous, and time-dependent historical data. Allowing for a comprehensive assessment of the model’s performance in real-world scenarios. All three datasets represent binary classification tasks.

3.1.1 Numerical Dataset

The Electrical Grid Stability dataset provides a fully numerical benchmark for evaluating the proposed method in a controlled setting. It consists of simulated measurements from a four-node star power system, with twelve continuous input features describing power balance and the response times of energy producers and consumers. The task is to predict whether the grid operates in a stable or unstable regime.

Because all features are continuous and generated through mathematical simulation, this dataset enables an isolated assessment of the elastic neighborhood expansion (k_{elastic}) in a purely Euclidean feature space. Decision boundaries in this setting arise from physical system dynamics rather than behavioral or demographic correlations, allowing the impact of the proposed inference-time rebalancing mechanism to be examined independently of feature heterogeneity.

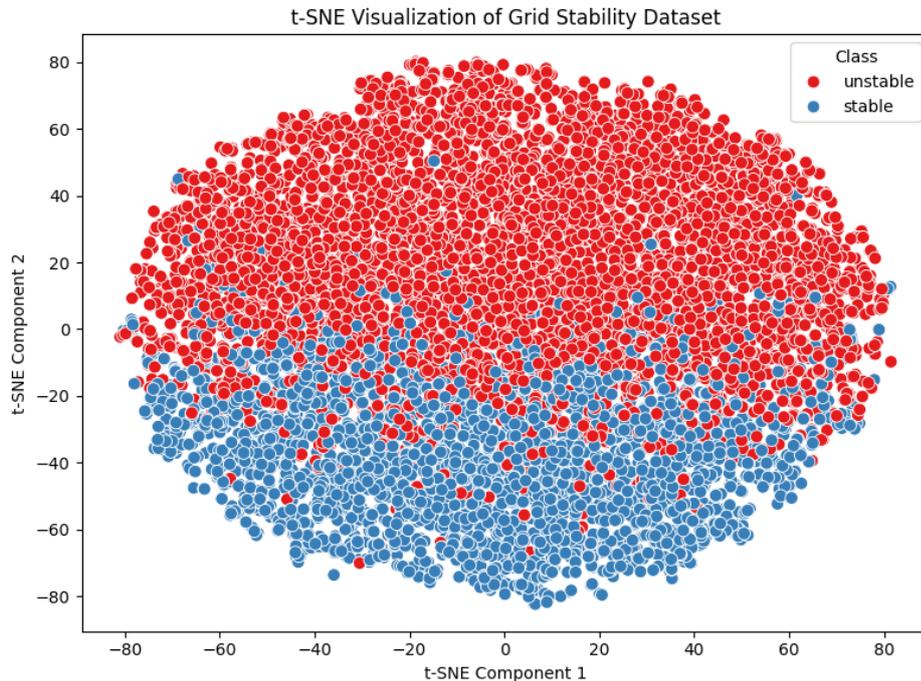


Figure 1. Two-dimensional t-SNE projection of the Electrical Grid Stability dataset

To visualize the dataset in a lower-dimensional space, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied. As shown in Figure 1, the two classes exhibit a relatively clear separation, indicating that the stability regimes are distinguishable in the feature space. Nevertheless, a small number of samples appear scattered between the main clusters, which may be interpreted

as borderline cases or numerical artifacts of the simulation. These points represent regions where local neighborhood composition may strongly influence classification outcomes.

3.1.2 Heterogeneous Dataset

The Census Income dataset, derived from the 1994 United States Census Bureau database and commonly known as the *Adult* dataset, is used as a primary benchmark for evaluating the performance of the proposed STIR-kNN framework on heterogeneous data. The dataset comprises fourteen attributes capturing a range of sociodemographic characteristics, including continuous numerical features such as age, education-num, and capital-gain, as well as categorical variables such as occupation, workclass, and marital status. The classification task is to predict whether an individual’s annual income exceeds \$50,000, thereby partitioning the population into two economic classes: high-income ($> \$50,000$) and low-income ($\leq \$50,000$).

From a machine learning perspective, this dataset is challenging due to the complex relationships between categorical attributes and income outcomes, which can induce irregular and noisy decision boundaries. In the original data distribution, the high-income class constitutes approximately 24% of the total instances, resulting in a moderately imbalanced classification problem. This imbalance can adversely affect distance-based classifiers such as kNN, as neighborhoods are often dominated by majority-class (low-income) instances, increasing the likelihood of minority-class underprediction.

By employing this dataset, we assess the ability of the STIR-kNN framework to adaptively rebalance local neighborhoods in mixed-attribute feature spaces, improving the detection of high-income individuals without compromising the integrity of the original data distribution.

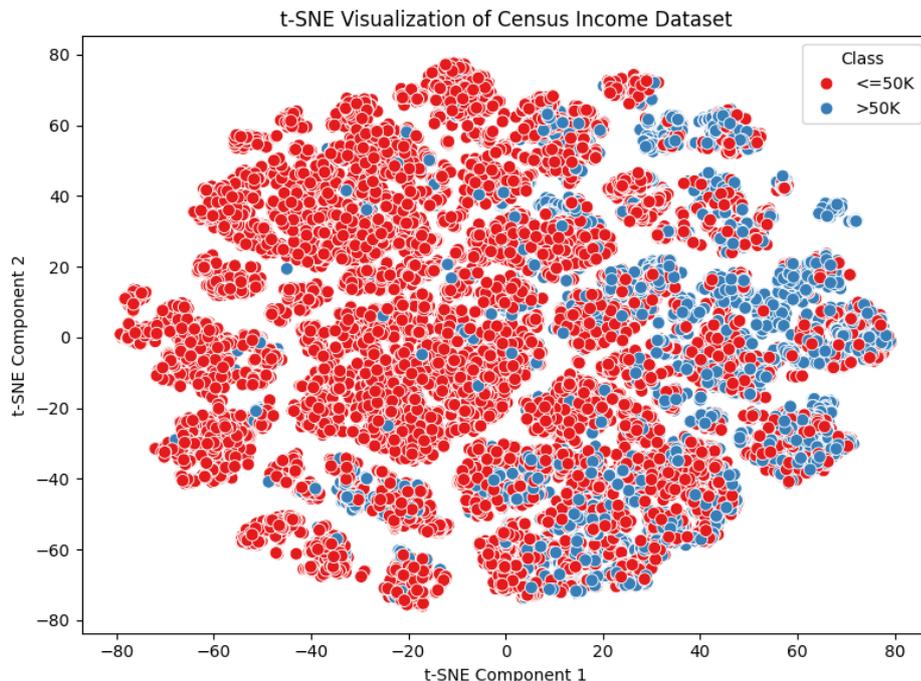


Figure 2. Two-dimensional t-SNE projection of the Census Income dataset

To gain qualitative insight into the structure of the dataset, a two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) projection is shown in Figure 2. While some degree of class separation is observable—particularly in regions dominated by low-income instances—there

remains substantial overlap between the two classes. This overlap reflects the inherent ambiguity of income prediction given the limited set of observed sociodemographic attributes, which do not fully capture the factors governing individual earnings. As a result, samples from different income classes may occupy similar regions of the feature space, leading to locally heterogeneous neighborhoods.

3.1.3 Time-dependent Historical Dataset

The Bank Marketing dataset captures the outcomes of direct phone-call marketing campaigns conducted by a Portuguese banking institution and exhibits a historical, time-dependent structure. Each record corresponds to a marketing contact with a client, and the target variable indicates whether the client subscribed to a bank term deposit following the campaign. The resulting binary classification task distinguishes between successful subscription (yes) and non-subscription (no) outcomes. In some cases, multiple contacts with the same client were required before a final subscription decision was reached.

The `duration` attribute was removed, as this variable is only observed after a call has concluded and would otherwise introduce data leakage, leading to artificially inflated performance. In addition, the attributes `campaign`, `previous`, `month`, and `day` were excluded, as they were determined to be either redundant or weakly informative for the task of predicting future client subscription behavior.

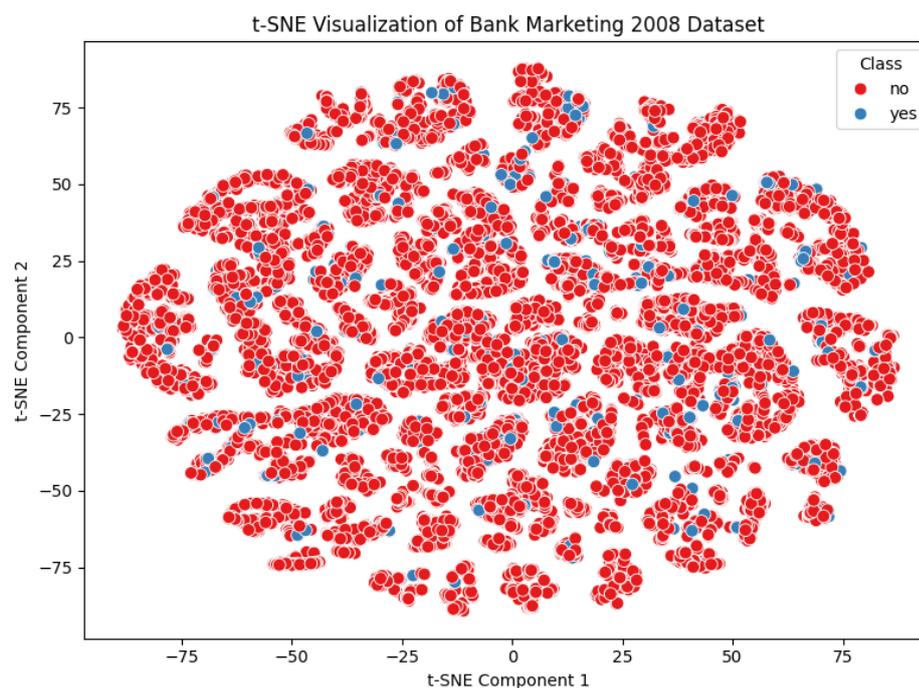


Figure 3. Two-dimensional t-SNE projection of the Bank Marketing dataset of 2008

To assess the framework’s ability to generalize across time, a chronological data split was employed rather than random shuffling. The training set comprises records collected in 2008, while the test set includes data from May 2009 onward. This temporal validation strategy reflects a realistic deployment scenario in which models are trained on historical campaign data to predict the outcomes of future marketing efforts. Because successful subscription events constitute a small fraction of the dataset, this benchmark presents a challenging imbalanced classification problem,

providing a rigorous testbed for evaluating the effectiveness of the STIR-kNN framework under strictly prospective conditions.

Because the train–test split differs from the previously described datasets, it is important to examine the data distributions in the training and test sets separately. Figure 3 illustrates marketing data collected in 2008.

The visualization highlights the strong class imbalance, with the majority class (no) dominating most regions of the feature space. The two classes exhibit substantial overlap, indicating that successful and unsuccessful campaign outcomes are often associated with similar feature representations.

Figure 4 shows data collected from May 2009 onward. Compared to the 2008 data, some localized clustering of the two classes can be observed, suggesting a modest increase in class separability. Nevertheless, considerable overlap between classes remains, and the decision boundary is still highly irregular. While the class imbalance in this period is less severe than in the 2008 data, the persistence of mixed local neighborhoods continues to complicate neighborhood-based classification.

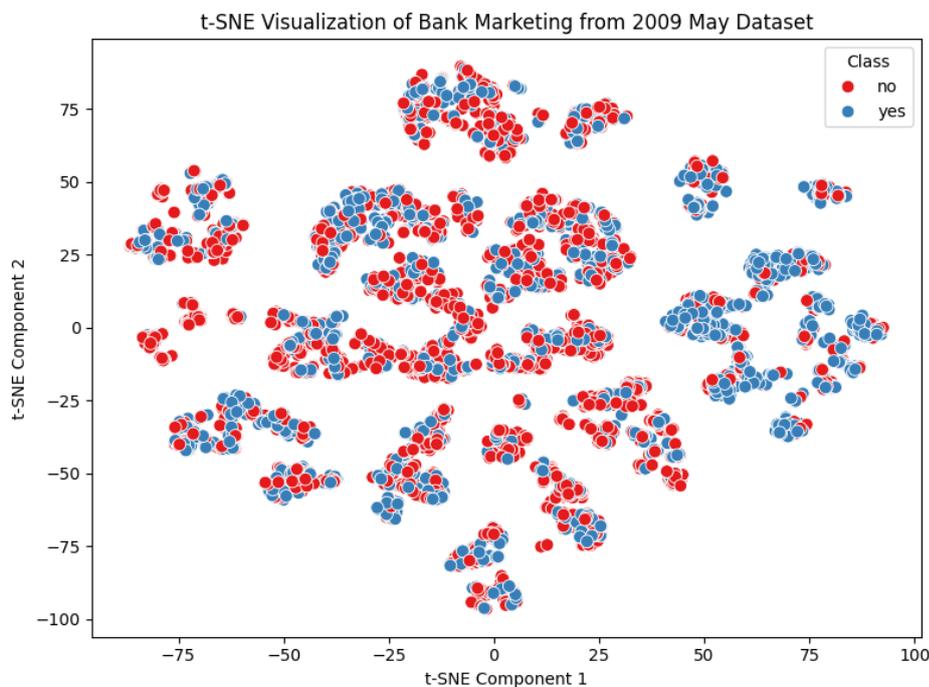


Figure 4. Two-dimensional t-SNE projection of the Bank Marketing dataset starting from May 2009

3.1.4 Summary

The key structural characteristics of the three benchmark datasets used in the experimental evaluation are summarized in Table 1.

Table 1. Summary of datasets used in the experimental evaluation.

Dataset	Type	Samples	Features	Minority (%)	Split Strategy
Grid Stability	Numerical	10,000	12	~36%	Random (train/test)
Census Income	Heterogeneous	32,561	14	~24%	Random (train/test)
Bank Marketing	Historical	45,208	12	~12%	Temporal (2008 / 2009+)

While the datasets differ in scale and feature composition, they also vary substantially in class imbalance and train–test splitting strategy, resulting in distinct neighborhood structures and learning challenges.

Together, these datasets provide complementary test conditions for analyzing the behavior of distance-based classifiers under varying levels of feature complexity and distributional shift. In particular, they enable a systematic examination of how local neighborhood composition changes across numerical, mixed-attribute, and temporally evolving data, and how effectively the proposed STIR-kNN framework adapts to these conditions at inference time.

3.2 Data Preprocessing

To ensure a consistent and reliable input for the STIR-kNN framework, a unified preprocessing pipeline was applied to all datasets. This pipeline addresses the requirements of distance-based algorithms while mitigating the risks of high-dimensionality and data leakage.

For the Census Income and Bank Marketing datasets, categorical variables were transformed using target encoding. Unlike one-hot encoding, which can lead to a significant expansion of the feature space and trigger the curse of dimensionality [13], target encoding maps categories to a single numerical value based on the target mean. This preserves the predictive signal of the categorical data while maintaining a compact dimensionality, which has been shown to perform effectively in benchmarking studies [15].

Following the encoding of categorical variables and the removal of non-informative attributes, all features across the three benchmark datasets were standardized. Since the k -nearest neighbors classifier relies on distance-based similarity, it is highly sensitive to disparities in feature scale, where features with larger numerical ranges can disproportionately influence neighborhood formation [10].

To mitigate this effect, each feature was standardized to have zero mean and unit variance ($\mu = 0, \sigma = 1$) using z-score normalization:

$$z = \frac{x - \mu}{\sigma}. \tag{18}$$

This preprocessing step ensures that all features contribute equally to the distance metric, preventing arbitrary units of measurement from biasing neighborhood selection. As a result, both the baseline kNN decision rule and the proposed stochastic inference-time rebalancing operate on a geometrically consistent feature space.

In this work, target encoding was implemented using the `MEstimateEncoder` class from the `category_encoders` Python library, while feature scaling was performed using the `StandardScaler` class from the `preprocessing` module of the `scikit-learn` Python library.

3.3 Evaluation Metrics

The performance of all evaluated models is assessed using accuracy, precision, recall, and the F1-score. While accuracy remains a commonly reported metric, it is insufficient in isolation for imbalanced classification problems, as it can obscure poor minority-class performance when the majority class dominates the dataset. Consequently, additional class-sensitive metrics are required to provide a more informative evaluation.

The primary objective of the proposed STIR-kNN framework is to improve the recall of the minority class, thereby reducing false negatives in scenarios where minority instances are systematically underrepresented. Minority-class recall directly measures the model’s ability to identify rare but critical samples and therefore serves as the principal indicator of the framework’s effectiveness.

However, optimizing recall alone may lead to excessive misclassification of majority-class instances. To ensure that performance gains for the minority class do not come at an unreasonable cost, precision is also reported, capturing the proportion of predicted minority instances that are correct. This allows for monitoring potential overcorrection effects introduced by the stochastic rebalancing mechanism.

Although accuracy is not the primary optimization objective, it provides context for the overall impact of inference-time rebalancing, to ensure that gains in minority-class performance do not come at the expense of a disproportionate degradation in majority-class predictions.

Finally, the F1-score is included as a harmonic mean of precision and recall, providing a balanced summary of classification performance in the presence of class imbalance. This metric is particularly valuable for comparing models that exhibit different trade-offs between recall and precision. Therefore, the F1 score reported in the performance summary in Section 4.3 as the median value across all ratios to provide a stable aggregate comparison.

Together, these metrics provide a comprehensive and principled evaluation framework that reflects both the targeted objectives of STIR-kNN and the need to maintain reasonable overall classification performance.

All evaluation metrics were computed using the `accuracy_score`, `precision_score`, `recall_score`, and `f1_score` functions from the `metrics` module of the `scikit-learn` Python library.

3.4 Experimental Setup

For each experiment, the dataset was partitioned into training and testing subsets using an 80/20 split. To systematically assess model performance under varying levels of class imbalance, the training dataset was resampled to achieve imbalance ratios ranging from 2 to 30, including the original class distribution. At each ratio, the models’ predictive performance was evaluated on the unchanged test set, ensuring a fair comparison of minority-class detection across experiments.

The number of neighbors k for kNN is a key hyperparameter that influences the classifier’s bias–variance trade-off. For every training set, the k was selected by evaluating candidate values in the range:

$$k \in [3, k_{\max}], \quad k_{\max} = \lfloor 2.5 \cdot \log(N) \rfloor, \quad (19)$$

where N is the number of training samples. Although our primary objective is to improve minority-class recall, the number of neighbors k is determined via cross-validation on the training set by maximizing accuracy. This is because k governs the geometric structure and stability of local neighborhoods in kNN, while minority-class performance is explicitly addressed by the proposed inference-time rebalancing mechanism.

The choice of maximum k balances the need for a sufficiently large neighborhood to capture global patterns while avoiding overly diffuse neighborhoods that may include distant, less relevant samples. Previous works on kNN have proposed scaling k with the number of training instances using sub-linear functions such as \sqrt{N} [5], but in our experiments, \sqrt{N} often produced excessively

large neighborhoods, leading to suboptimal accuracy. The logarithmic scaling factor ensures that k grows slowly with N , remaining large enough to stabilize predictions without compromising local sensitivity.

The evaluation considered three models: the proposed STIR-kNN, the classic kNN, and kNN trained on SMOTE-augmented data (hereafter referred to as SMOTE-kNN). These models were selected to represent distinct strategies for addressing class imbalance within the kNN framework. The classic kNN serves as a baseline that reflects the behavior of the standard algorithm in the absence of any imbalance correction. In contrast, SMOTE-kNN represents a widely adopted data-level approach, in which imbalance is addressed at training time through synthetic minority over-sampling. STIR-kNN, on the other hand, embodies an inference-time correction strategy that preserves the original training data and modifies the decision process dynamically at test time.

By keeping the test set fixed across all experiments, differences in performance can be attributed solely to the rebalancing strategy employed rather than variations in evaluation data. This experimental setup enables a focused analysis of how each approach responds to varying degrees of class imbalance, particularly in terms of minority-class recall, while simultaneously monitoring the impact on majority-class predictions.

4 Results and Analysis

4.1 Experimental Results

For each dataset, model performance is evaluated following the procedure described in Section 3.4. Results are presented using a consistent set of plots to enable direct comparison across datasets.

Each figure reports accuracy, precision, and recall results for STIR-kNN in comparison to baseline models - classical kNN and SMOTE-kNN. For each metric, the relative performance gain of STIR-kNN over the corresponding baseline is shown across varying imbalance ratios. In addition, for precision and recall, separate plots present the absolute performance values of the minority class for STIR-kNN and each baseline model. This combined visualization facilitates a detailed assessment of minority-class behavior under increasing imbalance. The vertical dashed line in each plot denotes the original imbalance ratio of the training-set prior to resampling.

Gains are visualized on a \log_2 scale, where positive values indicate improvement over the baseline and negative values indicate degradation. The \log_2 scale is employed to ensure numerical stability and comparability across imbalance regimes. In highly imbalanced settings, baseline models, particularly classical kNN, may yield near-zero or zero values for certain metrics. Using a logarithmic gain formulation prevents such cases from disproportionately distorting the visualization while preserving the direction and relative magnitude of performance changes.

To compute relative performance gains, we report the log-fold change between the proposed method and a baseline. For a given metric value M_{STIR} produced by STIR-kNN and the corresponding baseline value M_{base} , the gain is defined as:

$$\Delta_{\log_2} = \log_2 \left(\frac{M_{\text{STIR}} + \varepsilon}{M_{\text{base}} + \varepsilon} \right), \quad (20)$$

where ε is a small positive constant introduced for numerical stability. In all experiments, ε was fixed to 10^{-6} .

4.1.1 Electrical Grid Stability

For the Electrical Grid Stability dataset, we first examine minority-class recall, the primary optimization objective of the proposed STIR-kNN framework. As shown in Figure 5, classical kNN exhibits a quite sharp decline of minority-class (stable) in recall as the training-set imbalance ratio increases. This behavior is characteristic of distance-based classifiers, and especially kNN, where local neighborhoods become increasingly dominated by majority-class instances, causing minority samples to be systematically underrepresented in the voting process.

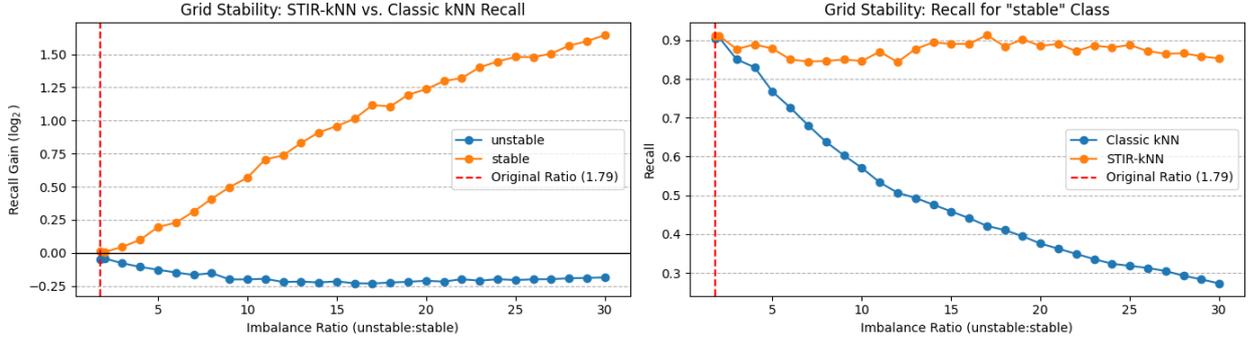


Figure 5. Recall comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Electrical Grid Stability dataset

In contrast, STIR-kNN maintains consistently high minority-class recall across the full range of imbalance ratios considered. As shown in Figure 5, the recall advantage over classical kNN increases monotonically as class imbalance becomes more severe. This behavior indicates that the inference-time rebalancing mechanism becomes progressively more influential as the dominance of the majority class intensifies. Specifically, it reflects the combined effect of stochastic intervention gating and elastic neighborhood expansion, which together allow the model to recover evidence of minority-class that would otherwise be excluded under a fixed neighborhood k .

The improvement in minority-class recall is accompanied by a reduction in majority-class (unstable) recall relative to classical kNN. However, the decline remains uniform across imbalance ratios and does not exhibit sudden degradation. This behavior represents an expected and controlled trade-off in settings with asymmetric misclassification costs, where prioritizing sensitivity to minority instances is operationally preferable to preserving majority-class recall.

At mild to moderate imbalance levels, SMOTE-kNN exhibits slightly higher minority-class recall than STIR-kNN, as shown in Figure 6. This behavior is consistent with SMOTE’s training-time oversampling strategy, which increases minority-class density in the feature space and therefore strengthens minority presence in local neighborhoods. Nevertheless, under more severe imbalance conditions, STIR-kNN attains higher minority-class recall, suggesting that inference-time rebalancing may outperform SMOTE-kNN when minority support becomes increasingly sparse.

By comparison with SMOTE-kNN, STIR-kNN preserves the original training distribution and relies exclusively on inference-time adjustment, demonstrating that strong minority sensitivity can be achieved without retraining or synthetic data generation. A corresponding difference is observed in majority-class recall: SMOTE-kNN retains marginally higher majority-class recall than STIR-kNN, but the gap remains small and stable across the evaluated imbalance range. However, compared to the classical kNN baseline, this trade-off is less pronounced, indicating that both STIR-kNN and SMOTE-kNN redistribute recall more evenly between classes than standard kNN.

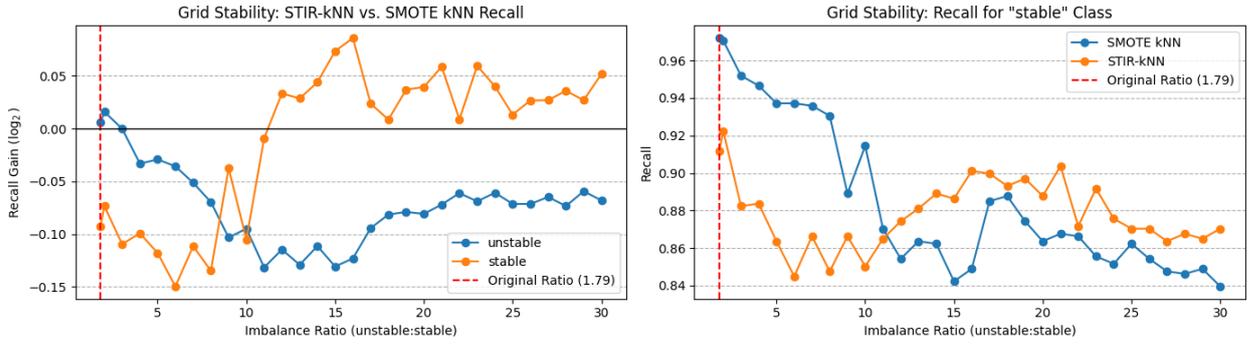


Figure 6. Recall comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Electrical Grid Stability dataset

The improvement in minority-class recall is accompanied by a controlled reduction in minority-class precision, as illustrated in Figure 7. This behavior reflects the intentional bias introduced by the stochastic label-injection mechanism, which prioritizes the reduction of false negatives at the expense of an increased false-positive rate. Importantly, the decline in minority-class precision remains gradual and stabilizes as imbalance increases, indicating that rebalancing is applied in a constrained and evidence-driven manner.

At the same time, majority-class precision exhibits a consistent improvement relative to classical kNN. This increase reflects the redistribution of decision boundaries induced by inference-time rebalancing, which reduces incorrect minority predictions in regions strongly dominated by the majority class. As a result, STIR-kNN achieves a more balanced precision profile across classes, reinforcing that recall gains are not obtained through indiscriminate label reassignment.

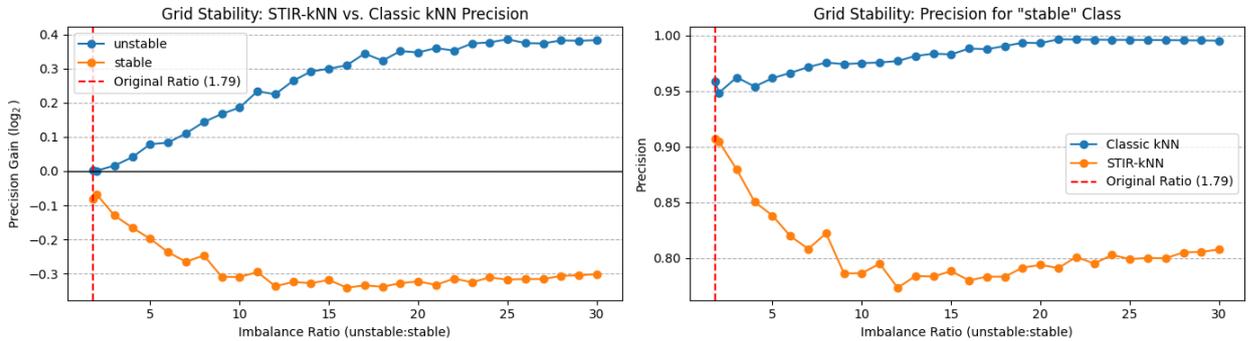


Figure 7. Precision comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Electrical Grid Stability dataset

Relative to SMOTE-kNN, STIR-kNN exhibits a smoother precision–recall trade-off, despite SMOTE-kNN achieving marginally higher minority-class recall in mild imbalance regimes. As shown in Figure 8, SMOTE-kNN achieves higher minority-class precision but experiences sharper declines as imbalance increases, a pattern consistent with the effects of aggressive training-time oversampling. In contrast, STIR-kNN demonstrates more gradual and stable precision behavior, reflecting the constrained nature of inference-time rebalancing.

Additionally, at more severe imbalance ratios, STIR-kNN attains slightly higher majority-class precision than SMOTE-kNN. This behavior indicates that preserving the original training distribution allows STIR-kNN to avoid excessive minority overprediction in regions strongly dominated by the majority class, thereby maintaining better precision for the majority class under extreme

imbalance.

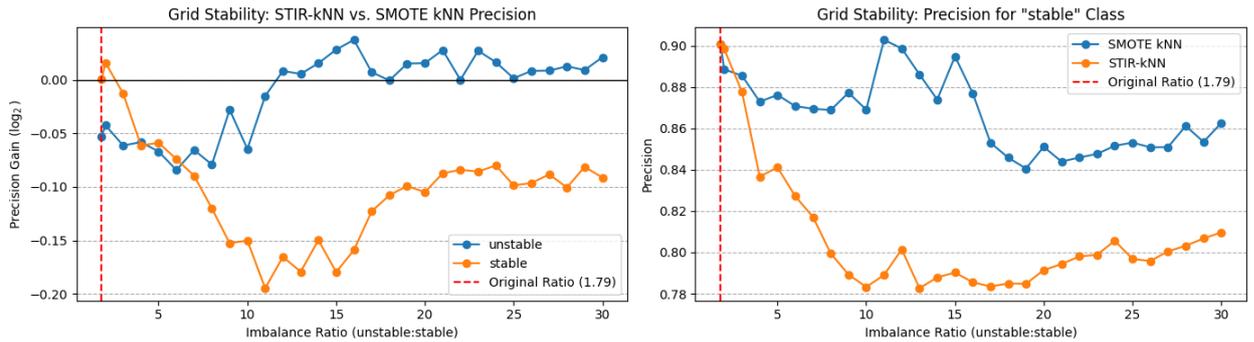


Figure 8. Precision comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Electrical Grid Stability dataset

Figure 9 reports accuracy trends for STIR-kNN relative to both classical kNN and SMOTE-kNN. Across the evaluated imbalance regimes, STIR-kNN maintains competitive accuracy and, relative to classical kNN, exhibits a consistent improvement as imbalance increases. This indicates that the gains in minority-class recall are not achieved through indiscriminate label flipping, but rather through informed local adjustments grounded in neighborhood evidence.

When compared to SMOTE-kNN, STIR-kNN yields slightly lower overall accuracy across most imbalance levels. However, the magnitude of this difference remains modest and stable, suggesting that the recall improvements obtained by STIR-kNN do not entail a disproportionate loss in global classification performance.

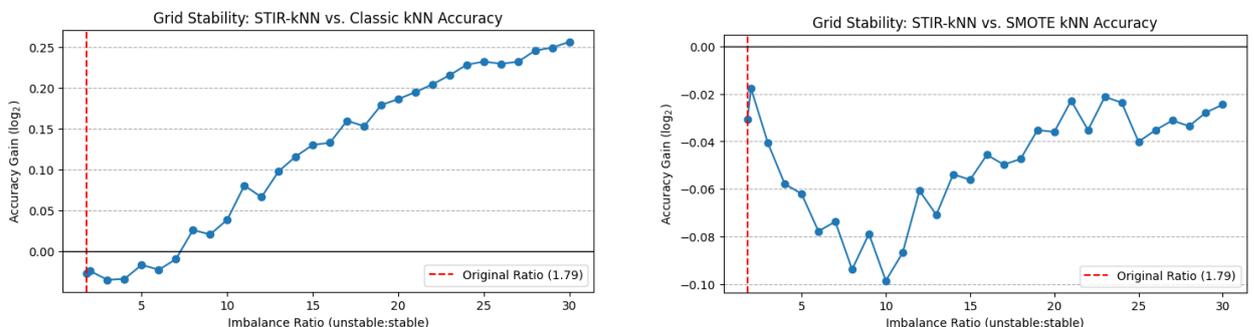


Figure 9. Overall accuracy on the Electrical Grid Stability dataset across increasing imbalance ratios. Left: comparison with classical kNN. Right: comparison with SMOTE-kNN

Overall, the Electrical Grid Stability results demonstrate that STIR-kNN effectively mitigates majority-class dominance in purely numerical settings governed by smooth physical dynamics, where class separation in the feature space is relatively clear. By adaptively rebalancing local neighborhoods at inference time, the framework achieves substantial improvements in minority-class recall while preserving controlled precision and competitive global accuracy. Importantly, these gains are obtained without modifying the underlying feature-space distribution, positioning STIR-kNN as a robust inference-time alternative to training-time oversampling methods such as SMOTE in continuous, simulation-driven domains.

4.1.2 Census Income

We next evaluate STIR-kNN on the Census Income dataset, focusing on recall for the minority high-income class ($> 50K$). As shown in Figure 10, classical kNN exhibits a pronounced and monotonic decline in recall as training-set imbalance increases. This degradation is particularly evident in the Census Income dataset due to its heterogeneous mix of categorical and numerical attributes, which leads to local neighborhoods dominated by high-frequency categorical values strongly associated with the majority ($\leq 50K$) class. As a result, minority-class instances are increasingly marginalized within local decision regions as imbalance grows.

STIR-kNN substantially mitigates this effect, maintaining consistently elevated recall across the full range of imbalance ratios. In contrast to the Electrical Grid Stability dataset, where class boundaries are shaped by smooth physical dynamics, the Census Income task is characterized by noisier and less separable decision regions. Under these conditions, inference-time rebalancing is particularly effective by amplifying weak minority-class signals that would otherwise be overwhelmed by majority-dominated neighborhoods formed around frequent categorical values.

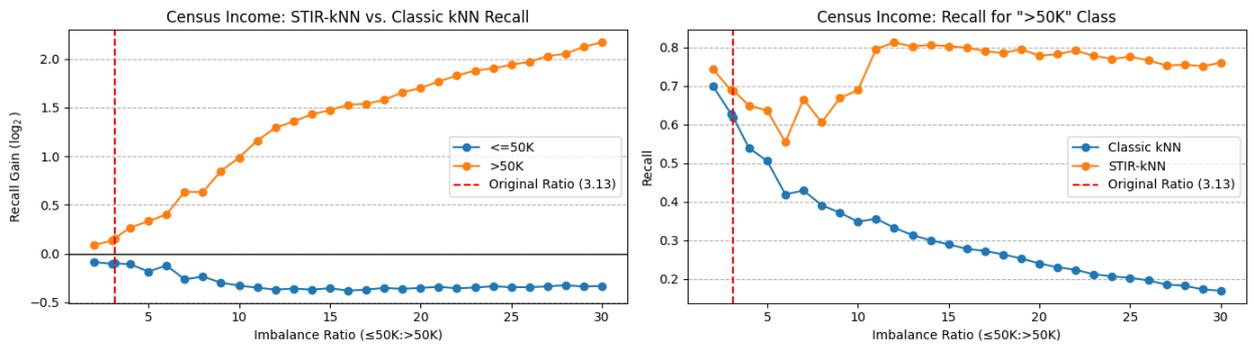


Figure 10. Recall comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Census Income dataset

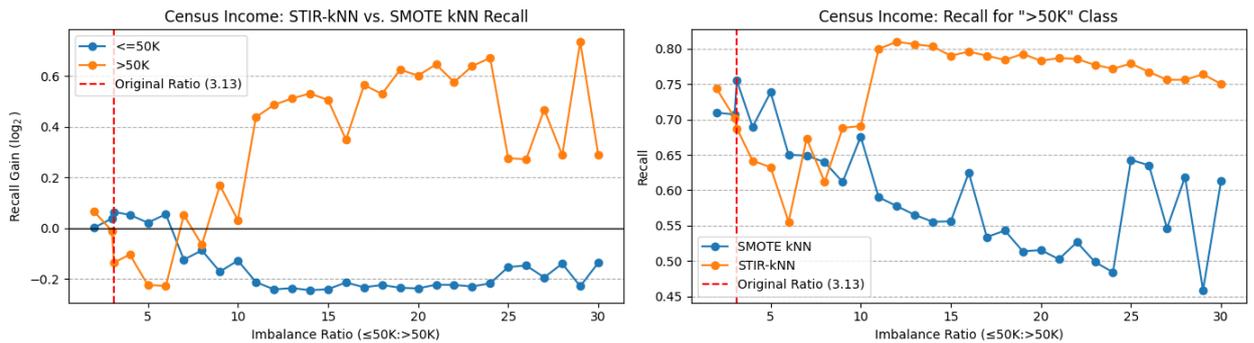


Figure 11. Recall comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Census Income dataset

Figure 11 further highlights the differences between STIR-kNN and SMOTE-kNN on the Census Income dataset. Across most imbalance ratios, STIR-kNN achieves higher minority-class recall, with relative gains emerging at lower imbalance levels than those observed for the Electrical Grid Stability dataset. Notably, under the most severe imbalance conditions, STIR-kNN attains substantially higher recall than SMOTE-kNN, indicating that inference-time rebalancing remains effective even when minority instances become extremely sparse. In these regimes, STIR-kNN

exhibits more stable and consistently elevated recall, whereas SMOTE-kNN shows increased variability.

While improvements in minority-class recall are accompanied by a reduction in majority-class recall relative to both classical kNN and SMOTE-kNN, this decline remains gradual and bounded, reflecting an intentional and controlled trade-off aligned with asymmetric misclassification costs. Consistent with the previously discussed dataset, the difference in majority-class recall between STIR-kNN and SMOTE-kNN is smaller than that observed relative to classical kNN.

The improvement in minority-class recall is accompanied by a measurable reduction in minority-class precision, as illustrated in Figure 12. Notably, this precision loss is more pronounced than that observed in the Electrical Grid Stability dataset, reflecting the increased sensitivity of STIR-kNN to minority-class instances in ambiguous regions of the feature space characterized by overlapping and heterogeneous attribute patterns. In contrast, majority-class precision relative to classical kNN is maintained or moderately improved, particularly at higher imbalance ratios. This consistency across datasets indicates that inference-time rebalancing redistributes classification uncertainty in a controlled manner rather than introducing indiscriminate degradation, yielding an asymmetric yet stable precision–recall trade-off.

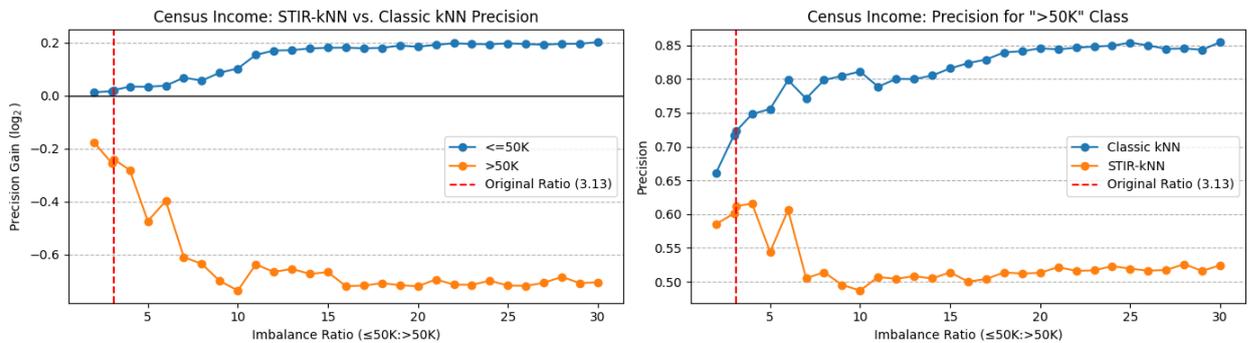


Figure 12. Precision comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Census Income dataset

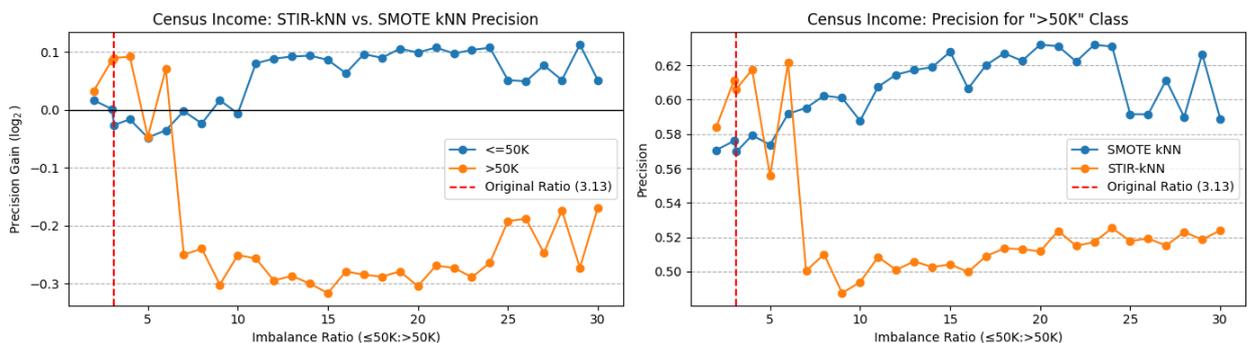


Figure 13. Precision comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Census Income dataset

Figure 13 illustrates minority-class precision trends for STIR-kNN and SMOTE-kNN. At mild imbalance levels, STIR-kNN occasionally achieves higher precision than SMOTE-kNN, indicating that inference-time rebalancing can preserve minority-class selectivity while improving recall. As imbalance severity increases, minority-class precision for STIR-kNN exhibits a sharper decline

before stabilizing, reflecting a controlled trade-off governed by local neighborhood evidence rather than global distributional modifications.

In contrast, SMOTE-kNN exhibits greater variability in precision across imbalance regimes. This behavior is consistent with the effects of training-time oversampling in heterogeneous, socio-demographic data, where synthetic minority instances may be generated in sparsely populated or weakly supported regions of the feature space. Such artifacts can distort local neighborhood structure and introduce unstable decision boundaries, particularly in the presence of high-cardinality categorical attributes.

These results indicate that STIR-kNN achieves a smoother and more predictable precision recall trade-off than SMOTE-kNN on the Census Income dataset. While minor variability is observed at low imbalance ratios, STIR-kNN exhibits notably more stable precision from moderately high to extreme imbalance levels, where inference-time rebalancing mitigates the volatility introduced by synthetic oversampling. By operating exclusively at inference time and preserving the empirical feature-space structure, STIR-kNN avoids the volatility associated with synthetic oversampling while maintaining competitive, and occasionally superior, precision performance.

With respect to majority-class precision, STIR-kNN exhibits a consistent improvement relative to classical kNN across imbalance regimes. When compared to SMOTE-kNN, majority-class precision under STIR-kNN is occasionally slightly lower at low imbalance ratios. However, as the imbalance increases, STIR-kNN progressively attains higher majority-class precision. This transition indicates that inference-time rebalancing increasingly mitigates overprediction of the minority class in majority-dominated regions, yielding improved majority-class selectivity at higher imbalance levels.

Accuracy trends for the Census Income dataset are reported in Figure 14 and are included to provide global performance context rather than as a primary optimization objective. As expected, given the strong emphasis on minority-class recall, STIR-kNN exhibits a modest reduction in overall accuracy relative to classical kNN across most imbalance regimes. However, as the imbalance ratio increases, the accuracy gap between classical kNN and STIR-kNN gradually narrows. In comparison with SMOTE-kNN, STIR-kNN achieves comparable overall accuracy, reflecting the fact that inference-time rebalancing remains largely inactive when class distributions are close to balanced. At lower imbalance ratios, STIR-kNN occasionally attains slightly higher accuracy than SMOTE-kNN. The decline in accuracy becomes more pronounced at moderate imbalance ratios, where inference-time rebalancing is already actively correcting majority-dominated neighborhoods and prioritizing minority sensitivity.

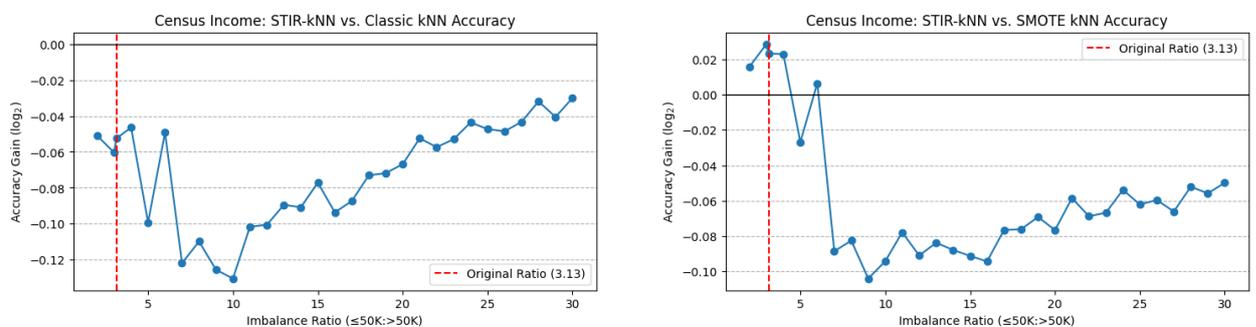


Figure 14. Overall accuracy on the Census Income dataset across increasing imbalance ratios. Left: comparison with classical kNN. Right: comparison with SMOTE-kNN.

Importantly, the observed accuracy degradation remains bounded and does not scale propor-

tionally with increasing imbalance. At higher imbalance levels, accuracy stabilizes and the gap relative to both classical kNN and SMOTE-kNN narrows, indicating that the recall gains achieved by STIR-kNN are not the result of indiscriminate label flipping but rather of targeted, evidence-driven local adjustments.

On the whole, the Census Income results demonstrate that STIR-kNN is well suited to heterogeneous, socio-demographic datasets characterized by weakly separable classes and high-cardinality categorical attributes. In such settings, classical distance-based voting is prone to majority dominance, leading to systematic under-detection of minority instances. The proposed inference-time rebalancing mechanism consistently counteracts this effect, yielding substantial and stable improvements in minority-class recall across imbalance regimes. Unlike training-time oversampling approaches such as SMOTE-kNN, which modify the original data distribution to increase minority density and may inadvertently amplify spurious correlations in socio-demographic data, STIR-kNN preserves the empirical feature-space structure and relies exclusively on inference-time adjustment. Importantly, these gains are achieved without excessive precision degradation or unstable accuracy behavior, positioning STIR-kNN as a robust and interpretable alternative to training-time oversampling methods for complex, socially derived data.

4.1.3 Bank Marketing

We finally evaluate STIR-kNN on the Bank Marketing dataset, which represents the most challenging scenario among the considered benchmarks due to its extreme class imbalance and time-dependent structure. The minority class (yes), corresponding to successful term deposit subscriptions, constitutes a sparse outcome with a baseline imbalance ratio of approximately 18.79. In this setting, false negatives are particularly costly, as missed positive responses directly correspond to lost business opportunities.

As shown in Figure 15, classical kNN exhibits a rapid collapse in minority-class recall as imbalance increases, approaching near-zero at higher ratios. This behavior reflects the overwhelming dominance of majority-class (no) instances in local neighborhoods, where sparse minority observations are systematically outvoted despite being temporally and behaviorally informative.

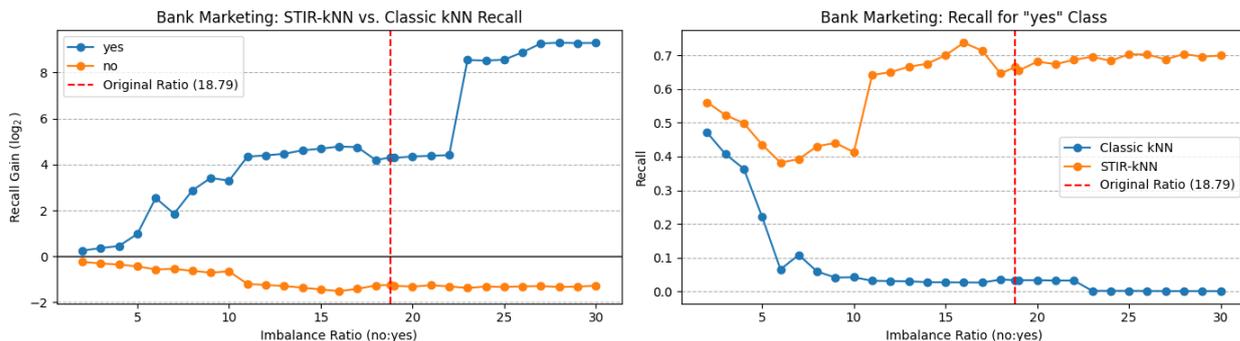


Figure 15. Recall comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Bank Marketing dataset

In contrast, STIR-kNN maintains substantially higher minority-class recall across the full range of imbalance ratios. The recall advantage over classical kNN increases sharply with imbalance severity, indicating that inference-time rebalancing becomes progressively more active as majority dominance intensifies. Unlike the previous datasets, the recall gains in Bank Marketing are partic-

ularly pronounced, highlighting the framework’s ability to recover rare but meaningful patterns in highly skewed, historically structured data.

These improvements are accompanied by a noticeable reduction in majority-class recall, particularly from moderate imbalance levels onward. Nevertheless, this decline remains stable and bounded across the imbalance spectrum and is outweighed by the substantial gains in minority-class sensitivity. In the context of marketing applications, where failing to identify potential subscribers is considerably more costly than contacting uninterested clients, this trade-off is both expected and operationally appropriate.

Figure 16 illustrates that STIR-kNN achieves competitive minority-class recall relative to SMOTE-kNN, particularly at moderate to severe imbalance ratios, where it consistently attains higher recall. By contrast, SMOTE-kNN, similarly to classical kNN, exhibits a sharp and sustained decline in minority-class recall as imbalance severity increases.

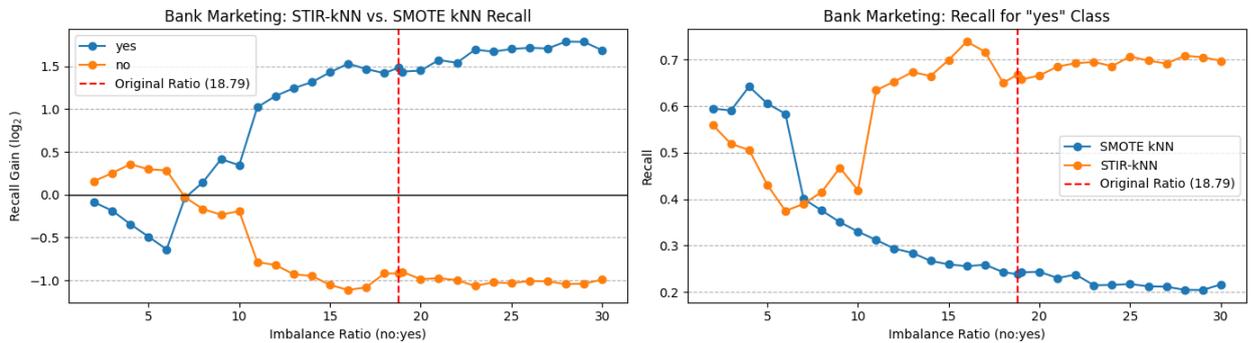


Figure 16. Recall comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Bank Marketing dataset

STIR-kNN preserves the original structural and statistical properties of the data by making all adjustments at inference time. This property is especially important in the Bank Marketing dataset, where the chronological ordering of observations reflects evolving customer behavior and campaign strategies. Synthetic oversampling in such settings risks distorting historical patterns and inflating minority prevalence in a manner that may not generalize to future operational conditions.

With respect to majority-class recall, STIR-kNN initially maintains higher performance than SMOTE-kNN at mild imbalance ratios. As imbalance severity increases, majority-class recall under STIR-kNN declines more noticeably, mirroring the behavior observed in the comparison with classical kNN. Nevertheless, this reduction remains stable across higher imbalance levels and represents a deliberate trade-off that is justified by the substantial gains in minority-class sensitivity in a domain where false negatives carry a significantly higher cost than false positives.

Figure 17 shows that the improved detection of minority-class instances comes with a moderate decline in minority-class precision. Rather than reflecting instability, this pattern is a direct consequence of the framework’s emphasis on sensitivity in highly skewed regions, where conservative decision rules would otherwise suppress rare but operationally significant outcomes. Importantly, minority-class precision under STIR-kNN remains well controlled and stabilizes at moderate imbalance ratios, indicating that the rebalancing mechanism does not lead to uncontrolled misclassification rates.

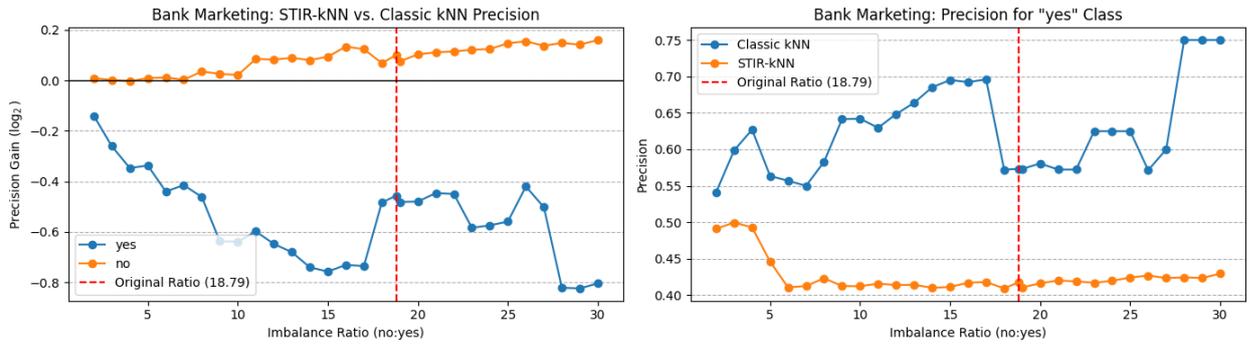


Figure 17. Precision comparison between STIR-kNN and classic kNN under varying class imbalance ratios on the Bank Marketing dataset

In contrast to the minority class, majority-class precision exhibits a different pattern. Relative to classical kNN, majority-class precision under STIR-kNN remains largely unchanged at lower imbalance ratios and shows a modest improvement as imbalance severity increases. This behavior indicates that inference-time rebalancing not only enhances minority-class recall but also yields more reliable majority-class predictions by limiting the misclassification of minority instances in majority-dominated regions. By comparison, classical kNN displays increasingly unstable precision behavior as minority recall collapses, emphasizing the advantage of controlled, evidence-driven adjustment at inference time.

Figure 18 illustrates the precision trends of STIR-kNN and SMOTE-kNN on the Bank Marketing dataset. Under mild imbalance conditions, STIR-kNN achieves higher minority-class precision than SMOTE-kNN, consistent with the conservative nature of inference-time rebalancing. As imbalance severity increases from moderate to extreme levels, minority-class precision under both methods converges, with SMOTE-kNN typically attaining slightly higher precision. This advantage, however, arises from training-time oversampling that increases minority prevalence in the feature space.

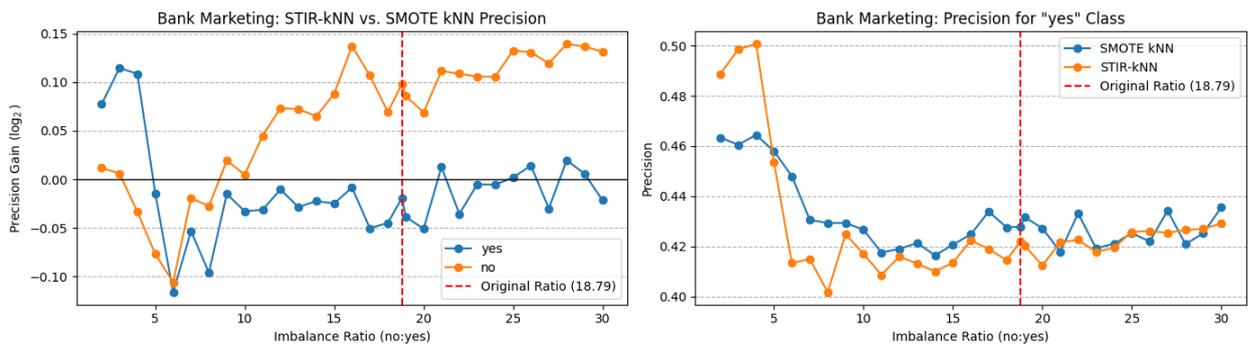


Figure 18. Precision comparison between STIR-kNN and SMOTE-kNN under varying class imbalance ratios on the Bank Marketing dataset

In comparison, majority-class precision follows an opposite trend. At lower imbalance ratios, STIR-kNN initially performs slightly worse than SMOTE-kNN. However, as imbalance severity increases, STIR-kNN exhibits improved and stabilized majority-class precision, ultimately surpassing SMOTE-kNN at higher imbalance ratios. This pattern indicates that inference-time rebalancing progressively suppresses incorrect minority assignments in majority-dominated regions, yielding more reliable majority-class decisions without altering the underlying data distribution.

As illustrated in Figure 19, STIR-kNN exhibits lower overall accuracy than both classical kNN and SMOTE-kNN, with the largest deviations observed at moderate to high imbalance ratios. This behavior is consistent with the model’s strong emphasis on minority detection in a domain where false negatives carry substantial operational cost.

Importantly, the observed reduction in accuracy stabilizes as imbalance severity increases and does not degrade proportionally with class imbalance. This pattern suggests that inference-time rebalancing induces structured, evidence-driven adjustments to local decision boundaries rather than uncontrolled label perturbations. In the context of marketing applications, where increased outreach to non-subscribing clients is preferable to missed subscription opportunities, the resulting accuracy trade-off remains operationally acceptable.

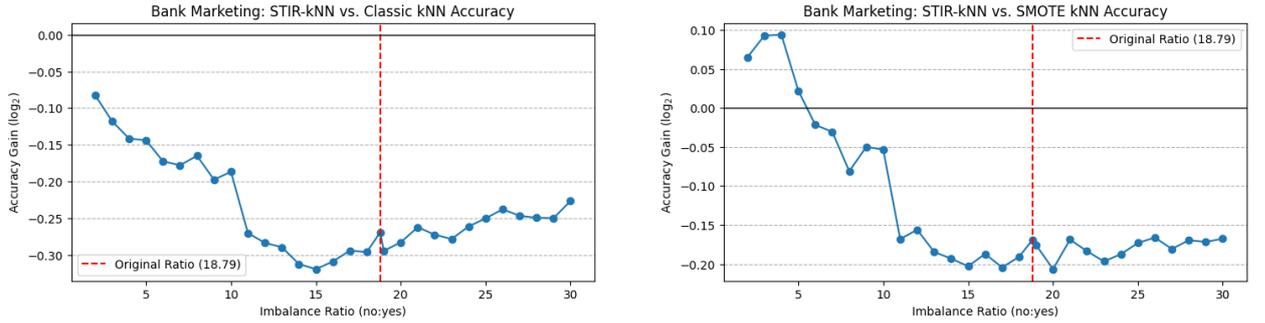


Figure 19. Overall accuracy on the Bank Marketing dataset across increasing imbalance ratios. Left: comparison with classical kNN. Right: comparison with SMOTE-kNN.

Overall, the Bank Marketing results demonstrate that STIR-kNN is particularly effective in high-risk, extremely imbalanced, and temporally structured domains where false negatives cause substantial operational costs. By performing inference-time rebalancing, the framework recovers a large proportion of minority-class instances that are systematically missed by classical kNN, even under severe class imbalance. These recall gains are accompanied by controlled and stable changes in precision, with minority-class precision remaining bounded and majority-class precision improving at higher imbalance levels. Although overall accuracy is reduced relative to classical kNN and SMOTE-kNN, this degradation remains stable and does not intensify with increasing imbalance, which is consistent with trends observed across previously examined datasets and indicates that performance gains are achieved through targeted, evidence-driven adjustments rather than indiscriminate label flipping. Collectively, these results further position STIR-kNN as a robust inference-time alternative to training-time oversampling methods in applications where preserving data integrity and prioritizing minority detection are critical.

4.2 Ablation Study: Contribution of STIR-kNN Components

To quantify the contribution of each component of the proposed STIR-kNN framework, we conduct a systematic ablation study in which individual phases of the method are selectively disabled. Specifically, we evaluate the impact of the probabilistic intervention gating, elastic evidence discovery, and stochastic label injection mechanisms by comparing the full model against variants in which each component is removed in isolation. In the following tables, these variants are denoted as No Gating, No Elastic, and No Injection, respectively.

The ablation analysis is performed across all three datasets under three imbalance settings, with imbalance ratios of 1.25, 10, and 25. These ratios are selected to represent mild, moderate, and

severe class imbalance, respectively, allowing us to examine how the contribution of each component evolves as imbalance increases. In particular, the near-balanced setting ($IR = 1.25$) provides a critical test of the gating mechanism, which is designed to suppress unnecessary intervention when class imbalance is low.

Since STIR-kNN is inherently stochastic, all experiments are repeated 50 times with different random seeds. Results are reported as averages over these runs, with the corresponding standard deviations shown in parentheses, ensuring statistically stable estimates and enabling fair and consistent comparisons between ablation variants.

4.2.1 Electrical Grid Stability

Table 2 summarizes the ablation results for the Electrical Grid Stability dataset across three imbalance regimes.

Near-balanced regime ($IR = 1.25$). Under mild class imbalance, disabling the intervention gating mechanism leads to the most substantial performance degradation, characterized by a pronounced reduction in majority-class recall and minority-class precision, alongside a modest increase in minority-class recall. This indicates that, in near-balanced settings, the gating mechanism plays a crucial role in suppressing unnecessary corrective intervention and preserving majority-class performance. In contrast, removing elastic evidence discovery or stochastic injection results in only negligible changes across all metrics, confirming that these components remain largely inactive when imbalance is low and do not introduce spurious bias.

Moderate imbalance ($IR = 10$). At a moderate imbalance ratio, the importance of elastic neighborhood expansion and stochastic injection becomes evident. Disabling elastic evidence discovery leads to severe losses in majority-class recall and minority-class precision, indicating that fixed-size neighborhoods are insufficient to recover minority-class evidence in this regime. Similarly, removing the stochastic injection mechanism causes a substantial drop in minority-class recall, accompanied by an increase in minority-class precision and majority-class recall. This reflects the expected precision–recall trade-off arising when minority-class reinforcement is absent, resulting in more conservative classification behavior. By comparison, disabling the gating mechanism produces only minor changes, suggesting that the intervention probability is already appropriately elevated at this imbalance level.

Severe imbalance ($IR = 25$). Under extreme class imbalance, elastic evidence discovery and stochastic injection dominate the behavior of STIR-kNN. Removing elastic expansion results in pronounced degradation across all metrics, particularly majority-class recall, demonstrating the necessity of adaptive neighborhood scaling in highly imbalanced regions of the feature space. The removal of stochastic injection produces the largest decline in minority-class recall observed across all settings, while simultaneously increasing minority-class precision and majority-class recall. This confirms that stochastic injection is the primary mechanism driving minority-class recovery under severe imbalance. As in the previous regimes, disabling the gating mechanism has only a marginal effect at high imbalance.

IR	Ablation	Δ Recall _{minority}	Δ Precision _{minority}	Δ Recall _{majority}	Δ Precision _{majority}
1.25	No Gating	+0.018 (0.005)	-0.093 (0.010)	-0.075 (0.008)	+0.009 (0.003)
	No Elastic	+0.001 (0.003)	-0.002 (0.007)	-0.001 (0.005)	+0.001 (0.002)
	No Injection	+0.000 (0.003)	+0.002 (0.007)	+0.001 (0.005)	+0.000 (0.002)
10	No Gating	+0.031 (0.012)	-0.010 (0.009)	-0.013 (0.007)	+0.017 (0.007)
	No Elastic	+0.037 (0.015)	-0.376 (0.007)	-0.617 (0.011)	-0.113 (0.018)
	No Injection	-0.228 (0.012)	+0.160 (0.009)	+0.115 (0.005)	-0.094 (0.006)
25	No Gating	+0.025 (0.007)	-0.002 (0.003)	-0.005 (0.002)	+0.015 (0.004)
	No Elastic	+0.095 (0.009)	-0.419 (0.003)	-0.824 (0.005)	-0.207 (0.047)
	No Injection	-0.488 (0.013)	+0.171 (0.009)	+0.126 (0.003)	-0.191 (0.006)

Table 2. Ablation results on the Electrical Grid Stability dataset. Paired differences (Δ) relative to the full STIR-kNN model (mean \pm std over 50 runs).

Overall, the ablation results on the Electrical Grid Stability dataset reveal a clear imbalance-dependent hierarchy among STIR-kNN components. Intervention gating is most critical in near-balanced conditions, while elastic neighborhood expansion and stochastic label injection become increasingly essential as class imbalance intensifies.

4.2.2 Census Income

Table 3 summarizes the ablation results for the Census Income dataset across the three imbalance regimes.

Near-balanced regime (IR = 1.25). Under mild class imbalance, the Census Income dataset exhibits behavior consistent with that observed for Electrical Grid Stability. Disabling the intervention gating mechanism results in the most pronounced performance degradation, marked by a substantial reduction in majority-class recall and minority-class precision, despite an increase in minority-class recall. This indicates that, as in the previous dataset, the gating mechanism is essential for preventing unnecessary intervention when class imbalance is low.

By contrast, removing elastic evidence discovery or stochastic injection leads to only negligible changes across all metrics, confirming that these components remain largely inactive in near-balanced conditions and do not introduce unintended bias.

Moderate imbalance (IR = 10). At a moderate imbalance ratio, the relative importance of elastic neighborhood expansion and stochastic injection becomes evident. Disabling elastic evidence discovery results in significant losses in majority-class recall and minority-class precision, demonstrating that fixed-size neighborhoods are insufficient to recover minority-class evidence in this dataset.

Similarly, removing the stochastic injection mechanism causes a pronounced decline in minority-class recall, accompanied by an increase in minority-class precision and majority-class recall. This reflects the expected conservative shift in predictions when minority-class reinforcement is absent. The impact of disabling the gating mechanism remains comparatively small, indicating that intervention is already frequent at this imbalance level.

Severe imbalance (IR = 25). Under severe class imbalance, elastic evidence discovery and stochastic injection dominate the behavior of STIR-kNN. Removing elastic expansion leads to substantial degradation across all metrics, particularly majority-class recall, highlighting the necessity of adaptive neighborhood scaling in highly imbalanced regions of the feature space.

The removal of stochastic injection produces the largest reduction in minority-class recall observed across all imbalance regimes, while simultaneously increasing minority-class precision and majority-class recall. This confirms that stochastic injection is the primary driver of minority-class recovery under extreme imbalance, at the cost of increased false positives. As in the Electrical Grid Stability dataset, disabling the gating mechanism has only a marginal effect at this imbalance level.

IR	Ablation	Δ Recall _{minority}	Δ Precision _{minority}	Δ Recall _{majority}	Δ Precision _{majority}
1.25	No Gating	+0.068 (0.006)	-0.079 (0.006)	-0.084 (0.005)	+0.016 (0.002)
	No Elastic	+0.001 (0.005)	-0.006 (0.003)	-0.004 (0.002)	-0.000 (0.002)
	No Injection	-0.000 (0.004)	+0.001 (0.004)	+0.001 (0.003)	-0.000 (0.001)
10	No Gating	+0.033 (0.013)	-0.013 (0.007)	-0.022 (0.006)	+0.008 (0.004)
	No Elastic	+0.053 (0.012)	-0.221 (0.006)	-0.391 (0.008)	-0.061 (0.005)
	No Injection	-0.230 (0.012)	+0.188 (0.011)	+0.151 (0.005)	-0.042 (0.003)
25	No Gating	+0.025 (0.004)	-0.002 (0.002)	-0.009 (0.001)	+0.008 (0.001)
	No Elastic	+0.194 (0.005)	-0.282 (0.002)	-0.740 (0.003)	-0.124 (0.023)
	No Injection	-0.485 (0.007)	+0.213 (0.009)	+0.189 (0.002)	-0.102 (0.002)

Table 3. Ablation results on the Census Income dataset. Paired differences (Δ) relative to the full STIR-kNN model (mean \pm std over 50 runs).

Overall, the ablation results on the Census Income dataset closely mirror those observed for Electrical Grid Stability, reinforcing the imbalance-dependent hierarchy among STIR-kNN components.

4.2.3 Bank Marketing

Table 4 summarizes the ablation results for the Bank Marketing dataset across three imbalance regimes.

Near-balanced regime (IR = 1.25). Under mild class imbalance, disabling the intervention gating mechanism results in the most substantial performance degradation, with a pronounced decrease in majority-class recall and minority-class precision, despite a large increase in minority-class recall. Compared to the previous datasets, this effect is more pronounced, indicating that the Bank Marketing dataset is particularly sensitive to unnecessary intervention in near-balanced settings. This highlights the importance of gating in preventing excessive corrective behavior when the underlying decision boundary is already relatively well-defined.

By contrast, removing elastic evidence discovery or stochastic injection leads to only negligible changes across all metrics, confirming that these components remain largely inactive when imbalance is mild.

Moderate imbalance (IR = 10). At a moderate imbalance ratio, the relative contributions of elastic neighborhood expansion and stochastic injection become apparent, although their effects are less extreme than in the Electrical Grid Stability and Census Income datasets. Disabling elastic evidence discovery leads to moderate losses in majority-class recall, indicating that fixed-size neighborhoods begin to limit minority evidence recovery.

Removing the stochastic injection mechanism causes a sharp reduction in minority-class recall, accompanied by an increase in majority-class recall and a modest improvement in majority-class precision. Compared to the previously reviewed datasets, the model exhibits a weaker conservative shift for the Bank Marketing dataset. Disabling the gating mechanism produces a comparable but slightly smaller effect, suggesting that both gating and injection contribute to balancing recall and precision in this regime.

Severe imbalance (IR = 25). Under severe class imbalance, elastic evidence discovery and stochastic injection dominate the behavior of STIR-kNN, although the magnitude of their effects differs from the previous datasets. Removing elastic expansion leads to a degradation in majority-class recall, confirming the necessity of adaptive neighborhood scaling in highly imbalanced regions. However, compared to the Electrical Grid Stability and Census Income datasets, the decline in majority-class recall is less severe, suggesting that the Bank Marketing dataset exhibits a more diffuse class structure in which expanded neighborhoods provide comparatively smaller gains.

The removal of stochastic injection produces the largest decline in minority-class recall observed across all datasets, accompanied by a dramatic increase in majority-class recall. This indicates that, for Bank Marketing, stochastic injection is the primary mechanism preventing the model from collapsing into majority-dominated predictions under extreme imbalance. As in the other datasets, disabling the gating mechanism has only a marginal effect at this imbalance level.

IR	Ablation	Δ Recall _{minority}	Δ Precision _{minority}	Δ Recall _{majority}	Δ Precision _{majority}
1.25	No Gating	+0.122 (0.007)	-0.039 (0.004)	-0.168 (0.006)	+0.005 (0.006)
	No Elastic	+0.001 (0.004)	-0.000 (0.003)	-0.001 (0.005)	+0.000 (0.003)
	No Injection	-0.001 (0.004)	-0.000 (0.003)	+0.000 (0.004)	-0.000 (0.003)
10	No Gating	+0.043 (0.010)	-0.001 (0.008)	-0.041 (0.010)	+0.001 (0.006)
	No Elastic	+0.041 (0.012)	-0.010 (0.009)	-0.056 (0.012)	-0.005 (0.006)
	No Injection	-0.231 (0.011)	+0.052 (0.016)	+0.244 (0.009)	+0.000 (0.005)
25	No Gating	+0.030 (0.003)	+0.000 (0.001)	-0.025 (0.002)	+0.009 (0.003)
	No Elastic	+0.255 (0.004)	-0.037 (0.002)	-0.356 (0.004)	-0.072 (0.033)
	No Injection	-0.627 (0.005)	+0.012 (0.016)	+0.541 (0.004)	-0.063 (0.003)

Table 4. Ablation results on the Bank Marketing dataset. Paired differences (Δ) relative to the full STIR-kNN model (mean \pm std over 50 runs).

Overall, the ablation results on the Bank Marketing dataset reinforce the imbalance-dependent hierarchy observed previously. At the same time, the results suggest that datasets exhibiting greater class overlap, such as Bank Marketing, are particularly sensitive to the intervention gating mechanism in near-balanced settings and to stochastic injection under severe imbalance.

4.3 Summary

Across all evaluated benchmarks, STIR-kNN consistently demonstrates its effectiveness as an inference-time strategy for addressing class imbalance in settings characterized by asymmetric misclassification costs. On the Electrical Grid Stability dataset, the framework improves minority-class recall in a purely numerical environment governed by smooth physical dynamics, while preserving stable precision and accuracy. In the Census Income dataset, which combines numerical features with high-cardinality categorical attributes, STIR-kNN effectively mitigates majority-dominated neighborhood effects, substantially improving minority-class recall while maintaining controlled precision and bounded accuracy degradation. The Bank Marketing dataset, which originates from a highly imbalanced distribution and incorporates temporal dependencies, further highlights the framework’s strengths - STIR-kNN recovers rare but operationally critical minority instances without altering the underlying data distribution or distorting historical patterns.

Importantly, across all datasets, improvements in minority-class recall under STIR-kNN are achieved through localized, evidence-driven adjustments rather than indiscriminate label perturbations. At mild to moderate imbalance levels, the training-time oversampling method SMOTE attains slightly better minority-class performance than STIR-kNN. As imbalance severity increases, however, STIR-kNN consistently surpasses SMOTE, maintaining higher minority-class recall and exhibiting more stable behavior.

These recall gains are accompanied by predictable trade-offs, including reductions in majority-class recall, moderate decreases in minority-class precision, and modest impacts on overall accuracy. Crucially, these effects remain gradual, bounded, and aligned with application-specific cost asymmetries, reflecting intentional and controlled rebalancing behavior. Compared to training-time oversampling approaches such as SMOTE, STIR-kNN provides a robust and computationally efficient alternative that preserves the original feature-space and temporal structure of the data while delivering competitive or superior minority-class sensitivity. Notably, STIR-kNN frequently improves majority-class precision at higher imbalance ratios, indicating more reliable majority-class predictions in majority-dominated regions.

Collectively, these results demonstrate that inference-time rebalancing provides a principled, computationally efficient alternative to training-time oversampling methods, preserving feature-space and temporal integrity while delivering competitive or superior performance in highly imbalanced classification tasks.

The ablation study further clarifies the mechanisms underlying these performance gains. Across all datasets, intervention gating is shown to be most critical in near-balanced regimes, where it prevents unnecessary corrective behavior and preserves majority-class performance. As class imbalance increases, elastic neighborhood expansion and stochastic label injection emerge as the dominant contributors to minority-class recovery, enabling STIR-kNN to identify sparse minority evidence and counteract majority-dominated neighborhoods. Notably, the relative strength of these components varies across datasets: while elastic expansion plays a decisive role in structured numerical settings such as Electrical Grid Stability, stochastic injection becomes particularly important in highly imbalanced and noisy domains such as Bank Marketing. These findings confirm that STIR-kNN’s improvements arise from targeted, imbalance-aware adaptations rather than from uniform reweighting or global data manipulation.

Table 5 summarizes the median precision, recall, and F1-score for both majority and minority classes across all imbalance ratios, providing a robust, aggregate view of model behavior that complements the detailed grid-stability analyses presented earlier. Median values are reported to

mitigate the influence of local fluctuations observed at extreme imbalance settings.

Across all three datasets, STIR-kNN consistently achieves the highest median minority-class recall, confirming that inference-time rebalancing effectively addresses systematic under-detection of rare classes. This trend is most pronounced in the Bank Marketing dataset, where STIR-kNN increases minority recall from 0.03 (classic kNN) and 0.26 (SMOTE-kNN) to 0.67, representing a substantial recovery of positive instances in a highly skewed and operationally sensitive domain. Correspondingly, STIR-kNN also attains the highest minority-class F1-score in this setting, indicating that recall gains are not achieved at the expense of excessive precision loss.

In the Grid Stability dataset, STIR-kNN improves minority recall relative to classical kNN (0.88 vs. 0.45) and closely matches the performance of SMOTE-kNN, while preserving the original training distribution. Although SMOTE-kNN attains the highest minority F1-score in this dataset, it does so through training-time oversampling, whereas STIR-kNN achieves competitive performance solely via inference-time adjustments.

For the Census Income dataset, STIR-kNN again delivers the strongest median minority recall (0.77), outperforming both classical kNN (0.28) and SMOTE-kNN (0.60). This improvement translates into the highest minority F1-score, highlighting the method’s effectiveness in heterogeneous, categorical-heavy settings where minority signals are weak and easily dominated by majority-class patterns.

Dataset	Method	Majority Class			Minority Class		
		Precision	Recall	F1	Precision	Recall	F1
Grid Stability	Classic kNN	0.75	1.00	0.86	0.99	0.45	0.62
	SMOTE-kNN	0.92	0.92	0.92	0.87	0.87	0.87
	STIR-kNN	0.92	0.87	0.89	0.80	0.88	0.83
Census Income	Classic kNN	0.82	0.98	0.89	0.82	0.28	0.42
	SMOTE-kNN	0.88	0.88	0.88	0.61	0.60	0.60
	STIR-kNN	0.92	0.78	0.84	0.52	0.77	0.62
Bank Marketing	Classic kNN	0.62	0.99	0.76	0.62	0.03	0.06
	SMOTE-kNN	0.62	0.78	0.70	0.43	0.26	0.32
	STIR-kNN	0.66	0.41	0.51	0.42	0.67	0.51

Table 5. Median performance across imbalance ratios for three classification methods. Best results per dataset and class are highlighted in bold

With respect to majority-class performance, classical kNN generally attains the highest recall, reflecting its strong bias toward the dominant class. In contrast, STIR-kNN exhibits a controlled reduction in majority recall across datasets, consistent with its intentional prioritization of minority sensitivity. Importantly, the majority-class precision under STIR-kNN remains competitive and, in several cases, improves relative to classical kNN, indicating that the rebalancing mechanism does not indiscriminately degrade majority predictions.

Overall, the median results demonstrate that STIR-kNN offers a favorable and consistent precision–recall trade-off across diverse imbalance regimes and data modalities. By operating exclusively at inference time, it achieves substantial improvements in minority-class detection without retraining or synthetic data generation, positioning it as a robust and practically deployable alternative to training-time oversampling methods such as SMOTE.

Conclusions and Recommendations

This work introduced STIR-kNN, a stochastic inference-time rebalancing framework for kNN designed for imbalanced classification under asymmetric misclassification costs. The core contribution is a lightweight probabilistic wrapper that modifies the local voting evidence at prediction time through three components: intervention gating, elastic evidence discovery, and stochastic label injection (see Section 2.3 and Algorithm 1). Unlike training-time approaches, STIR-kNN preserves the original training distribution and does not require retraining, making it particularly suitable for settings where altering the data distribution is undesirable.

Across all evaluated benchmarks, the empirical findings support three consistent conclusions. First, STIR-kNN reliably improves minority-class recall as imbalance severity increases, demonstrating that inference-time neighborhood rebalancing can counteract the majority-dominance failure mode of classical kNN. Second, the observed recall improvements stem from localized, evidence-driven adjustments rather than indiscriminate label perturbations. Interventions are activated only when minority-class evidence is present in the local neighborhood, and the strength of rebalancing increases sub-linearly with the imbalance ratio. Third, these improvements introduce predictable performance trade-offs, including reduced majority-class recall, moderate decreases in minority-class precision, and modest impacts on overall accuracy. Importantly, these effects remain gradual and bounded across imbalance regimes, consistent with the asymmetric-cost motivation of the method.

The comparative analysis also clarifies the regime in which STIR-kNN is most advantageous. At mild to moderate imbalance levels, SMOTE-based training-time oversampling can match or occasionally surpass STIR-kNN on minority-focused metrics, reflecting the benefits of increasing minority density during model fitting. However, under severe imbalance, STIR-kNN consistently provides higher and more stable minority-class recall than SMOTE-kNN, suggesting that inference-time rebalancing becomes increasingly effective when minority evidence is sparse and oversampling is less reliable. In addition, STIR-kNN frequently improves majority-class precision at higher imbalance ratios, indicating more reliable majority decisions in majority-dominated regions. These patterns collectively position STIR-kNN as a robust alternative when the primary objective is minority detection under increasing imbalance and when preserving the empirical structure of the data is important.

The results indicate that STIR-kNN is particularly appropriate when false negatives are substantially more costly than false positives, decision boundaries are locally ambiguous due to feature overlap or heterogeneous attributes, and training-time resampling may distort meaningful structure (e.g., socio-demographic relationships or chronological patterns). In such settings, STIR-kNN provides a practical mechanism to bias kNN toward minority sensitivity without modifying the training distribution or introducing synthetic samples.

Several limitations should be acknowledged. As STIR-kNN is stochastic by design, performance can vary across runs. During internal testing and exploratory experimentation, no substantial qualitative differences in performance trends were observed. Nevertheless, a comprehensive statistical evaluation across multiple random seeds remains an important direction for future work. In addition, the method introduces inference-time overhead due to conditional neighborhood expansion, which may be non-trivial for large-scale deployments without approximate nearest-neighbor search. Finally, although evidence gating constrains unnecessary interventions, minority-class precision may decrease in datasets characterized by significant class overlap, reflecting an inherent trade-off between enhanced minority detection and selective classification.

Overall, the findings of this thesis demonstrate that inference-time neighborhood rebalancing is a principled and practical mechanism for improving kNN in imbalanced binary classification settings. By preserving the empirical data distribution while selectively amplifying minority evidence at prediction time, STIR-kNN offers a compelling alternative to training-time oversampling in applications where minority detection and data integrity are critical.

Future Works

This work motivates several directions for further research. First, STIR-kNN should be compared against additional imbalance-aware kNN variants beyond SMOTE-kNN, including cost-sensitive kNN and kNN with balanced prior probabilities. Such baselines would provide a clearer understanding of when inference-time rebalancing is preferable to explicitly reweighting decisions at training or prediction time.

Second, the functional forms used in STIR-kNN offer opportunities for systematic improvement. The intervention probability function could be redesigned to activate earlier under moderate imbalance or to incorporate local uncertainty rather than relying solely on the global imbalance ratio. Similarly, the elastic neighborhood expansion could be adapted using alternative growth schedules or distance-weighted evidence discovery to better respect local geometry in high-dimensional spaces. The stochastic label injection rule could also be generalized to use distance-based weighting, class-conditional confidence, or calibrated targets derived from validation data, thereby improving precision control while retaining recall gains.

Third, future work should examine robustness and practical deployment aspects of STIR-kNN, including its sensitivity to the choice of the neighborhood size k , the influence of different distance metrics for mixed numerical and categorical features, and scalable implementations based on approximate nearest-neighbor search.

Declaration of Artificial Intelligence

This work was prepared with limited assistance from ChatGPT (OpenAI) for grammar and language checking, table formatting, and preliminary information searches.

References

- [1] Ibraheem M Alkhalwaldeh, Ibrahim Albalkhi, and Abdulqadir Jeprel Naswhan. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World journal of methodology*, 13(5):373, 2023.
- [2] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.
- [3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [5] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
- [6] S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(4):325–327, 1976.
- [7] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.
- [8] Jonatan Møller Nuutinen Gøttcke and Arthur Zimek. Handling class imbalance in k-nearest neighbor classification by balancing prior probabilities. In *International Conference on Similarity Search and Applications*, pages 247–261. Springer, 2021.
- [9] Rajib Kumar Halder, Mohammed Nasir Uddin, Md Ashraf Uddin, Sunil Aryal, and Ansam Khraisat. Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1):113, 2024.
- [10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.
- [11] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- [12] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The uci machine learning repository. <https://archive.ics.uci.edu>, 2023. Accessed: 2025-12-29.
- [13] Nikolaos Kouiroukidis and Georgios Evangelidis. The effects of dimensionality curse in high dimensional knn search. In *2011 15th Panhellenic Conference on Informatics*, pages 41–45, 2011.
- [14] Evan Kriminger, Jose C Principe, and Choudur Lakshminarayan. Nearest neighbor distributions for imbalanced classification. In *The 2012 international joint conference on neural networks (IJCNN)*, pages 1–5. IEEE, 2012.

- [15] Florian Pargent, Bernd Bischl, and Janek Thomas. A benchmark experiment on how to encode categorical features in predictive modeling. *München: Ludwig-Maximilians-Universität München*, 2019.
- [16] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
- [17] Claude Sammut and Geoffrey I. Webb, editors. *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017.
- [18] Shashank Singh and Justin T Khim. Optimal binary classification beyond accuracy. *Advances in Neural Information Processing Systems*, 35:18226–18240, 2022.
- [19] Bo Sun and Haiyan Chen. A survey of k nearest neighbor algorithms for solving the class imbalanced problem. *Wireless Communications and Mobile Computing*, 2021(1):5520990, 2021.
- [20] Shichao Zhang. Cost-sensitive knn classification. *Neurocomputing*, 391:234–242, 2020.
- [21] Shichao Zhang. Challenges in knn classification, 2022.

A Appendix

A.1 STIR-kNN

```
1 def stir_knn(knn, k, y_train, x_test, y_test, minority_class):
2
3     y_pred = knn.predict(x_test)
4     y_corrected = pd.Series(y_pred).copy()
5     _, indices = knn.kneighbors(x_test)
6
7     ratio = y_train.value_counts().max() / y_train.value_counts().min()
8     p_intervention = 1 - (1/ratio)
9     elastic_k = max(k + 1, int(k * np.sqrt(ratio)))
10    for i, (indexes, pred) in enumerate(zip(indices, y_pred)):
11        if pred == minority_class:
12            continue
13        if random.random() < p_intervention:
14            neighbor_classes = y_train.iloc[indexes].to_list()
15
16            has_minority = minority_class in neighbor_classes
17
18            if not has_minority:
19                _, wider_idx = knn.kneighbors(x_test.iloc[i:i+1], n_neighbors=elastic_k)
20                wider_labels = y_train.iloc[wider_idx[0]].to_list()
21                if minority_class in wider_labels:
22                    has_minority = True
23
24            if has_minority:
25                continuous_target = np.log2(ratio)
26                base = int(np.floor(continuous_target))
27                chance = continuous_target - base
28                total_to_inject = base + (1 if random.random() < chance else 0)
29                current_min = neighbor_classes.count(minority_class)
30                if current_min < total_to_inject:
31                    needed = total_to_inject - current_min
32                    maj_indices = [idx for idx, label in enumerate(neighbor_classes)
33                                 if label != minority_class]
34
35                    for _ in range(min(needed, len(maj_indices))):
36                        flip_idx = maj_indices.pop(random.randrange(len(maj_indices)))
37                        neighbor_classes[flip_idx] = minority_class
38
39            y_corrected.iloc[i] = random.choice(neighbor_classes)
```