The Role of Synthetic Data in Improving Supervised Learning Methods

The Case of Land Use/Land Cover Classification

PhD Examination

João Pedro Martins Ribeiro da Fonseca
jpfonseca@novaims.unl.pt

Supervisor: Doctor Fernando José Ferreira Lucas Bação (bacao@novaims.unl.pt)

NOVA Information Management School
Doctoral Program in Information Management
October 12th, 2023

NOVA IMS Information Management School

Agenda

1. **About me**
2. **Introduction**
3. Tabular and Latent Space Synthetic Data Generation: A **Literature Review**
4. **Geometric SMOTE** for Imbalanced Datasets **with Nominal and Continuous Features**
5. Improving Imbalanced **Land Cover Classification with K-means SMOTE**: Detecting and Oversampling Distinctive Minority Spectral Signatures
6. Increasing the Effectiveness of Active Learning: Introducing **Artificial Data Generation in Active Learning** for Land Use/Land Cover Classification
7. Improving **Active Learning** Performance Through the Use of **Data Augmentation**
8. **Conclusions**

NOVA IMS Information Management School

About me

Education:

- Bsc in Economics @ NOVA SBE (2016)
- Msc in Management @ NOVA SBE (2019)
- Msc in Information Management @ NOVA IMS (2019)

Academic Positions:

- Invited teaching assistant @ NOVA IMS (2019 – 2023)
- Researcher @ NOVA IMS (2019 – 2020)
- PhD Candidate @ NOVA IMS (2020 – 2023)
- Research Intern @ NYU (early 2023)
- Invited Assistant Professor @ NOVA IMS (Current)
- Visiting Research Scholar @ NYU (Current)

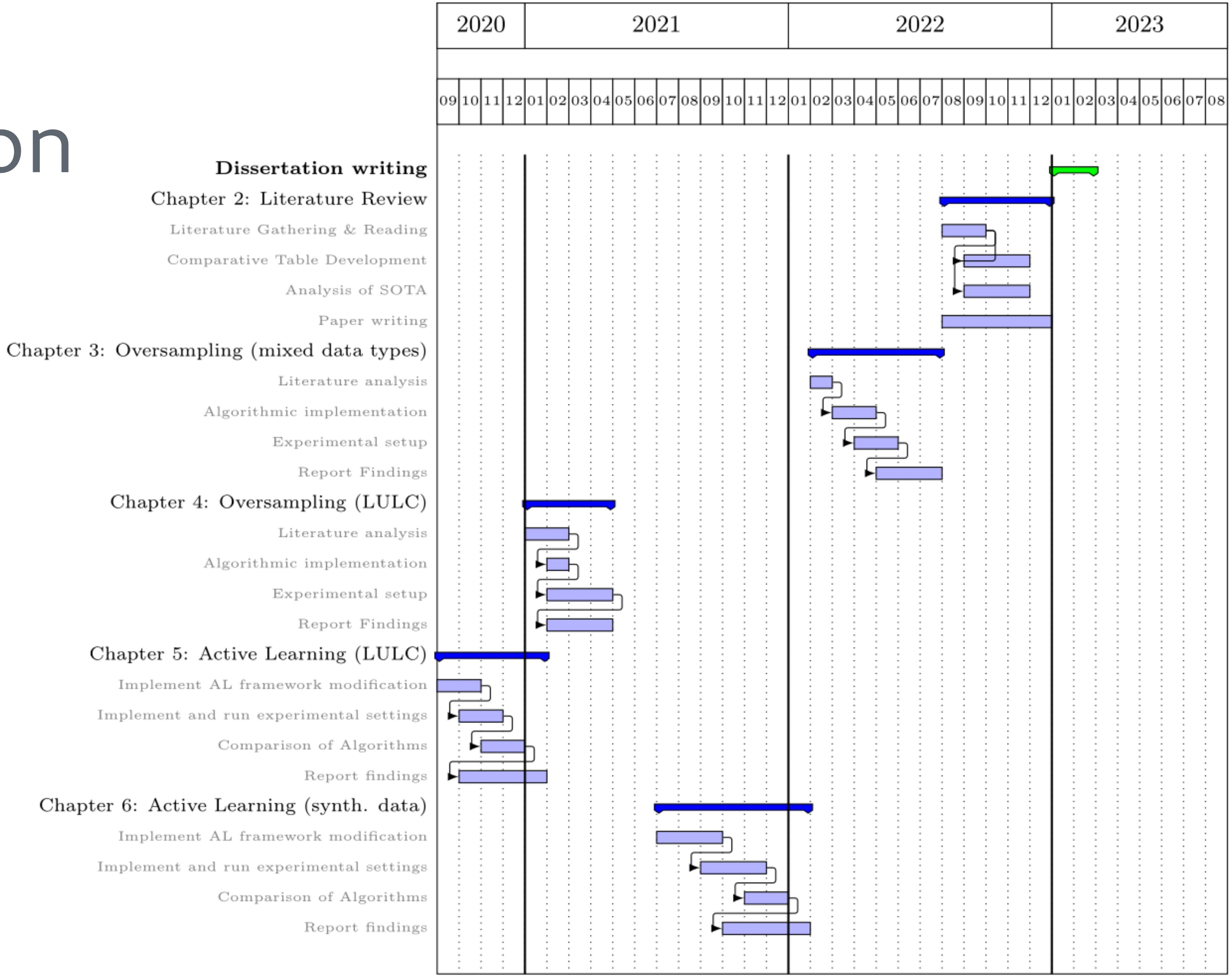
Full bio: joaopfonseca.github.io

Publications based on doctoral work

Chapter	RQ	Study Name	Current stage
2	1	Tabular and Latent Space Synthetic Data Generation: A Literature Review	~~Under Review~~ **Published in "Journal of Big Data"**
3	2	Geometric SMOTENC: A geometrically enhanced drop-in replacement for SMOTENC	~~Under Review~~ **Published in "Expert Systems with Applications"**
4	3	Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures	Published in the journal Information
5	4	Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification	Published in the journal Remote Sensing
6	4	Improving Active Learning Performance Through the Use of Data Augmentation	Published in International Journal of Intelligent Systems

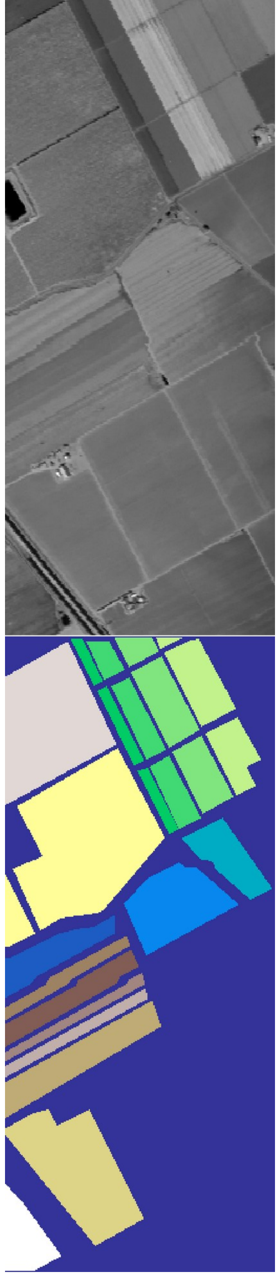
Introduction

- This dissertation covers three main topics:
  1. Synthetic Data Generation
  2. Active Learning
  3. Land Use/Land Cover Classification



Introduction - Land Use/Land Cover Classification

- LULC maps are essential for a variety of tasks
- Production of LULC maps are costly and rarely updated
- Automatic classification of LULC maps attempt to address this problem
- But it is hindered by several limitations (e.g., Imbalanced training data, labeling errors and annotation cost)



Salinas Dataset

Introduction - Active Learning

- Collecting large volumes of training data becomes a challenge when its annotation is labor and time intensive
- Active Learning (AL) reduces the cost for annotating training datasets
- AL aims to find the most informative observations to annotate, maximizing the predictive power of a classifier, given a certain annotation budget.
- AL faces two main challenges: **Consistency** and **Efficiency**

Introduction - Synthetic Data Generation

- Synthetic data is essential for several tasks:
  - Regularization (i.e., Data Augmentation)
  - Imbalanced learning (i.e., Oversampling)
  - Data anonymization
  - Semi-supervised learning
  - Self-supervised learning
- Expands the training dataset by introducing synthetic observations, based on a prior data distribution

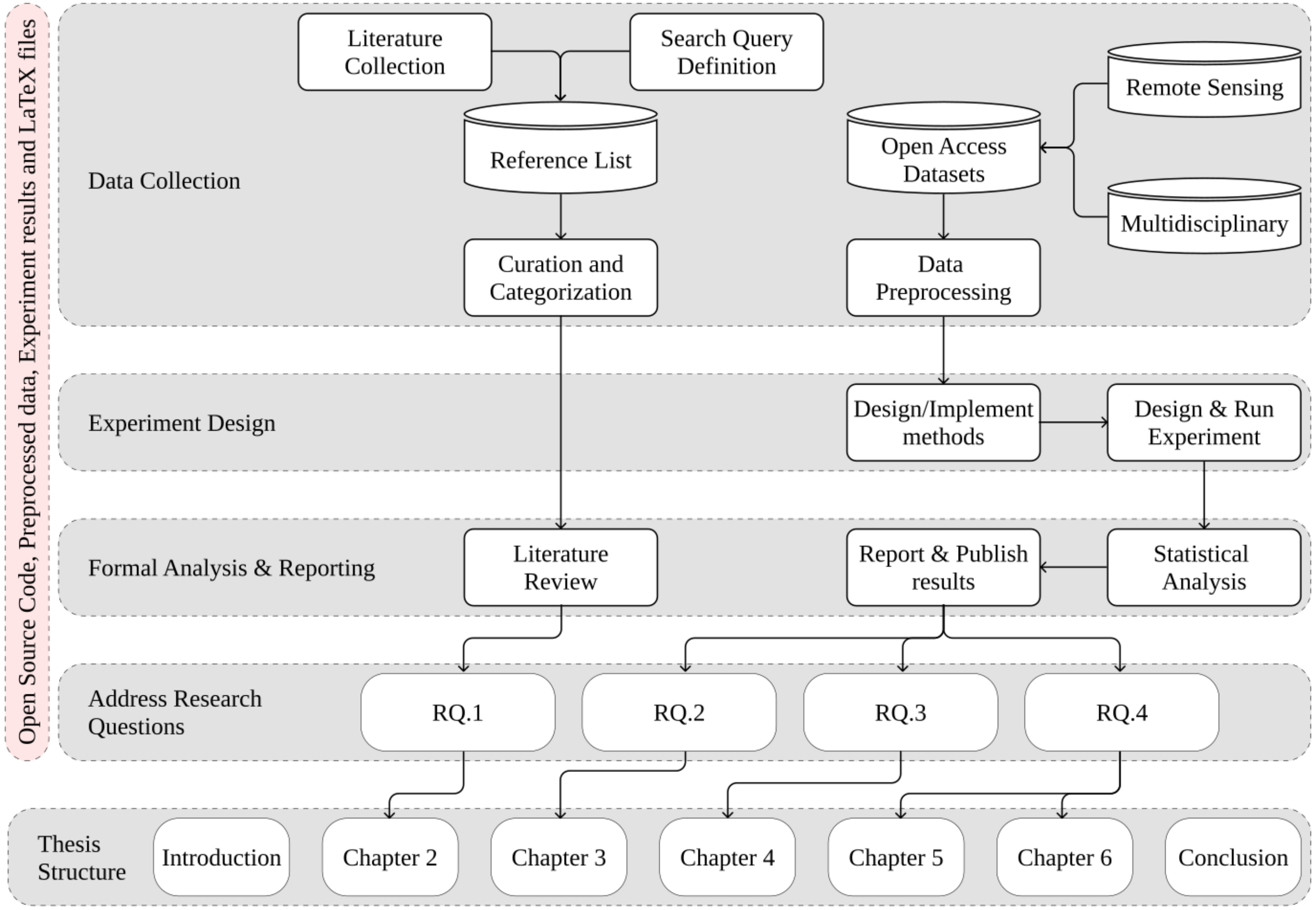
Introduction – Research Questions

1. What are the main research lines in synthetic data generation?
   - Development of a literature review to study existing synthetic data generation methods and the core fields where they are being used.
2. How can one oversample data with both continuous and categorical features?
   - Development of an improved oversampling method to be used with mixed data types.
3. How can the quality and consistency of automatic LULC mapping be enhanced?
   - Exploration of imbalanced learning methods in the context of LULC.
4. How can efficient automated LULC mapping be achieved with limited availability of ground-truth data?
   - Development of an improved active learning framework in the remote sensing domain using artificial data generation.

Introduction – State of Research

Key principles:

- Common methodological approach across chapters
- Use of publicly available data
- Ensure reproducibility of every step in the process
- Make every research output available to the community



Structure and methodological approach used in this dissertation.

Introduction – State of Research

Key principles:

- Common methodological approach across chapters
- Use of publicly available data
- Ensure reproducibility of every step in the process
- Make every research output available to the community

ML-Research

Build passing | codecov 92% | docs passing | code style black | python 3.8 | 3.9 | 3.10 | 3.11 | DOI 10.1155/2023/7941878

PyPI	pypi package 0.4.2 downloads 15k
Anaconda	conda-forge v0.4.2 downloads 7k

Installable via “pip” and “conda”:

```
pip install ml-research
```

```
conda install -c conda-forge ml-research
```

NOVA IMS Information Management School

Introduction – Methodological approach

- **Data collection**
- Performance metrics
- Experimental procedure
- Ranking tables
- Statistical analysis

- **Source (LULC datasets)**
  - Universidad del País Vasco - Grupo de Inteligencia Computacional
  - http://www.ehu.eus/ccwintco
- **Source (Multidisciplinary)**
  - UC Irvine Machine Learning Repository
  - https://archive.ics.uci.edu/

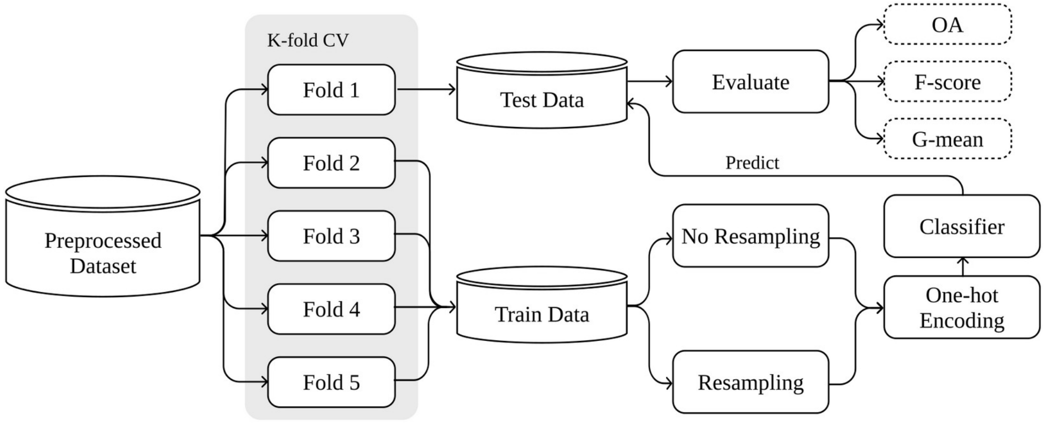
Introduction – Methodological approach

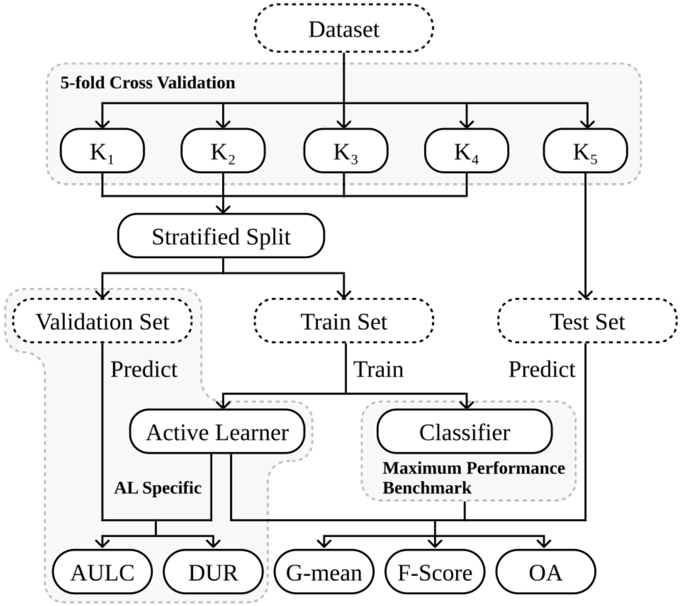
- Data collection
- **Performance metrics**
- Experimental procedure
- Ranking tables
- Statistical analysis

- **Standard performance metrics**
  - Overall Accuracy
  - Geometric mean score (root of the product of class-wise sensitivity/recall)
  - F1-Score
- **Active Learning metrics**
  - AULC – Area Under the Learning Curve
  - DUR – Data Utilization Rate

Introduction – Methodological approach

- Data collection
- Performance metrics
- **Experimental procedure**
- Ranking tables
- Statistical analysis





Introduction – Methodological approach

- Data collection
- Performance metrics
- Experimental procedure
- **Ranking tables**
- Statistical analysis

- **Mean rankings will be used to aggregate results across multiple datasets**
  - Fluctuations of performance scores between datasets make the analysis of mean scores less accurate.
  - The use of mean rankings are generally recommended in this scenario

Introduction – Methodological approach

- Data collection
- Performance metrics
- Experimental procedure
- Ranking tables
- **Statistical analysis**

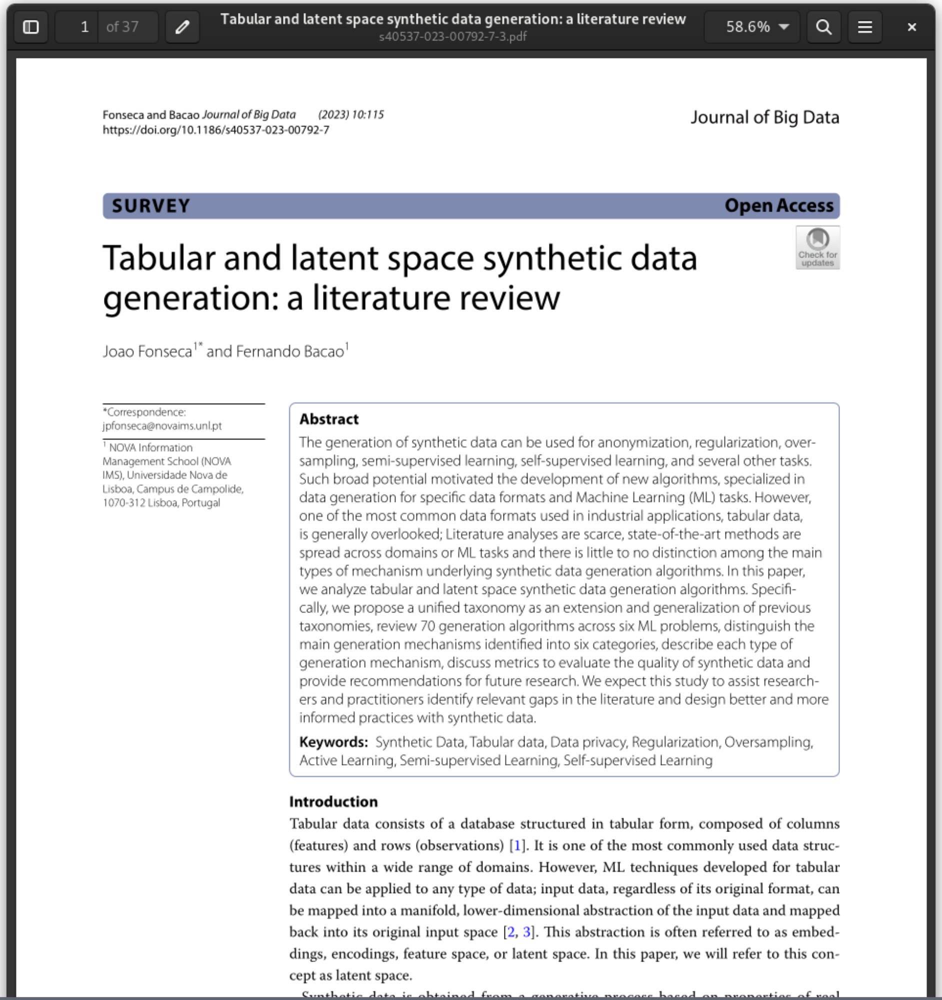
- **Methods used must account for the multiple comparison problem**
- **Two tests used:**
  - Friedman test: Understand whether there is a difference across methods
  - Holm-Bonferroni or Wilcoxon signed-rank test

NOVA IMS Information Management School

Tabular and Latent Space Synthetic Data Generation: A Literature Review

Published as:

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. Journal of Big Data, 10(1), 115.



Fonseca and Bacao *Journal of Big Data* (2023) 10:115
https://doi.org/10.1186/s40537-023-00792-7

Journal of Big Data

SURVEY Open Access

Tabular and latent space synthetic data generation: a literature review

Joao Fonseca[1*] and Fernando Bacao[1]

*Correspondence: jpfonseca@novaims.unl.pt

[1] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

Abstract

The generation of synthetic data can be used for anonymization, regularization, oversampling, semi-supervised learning, self-supervised learning, and several other tasks. Such broad potential motivated the development of new algorithms, specialized in data generation for specific data formats and Machine Learning (ML) tasks. However, one of the most common data formats used in industrial applications, tabular data, is generally overlooked; Literature analyses are scarce, state-of-the-art methods are spread across domains or ML tasks and there is little to no distinction among the main types of mechanism underlying synthetic data generation algorithms. In this paper, we analyze tabular and latent space synthetic data generation algorithms. Specifically, we propose a unified taxonomy as an extension and generalization of previous taxonomies, review 70 generation algorithms across six ML problems, distinguish the main generation mechanisms identified into six categories, describe each type of generation mechanism, discuss metrics to evaluate the quality of synthetic data and provide recommendations for future research. We expect this study to assist researchers and practitioners identify relevant gaps in the literature and design better and more informed practices with synthetic data.

Keywords: Synthetic Data, Tabular data, Data privacy, Regularization, Oversampling, Active Learning, Semi-supervised Learning, Self-supervised Learning

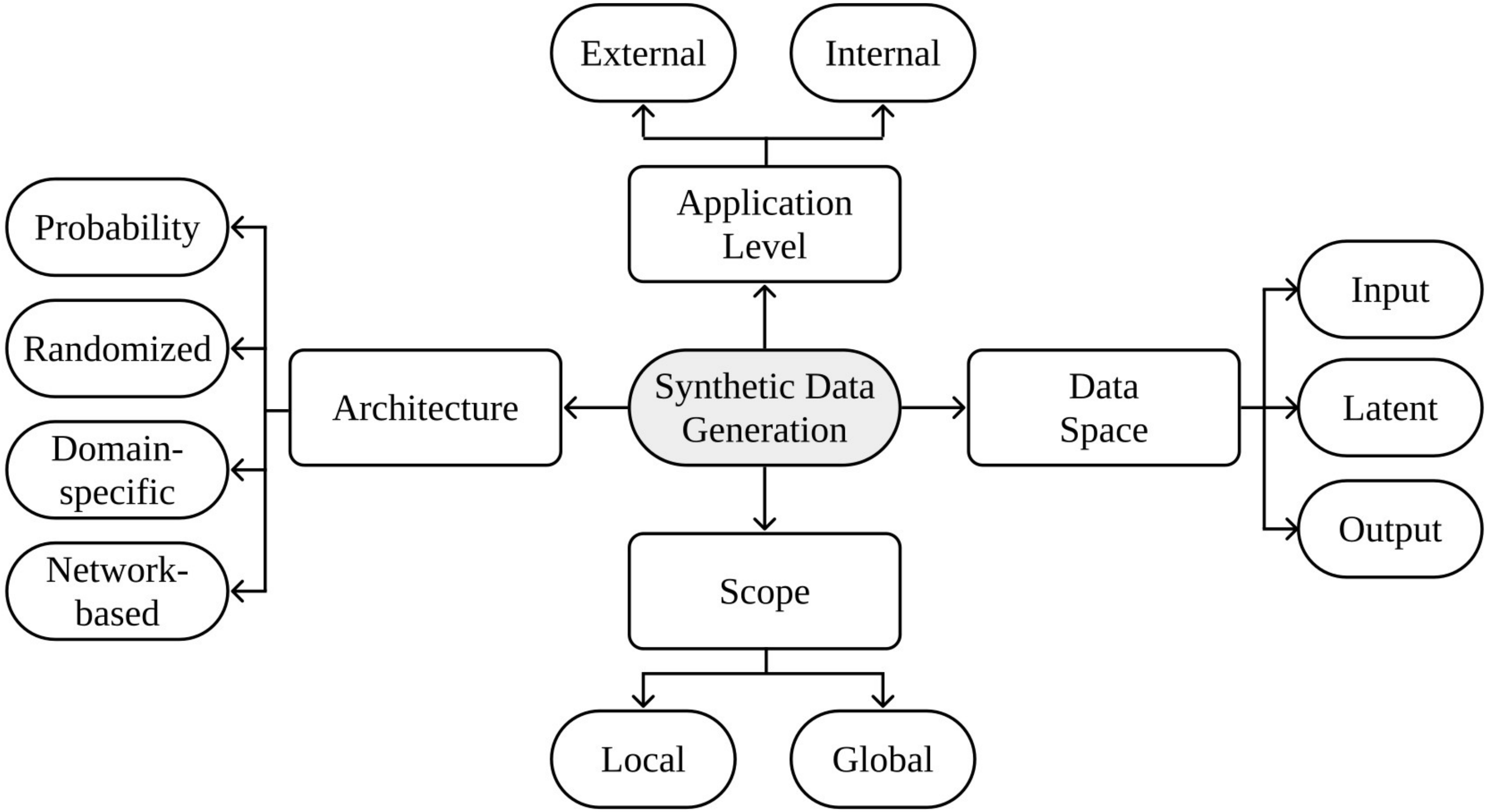
Introduction

Tabular data consists of a database structured in tabular form, composed of columns (features) and rows (observations) [1]. It is one of the most commonly used data structures within a wide range of domains. However, ML techniques developed for tabular data can be applied to any type of data; input data, regardless of its original format, can be mapped into a manifold, lower-dimensional abstraction of the input data and mapped back into its original input space [2, 3]. This abstraction is often referred to as embeddings, encodings, feature space, or latent space. In this paper, we will refer to this concept as latent space.

Tabular and Latent Space Synthetic Data Generation: A Literature Review

- The broad potential of **SDG** motivated the development of new algorithms, **specialized in data generation for specific data formats and ML tasks**.
- However, one of the most common data formats used in industrial applications, tabular data, is generally overlooked.
- In this paper, we analyze tabular and latent space synthetic data generation algorithms.
  - Propose **a unified and extended taxonomy** for SDG
  - **Review 70 SDG** algorithms across 6 ML tasks
  - Distinguish **generation mechanisms into 6 categories**
  - Review **performance metrics to evaluate the quality of synthetic data**

Tabular and Latent Space Synthetic Data Generation: A Literature Review



General taxonomy of data generation mechanisms proposed in this paper.

NOVA IMS Information Management School

Tabular and Latent Space Synthetic Data Generation: A Literature Review

Type	Mechanism	Smoothness	Manifold	Priv.	Reg.	Ovs.	AL	Semi-SL	Self-SL
Perturbation	Random	✓	✓	×	×	✓	×	×	×
	Laplace	✓	✓	✓	×	×	×	×	×
	Gaussian	✓	✓	✓	✓	×	×	✓	✓
	Swap-noise	×	×	×	×	×	×	✓	✓
	Zero-out noise	×	×	×	×	×	×	×	✓
PDF	Gaussian Gen.	×	✓	✓	×	✓	×	×	×
	Gaussian Mix.	×	✓	✓	×	✓	×	×	×
	KDE	×	✓	×	×	✓	×	×	×
PGM	Bayesian Net.	×	×	✓	✓	×	×	×	×
	Gibbs	×	×	×	✓	✓	×	×	×
	Random Walk	×	×	×	×	✓	×	×	×
Linear	Between-class Int.	×	✓	×	✓	×	✓	✓	×
	Within-class Int.	✓	✓	×	✓	✓	✓	✓	×
	Extrapolation	✓	✓	×	✓	✓	×	×	×
	Hard Extra.	✓	✓	×	✓	✓	×	×	×
	Inter.+Extra.	✓	✓	×	×	✓	×	×	×
	Difference Transf.	✓	✓	×	✓	×	×	×	×
Geometric	Hypersphere	✓	✓	×	×	✓	✓	×	×
	Triangular	✓	✓	×	×	×	×	✓	×
	Hyperrectangle	×	✓	×	✓	×	×	×	×
Neural nets.	GAN	×	×	✓	✓	✓	✓	×	×
	AE	×	×	×	✓	✓	✓	✓	×
Others	Exponential M.	×	×	✓	×	×	×	×	×
	Reconstruction err.	×	×	×	×	✓	×	×	×

Analysis of synthetic data generation mechanisms.

Tabular and Latent Space Synthetic Data Generation: A Literature Review

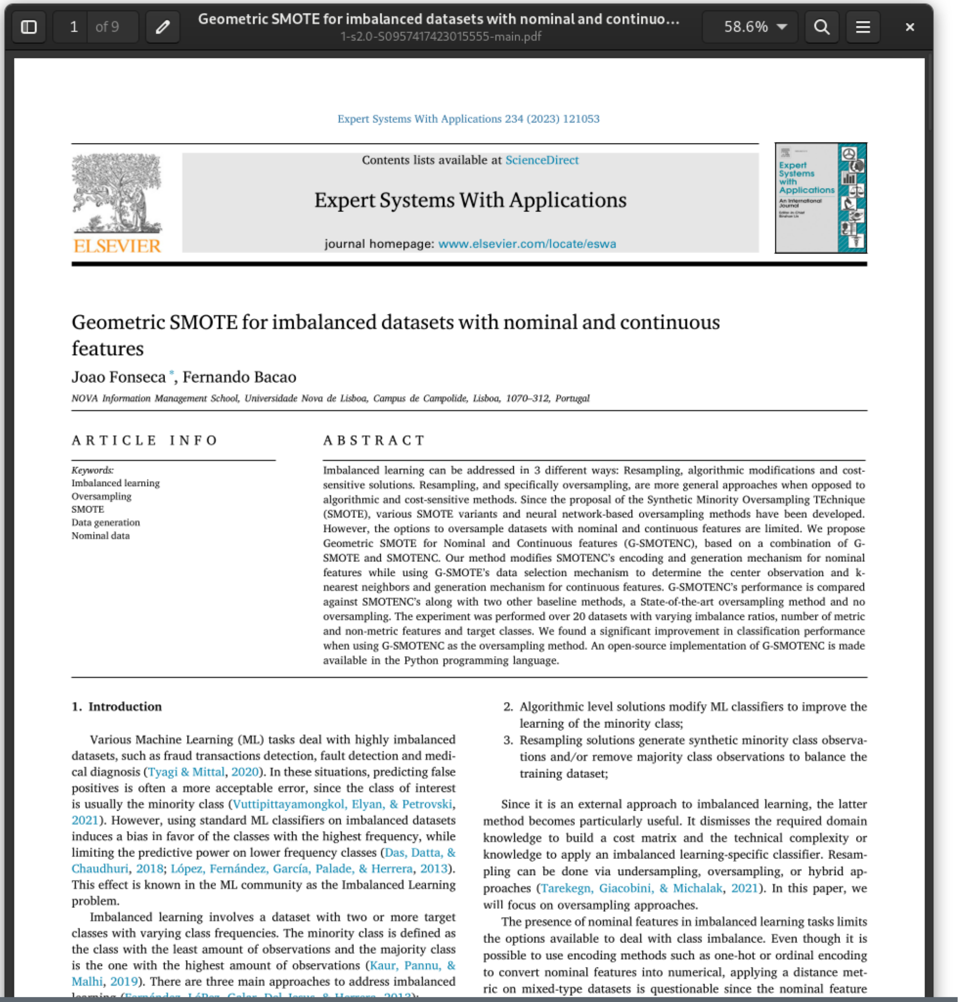
Conclusions:

- There are several approaches to generate synthetic data across a wide range of tasks
- However, there are several limitations in the literature regarding:
  - Latent space learning
  - Selection of generation mechanisms
  - Data privacy mechanisms
  - Analysis of quality of synthetic data for regularization techniques
  - Consistency and interpretability of generative neural network methods
  - Ensemble techniques for tabular data
  - Oversampling tabular data with mixed data types
  - Lack of research of synthetic data generation in tabular few-shot learning
  - Lack of research of the effect of synthetic data towards model fairness and bias.

Geometric SMOTE for Imbalanced Datasets with Nominal and Continuous Features

Published as:

Fonseca, J., & Bacao, F. (2023). Geometric SMOTE for imbalanced datasets with nominal and continuous features. Expert Systems with Applications, 234, 121053.



Expert Systems With Applications 234 (2023) 121053

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Geometric SMOTE for imbalanced datasets with nominal and continuous features

Joao Fonseca *, Fernando Bacao

NOVA Information Management School, Universidade Nova de Lisboa, Campus de Campolide, Lisboa, 1070–312, Portugal

ARTICLE INFO

Keywords:
Imbalanced learning
Oversampling
SMOTE
Data generation
Nominal data

ABSTRACT

Imbalanced learning can be addressed in 3 different ways: Resampling, algorithmic modifications and cost-sensitive solutions. Resampling, and specifically oversampling, are more general approaches when opposed to algorithmic and cost-sensitive methods. Since the proposal of the Synthetic Minority Oversampling TEchnique (SMOTE), various SMOTE variants and neural network-based oversampling methods have been developed. However, the options to oversample datasets with nominal and continuous features are limited. We propose Geometric SMOTE for Nominal and Continuous features (G-SMOTENC), based on a combination of G-SMOTE and SMOTENC. Our method modifies SMOTENC's encoding and generation mechanism for nominal features while using G-SMOTE's data selection mechanism to determine the center observation and k-nearest neighbors and generation mechanism for continuous features. G-SMOTENC's performance is compared against SMOTENC's along with two other baseline methods, a State-of-the-art oversampling method and no oversampling. The experiment was performed over 20 datasets with varying imbalance ratios, number of metric and non-metric features and target classes. We found a significant improvement in classification performance when using G-SMOTENC as the oversampling method. An open-source implementation of G-SMOTENC is made available in the Python programming language.

1. Introduction

Various Machine Learning (ML) tasks deal with highly imbalanced datasets, such as fraud transactions detection, fault detection and medical diagnosis (Tyagi & Mittal, 2020). In these situations, predicting false positives is often a more acceptable error, since the class of interest is usually the minority class (Vuttipittayamongkol, Elyan, & Petrovski, 2021). However, using standard ML classifiers on imbalanced datasets induces a bias in favor of the classes with the highest frequency, while limiting the predictive power on lower frequency classes (Das, Datta, & Chaudhuri, 2018; López, Fernández, García, Palade, & Herrera, 2013). This effect is known in the ML community as the Imbalanced Learning problem.

Imbalanced learning involves a dataset with two or more target classes with varying class frequencies. The minority class is defined as the class with the least amount of observations and the majority class is the one with the highest amount of observations (Kaur, Pannu, & Malhi, 2019). There are three main approaches to address imbalanced learning

2. Algorithmic level solutions modify ML classifiers to improve the learning of the minority class;
3. Resampling solutions generate synthetic minority class observations and/or remove majority class observations to balance the training dataset;

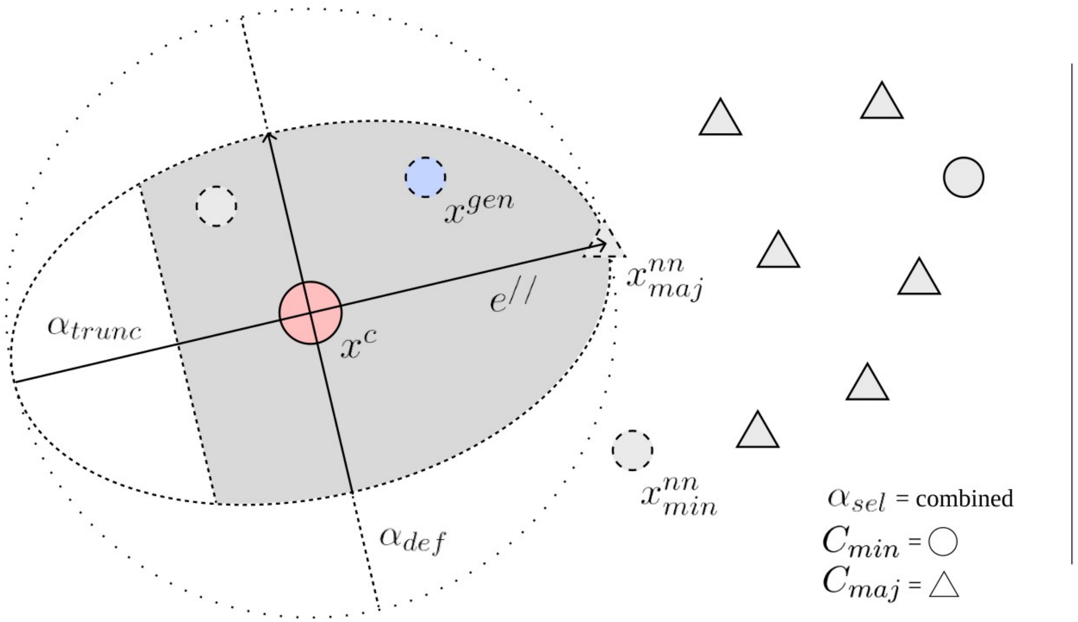
Since it is an external approach to imbalanced learning, the latter method becomes particularly useful. It dismisses the required domain knowledge to build a cost matrix and the technical complexity or knowledge to apply an imbalanced learning-specific classifier. Resampling can be done via undersampling, oversampling, or hybrid approaches (Tarekegn, Giacobini, & Michalak, 2021). In this paper, we will focus on oversampling approaches.

The presence of nominal features in imbalanced learning tasks limits the options available to deal with class imbalance. Even though it is possible to use encoding methods such as one-hot or ordinal encoding to convert nominal features into numerical, applying a distance metric on mixed-type datasets is questionable since the nominal feature

Geometric SMOTE for Imbalanced Datasets with Nominal and Continuous Features

- Since the proposal of the Synthetic Minority Oversampling TEchnique (SMOTE), various SMOTE variants and neural network-based oversampling methods have been developed.
- However, the options to oversample datasets with nominal and continuous features are limited.
- **We propose Geometric SMOTE for Nominal and Continuous features (G-SMOTENC)**
- Our method modifies SMOTENC's encoding and generation mechanism for nominal features while using G-SMOTE's data selection mechanism

Geometric SMOTE for Imbalanced Datasets with Nominal and Continuous Features



X^{nn}		f_1^{nom}	f_2^{nom}	f_3^{nom}
x^{nn}_{maj}	△	A	C	B
x^{nn}_{min}	○	B	B	B
n/a	○	B	C	A
x^{gen}_{nom}	○	B	C	B

Visual depiction of G-SMOTENC

Geometric SMOTE for Imbalanced Datasets with Nominal and Continuous Features

Classifier	Metric	G-SMOTENC	NONE	SMOTENC	ROS	RUS	SMOTE-ENC
DT	F-Score	**1.32 ± 0.11**	3.84 ± 0.40	3.13 ± 0.20	4.32 ± 0.19	5.47 ± 0.23	2.92 ± 0.34
DT	G-Mean	**1.68 ± 0.24**	5.84 ± 0.09	2.82 ± 0.21	2.95 ± 0.32	4.26 ± 0.32	3.45 ± 0.30
KNN	F-Score	**1.37 ± 0.16**	3.95 ± 0.35	3.11 ± 0.29	3.47 ± 0.36	5.53 ± 0.23	3.58 ± 0.23
KNN	G-Mean	**1.74 ± 0.17**	5.84 ± 0.12	2.89 ± 0.23	3.76 ± 0.33	3.00 ± 0.45	3.76 ± 0.23
LR	F-Score	**2.11 ± 0.24**	4.53 ± 0.35	2.37 ± 0.28	3.47 ± 0.32	5.21 ± 0.27	3.32 ± 0.38
LR	G-Mean	2.13 ± 0.26	6.00 ± 0.00	3.61 ± 0.21	**2.11 ± 0.23**	3.32 ± 0.40	3.84 ± 0.28
RF	F-Score	**1.32 ± 0.13**	5.05 ± 0.31	3.16 ± 0.22	3.05 ± 0.31	5.37 ± 0.14	3.05 ± 0.27
RF	G-Mean	**1.68 ± 0.22**	5.79 ± 0.21	3.26 ± 0.28	2.47 ± 0.30	3.89 ± 0.35	3.89 ± 0.19

Mean rankings over the different datasets (20), folds (5) and runs (3) used in the experiment.

NOVA IMS Information Management School

Geometric SMOTE for Imbalanced Datasets with Nominal and Continuous Features

Conclusions:

- The proposed method can be considered a generalization of the classical SMOTENC approach, since a specific parametrization of this algorithm will replicate SMOTENC's behavior
- The proposed method allows a significantly wider array of possibilities and higher variability in the synthetic data being generated.
- Can be applied before any type of categorical feature encoding.
- The proposed method consistently outperformed the remaining oversamplers, including SMOTENC and SMOTE-ENC

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Published as:

Fonseca, J., Douzas, G., Bacao, F. (2021). Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures. Information, 12(7), 266.

information MDPI

Article

Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Joao Fonseca *, Georgios Douzas and Fernando Bacao

NOVA Information Management School (NOVA IMS), Campus de Campolide, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal; gdouzas@novaims.unl.pt (G.D.); bacao@novaims.unl.pt (F.B.)
* Correspondence: jpfonseca@novaims.unl.pt

Abstract: Land cover maps are a critical tool to support informed policy development, planning, and resource management decisions. With significant upsides, the automatic production of Land Use/Land Cover maps has been a topic of interest for the remote sensing community for several years, but it is still fraught with technical challenges. One such challenge is the imbalanced nature of most remotely sensed data. The asymmetric class distribution impacts negatively the performance of classifiers and adds a new source of error to the production of these maps. In this paper, we address the imbalanced learning problem, by using K-means and the Synthetic Minority Oversampling Technique (SMOTE) as an improved oversampling algorithm. K-means SMOTE improves the quality of newly created artificial data by addressing both the between-class imbalance, as traditional oversamplers do, but also the within-class imbalance, avoiding the generation of noisy data while effectively overcoming data imbalance. The performance of K-means SMOTE is compared to three popular oversampling methods (Random Oversampling, SMOTE and Borderline-SMOTE) using seven remote sensing benchmark datasets, three classifiers (Logistic Regression, K-Nearest Neighbors and Random Forest Classifier) and three evaluation metrics using a five-fold cross-validation approach with three different initialization seeds. The statistical analysis of the results show that the proposed method consistently outperforms the remaining oversamplers producing higher quality land cover classifications. These results suggest that LULC data can benefit significantly from the use of more sophisticated oversamplers as spectral signatures for the same class can vary according to geographical distribution.

Keywords: LULC classification; imbalanced learning; oversampling; data augmentation; clustering

check for updates

Citation: Fonseca, J.; Douzas, G.; Bacao, F. Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures. *Information* **2021**, *12*, 266. https://doi.org/10.3390/info12070266

Academic Editor: Amir H. Gandomi

Received: 7 May 2021
Accepted: 25 June 2021
Published: 29 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in

1. Introduction

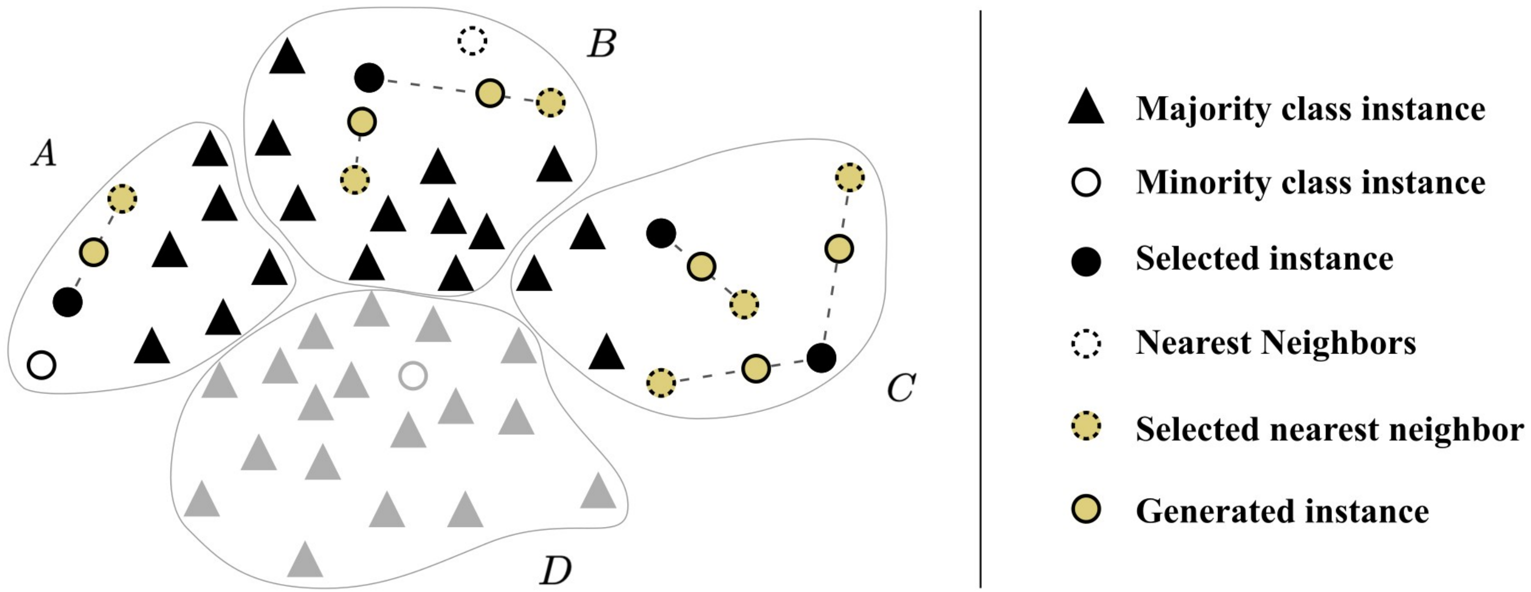
The increasing amount of remote sensing missions granted the access to dense time series (TS) data at a global level and provides up-to-date, accurate land cover information [1].

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

- Land cover maps are a critical tool to support informed policy development, planning, and resource management decisions
- However, the automatic production of Land Use/Land Cover maps is fraught with several limitations
- One such challenge is the imbalanced nature of most remotely sensed data
- **We address the imbalanced learning problem, by using K-means and the Synthetic Minority Oversampling TEchnique (K-means SMOTE)**

NOVA IMS Information Management School

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures



Example of K-means SMOTE's data generation process.

NOVA IMS Information Management School

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
LR	Accuracy	0.906 ± 0.039	0.904 ± 0.04	0.904 ± 0.04	0.901 ± 0.04	**0.909 ± 0.038**
LR	F-score	0.891 ± 0.041	0.893 ± 0.042	0.893 ± 0.042	0.890 ± 0.042	**0.898 ± 0.04**
LR	G-mean	0.936 ± 0.025	0.940 ± 0.025	0.940 ± 0.025	0.937 ± 0.025	**0.943 ± 0.024**
KNN	Accuracy	0.879 ± 0.043	0.865 ± 0.048	0.867 ± 0.05	0.862 ± 0.054	**0.881 ± 0.045**
KNN	F-score	0.859 ± 0.05	0.853 ± 0.049	0.861 ± 0.047	0.851 ± 0.053	**0.866 ± 0.048**
KNN	G-mean	0.919 ± 0.03	0.920 ± 0.029	0.926 ± 0.027	0.918 ± 0.03	**0.927 ± 0.027**
RF	Accuracy	0.898 ± 0.032	0.901 ± 0.031	0.900 ± 0.031	0.898 ± 0.032	**0.905 ± 0.031**
RF	F-score	0.879 ± 0.041	0.885 ± 0.037	0.887 ± 0.036	0.883 ± 0.037	**0.891 ± 0.036**
RF	G-mean	0.930 ± 0.024	0.935 ± 0.022	0.937 ± 0.021	0.935 ± 0.021	**0.939 ± 0.02**

Mean cross-validation scores of oversamplers.

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Conclusions:

- A distinctive characteristic of LULC classification is the potential for some classes to contain significantly different spectral signatures (e.g., forests composed majorly by pine trees vs spruce trees)
- Clustering-based synthetic data generation assists in distinguishing these differences within a minority class among clusters and avoid the generation of noisy synthetic data
- To do this, we introduced K-means SMOTE in LULC, which proved to be effective across several LULC classification problems

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification

Published as:

Fonseca, J., Douzas, G., Bacao, F. (2021). Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification. Remote Sensing, 13(13), 2619.

remote sensing
MDPI

Article

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification

Joao Fonseca *, Georgios Douzas and Fernando Bacao

Campus de Campolide, NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal; gdouzas@novaims.unl.pt (G.D.); bacao@novaims.unl.pt (F.B.)
* Correspondence: jpfonseca@novaims.unl.pt

Abstract: In remote sensing, Active Learning (AL) has become an important technique to collect informative ground truth data "on-demand" for supervised classification tasks. Despite its effectiveness, it is still significantly reliant on user interaction, which makes it both expensive and time consuming to implement. Most of the current literature focuses on the optimization of AL by modifying the selection criteria and the classifiers used. Although improvements in these areas will result in more effective data collection, the use of artificial data sources to reduce human–computer interaction remains unexplored. In this paper, we introduce a new component to the typical AL framework, the data generator, a source of artificial data to reduce the amount of user-labeled data required in AL. The implementation of the proposed AL framework is done using Geometric SMOTE as the data generator. We compare the new AL framework to the original one using similar acquisition functions and classifiers over three AL-specific performance metrics in seven benchmark datasets. We show that this modification of the AL framework significantly reduces cost and time requirements for a successful AL implementation in all of the datasets used in the experiment.

Keywords: active learning; artificial data generation; land use/land cover classification; oversampling; SMOTE

check for updates

Citation: Fonseca, J.; Douzas, G.; Bacao, F. Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification . *Remote Sens.* **2021**, *13*, 2619. https://doi.org/10.3390/rs13132619

Academic Editors: Danfeng Hong and Dino Ienco

Received: 26 April 2021
Accepted: 30 June 2021
Published: 4 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in

1. Introduction

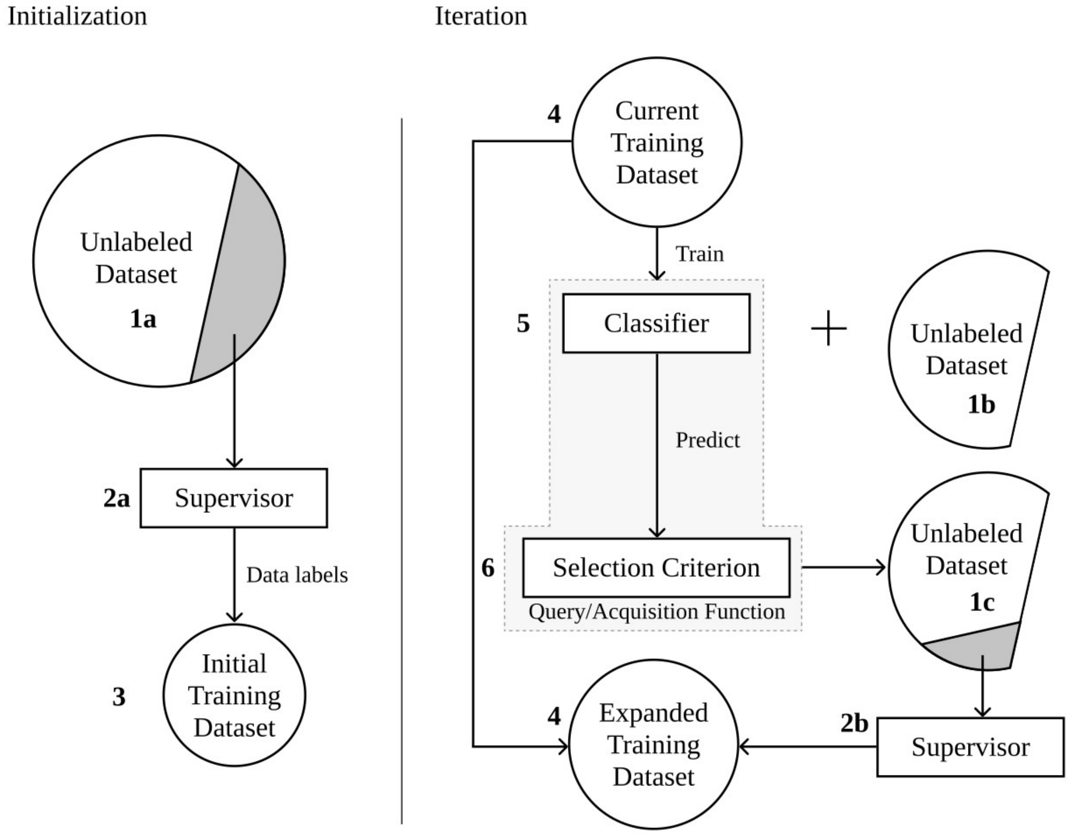
The technological development of air and spaceborne sensors and the increasing number of remote sensing missions have allowed the continuous collection of large amounts of high quality remotely sensed data. These data are often composed of multi- and hyperspectral satellite imagery, essential for numerous applications, such as Land Use/Land Cover (LULC) change detection, ecosystem management [1], agricultural management [2], water resource management [3], forest management, and urban monitoring [4]. Despite LULC maps being essential for most of these applications, their production is still a challenging task [5,6]. They can be updated using one of the following strategies:

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification

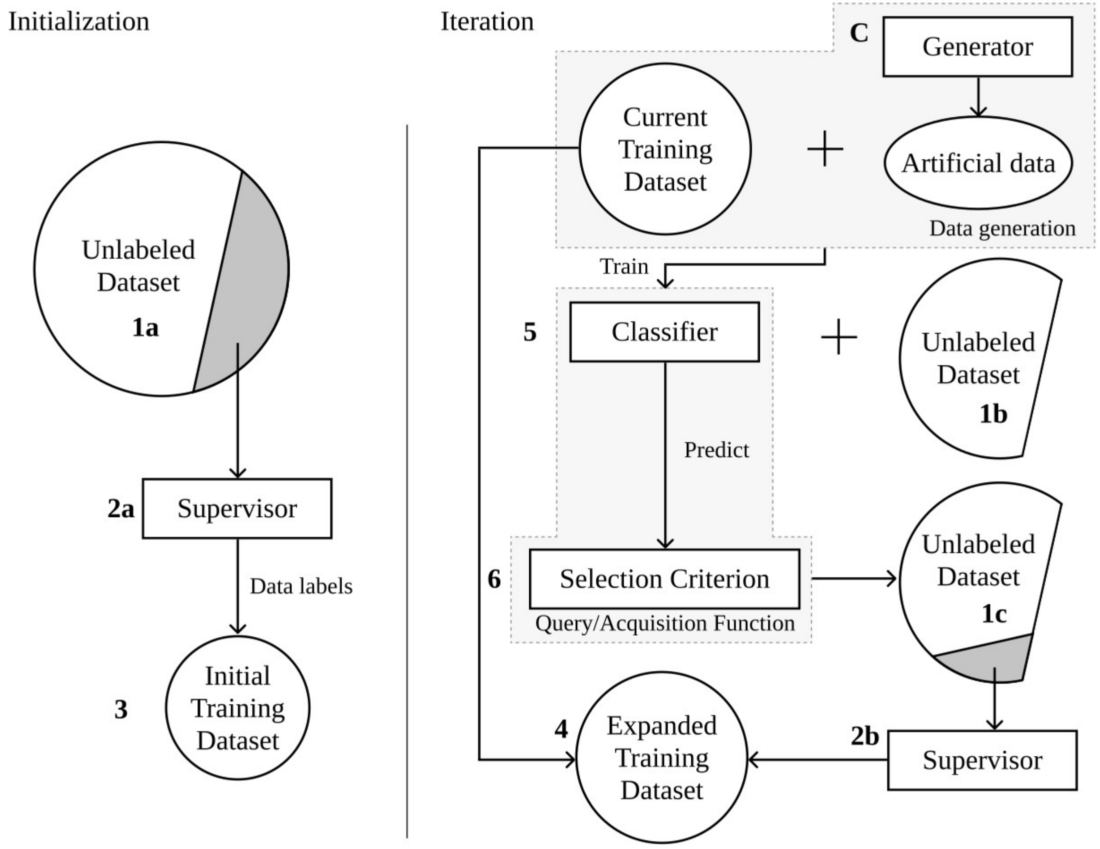
- In remote sensing, Active Learning (AL) has become an important technique to collect informative ground truth data.
- It relies on user interaction, making it expensive and time consuming.
- Most of the current literature attempt to optimize AL by modifying the selection criteria and the classifiers used.
- **In this paper, we introduce a new component to the typical AL framework, the data generator.**

NOVA IMS Information Management School

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification



Typical AL framework



Proposed method

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification

Classifier	Evaluation Metric	Standard	Proposed
KNN	F-score	0.762 ± 0.131	**0.794 ± 0.123**
KNN	G-mean	0.864 ± 0.079	**0.886 ± 0.073**
LR	F-score	0.839 ± 0.119	**0.843 ± 0.116**
LR	G-mean	0.907 ± 0.074	**0.911 ± 0.071**
RF	F-score	0.810 ± 0.109	**0.819 ± 0.1**
RF	G-mean	0.890 ± 0.068	**0.901 ± 0.059**

Average AULC of each AL configuration tested.

Classifier	Evaluation Metric	MP	Standard	Proposed
KNN	F-score	0.838 ± 0.106	0.835 ± 0.115	**0.843 ± 0.105**
KNN	G-mean	0.907 ± 0.063	0.904 ± 0.069	**0.912 ± 0.061**
LR	F-score	**0.890 ± 0.084**	0.883 ± 0.096	0.887 ± 0.097
LR	G-mean	0.935 ± 0.052	0.931 ± 0.059	**0.938 ± 0.055**
RF	F-score	0.859 ± 0.083	0.866 ± 0.081	**0.869 ± 0.08**
RF	G-mean	0.918 ± 0.051	0.921 ± 0.051	**0.930 ± 0.043**

Optimal classification scores.

NOVA IMS Information Management School

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification

Conclusions:

- The proposed AL framework was one of the first methods to implement synthetic data into AL with tabular data
- The proposed framework showed a significant reduction of the amount of labeled data required to reach any desired level of performance
- This method may be used to reduce the labeling cost when preparing a training dataset without affecting classification performance

NOVA IMS Information Management School

Improving Active Learning Performance Through the Use of Data Augmentation

Published as:

Fonseca, J., & Bacao, F. (2023). Improving Active Learning Performance through the Use of Data Augmentation. International Journal of Intelligent Systems, 2023.

IJIS7941878 1..17
7941878.pdf
1 of 17 58.2%

Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 7941878, 17 pages
https://doi.org/10.1155/2023/7941878

WILEY | Hindawi

Research Article

Improving Active Learning Performance through the Use of Data Augmentation

Joao Fonseca and Fernando Bacao

NOVA Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

Correspondence should be addressed to Joao Fonseca; jpfonseca@novaims.unl.pt

Received 25 August 2022; Revised 12 December 2022; Accepted 27 December 2022; Published 20 February 2023

Academic Editor: Frederick E. Petry

Active learning (AL) is a well-known technique to optimize data usage in training, through the interactive selection of unlabeled observations, out of a large pool of unlabeled data, to be labeled by a supervisor. Its focus is to find the unlabeled observations that, once labeled, will maximize the informativeness of the training dataset, therefore reducing data-related costs. The literature describes several methods to improve the effectiveness of this process. Nonetheless, there is a paucity of research developed around the application of artificial data sources in AL, especially outside image classification or NLP. This paper proposes a new AL framework, which relies on the effective use of artificial data. It may be used with any classifier, generation mechanism, and data type and can be integrated with multiple other state-of-the-art AL contributions. This combination is expected to increase the ML classifier's performance and reduce both the supervisor's involvement and the amount of required labeled data at the expense of a marginal increase in computational time. The proposed method introduces a hyperparameter optimization component to improve the generation of artificial instances during the AL process as well as an uncertainty-based data generation mechanism. We compare the proposed method to the standard framework and an oversampling-based active learning method for more informed data generation in an AL context. The models' performance was tested using four different classifiers, two AL-specific performance metrics, and three classification performance metrics over 15 different datasets. We demonstrated that the proposed framework, using data augmentation, significantly improved the performance of AL, both in terms of classification performance and data selection efficiency (all the codes and preprocessed data developed for this study are available at https://github.com/joaopfonseca/publications/).

1. Introduction

The importance of training robust ML models with minimal data requirements is substantially increasing [1–3]. Al-

(SSL) and active learning (AL). Even though they address the same problem, the two follow different approaches. SSL focuses on observations with the most certain predictions, whereas AL focuses on observations with the least certain

Improving Active Learning Performance Through the Use of Data Augmentation

- There is a paucity of research developed around the application of artificial data sources in AL, especially outside image classification or NLP.
- We propose an improved AL framework, which relies on the effective use of artificial data.
- It may be used with any classifier, generation mechanism and data type, and can be integrated with multiple other state-of-the-art AL contributions.
- The proposed method introduces a hyperparameter optimization component to improve the generation of artificial instances during the AL process, as well as an uncertainty-based data generation mechanism.

Improving Active Learning Performance Through the Use of Data Augmentation

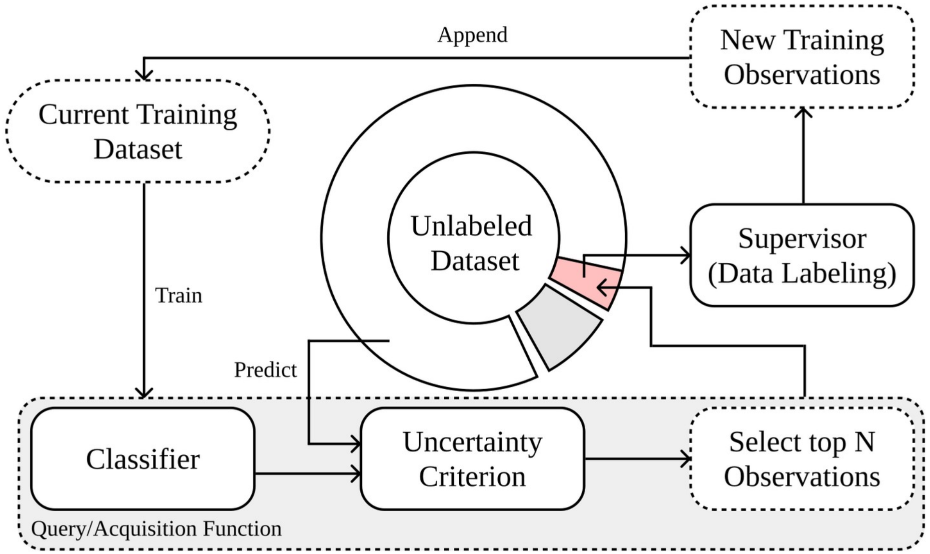


Diagram depicting an AL iteration.

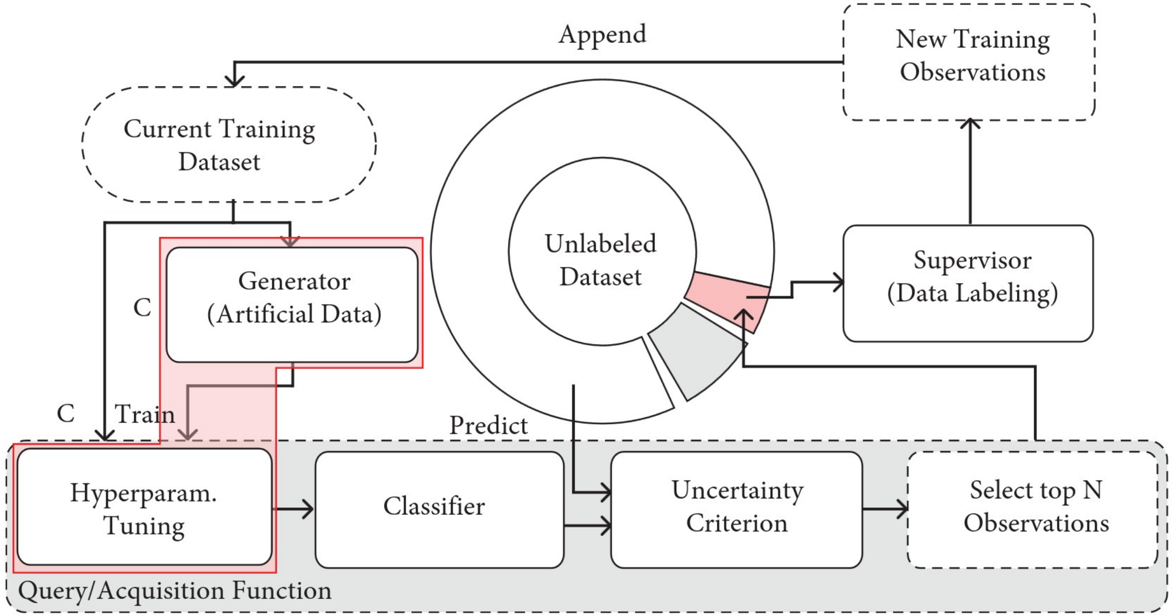


Diagram depicting the proposed AL iteration. The proposed modifications are comprised within the red polygon and marked with a boldface "C."

Improving Active Learning Performance Through the Use of Data Augmentation

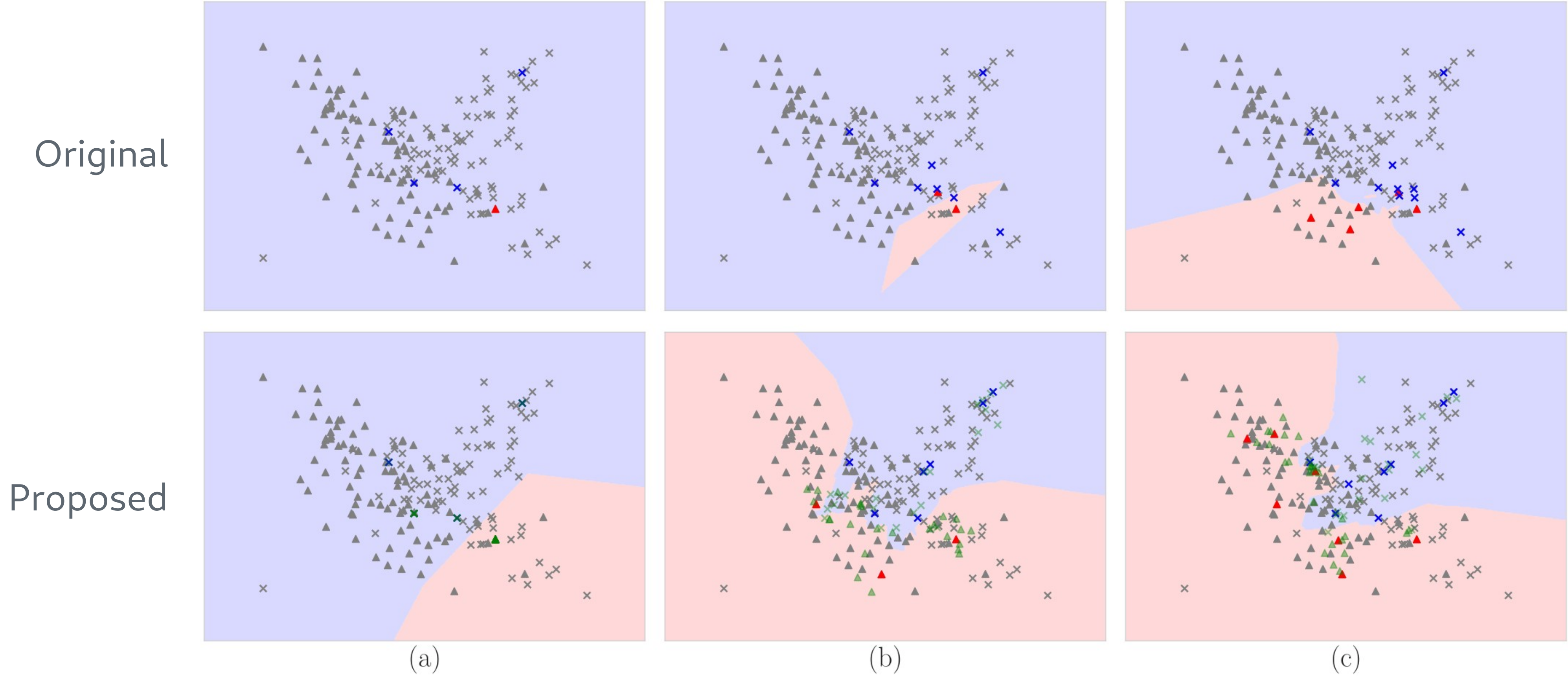


Illustration of the different acquisition processes in AL (**rows**) using a K-Nearest Neighbors classifier and Shannon's entropy as the uncertainty estimation function, with five observations being collected and labeled per iteration (**columns**).

NOVA IMS Information Management School

Improving Active Learning Performance Through the Use of Data Augmentation

Classifier	Evaluation Metric	Standard	Oversampling	Proposed
DT	Accuracy	2.13 ± 0.96	2.40 ± 0.49	**1.47 ± 0.62**
DT	F-score	2.47 ± 0.81	2.20 ± 0.40	**1.33 ± 0.70**
DT	G-mean	2.73 ± 0.57	1.93 ± 0.44	**1.33 ± 0.70**
KNN	Accuracy	2.07 ± 0.93	2.07 ± 0.68	**1.87 ± 0.81**
KNN	F-score	2.47 ± 0.81	1.87 ± 0.50	**1.67 ± 0.87**
KNN	G-mean	2.87 ± 0.34	**1.47 ± 0.50**	1.67 ± 0.70
LR	Accuracy	2.13 ± 0.88	2.20 ± 0.65	**1.67 ± 0.79**
LR	F-score	2.80 ± 0.40	1.87 ± 0.50	**1.33 ± 0.70**
LR	G-mean	2.80 ± 0.40	1.80 ± 0.54	**1.40 ± 0.71**
RF	Accuracy	2.27 ± 0.85	**1.87 ± 0.50**	**1.87 ± 0.96**
RF	F-score	2.73 ± 0.57	1.80 ± 0.54	**1.47 ± 0.72**
RF	G-mean	2.87 ± 0.34	**1.53 ± 0.50**	1.60 ± 0.71

Mean rankings of the AULC metric over the different datasets (15), folds (5), and runs (3) used in the experiment.

Improving Active Learning Performance Through the Use of Data Augmentation

Conclusions:

- The proposed method consists of a generalization of the AL framework previously proposed regarding the domain of application and data generation policy
- The proposed AL framework further reduced the amount of data labeling required to achieve a comparable classification performance
- The performance of the resulting classifiers also outperformed the ones trained on the entire dataset

Closing remarks

- The main focus of these questions was to address common scenarios where typical ML techniques will not work as intended:
  1. Imbalanced learning
  2. Supervised learning with scarcity of labeled data
- The main contributions of this dissertation can be summarized as follows:
  1. Address the limitation of oversampling on datasets with mixed data types
  2. Improvement of AL frameworks using on synthetic data
  3. Improvement of imbalanced LULC classification via oversampling

Current and future work

Selected recent working papers (not part of the PhD Thesis):

- Bell, A.*, **Fonseca, J.***, Abrate, C., Bonchi, F., Stoyanovich, J. (2023). **Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity**. Working paper.
- **Fonseca, J.** (2023). **A General Purpose Synthetic Data Generation Ensemble Mechanism for Fully, Semi and Self-Supervised Learning.** Working paper.
- Pliatsika, V., **Fonseca, J.**, Wang, T., Stoyanovich, J. (2023). **ShaRP: Explaining Rankings with Shapley Values**. Under submission.
- **Fonseca, J.***, Bell, A.*, Abrate, C., Bonchi, F., Stoyanovich, J. (2023). **Setting the Right Expectations: Algorithmic Recourse Over Time**. Accepted for the third ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO'23).

*** – Equal contribution**

Thank you!

Questions?

NOVA Information Management School
Doctoral Program in Information Management
October 12th, 2023

Appendix

Quick access links:

Chapter 1 - Introduction
Chapter 2 - Literature Review
Chapter 3 - G-SMOTENC
Chapter 4 - K-means SMOTE (LULC)
Chapter 5 - Active Learning (LULC)
Chapter 6 - Active Learning (Synth. Data)

NOVA Information Management School
Doctoral Program in Information Management

October 12th, 2023

Chapter 1 – Introduction

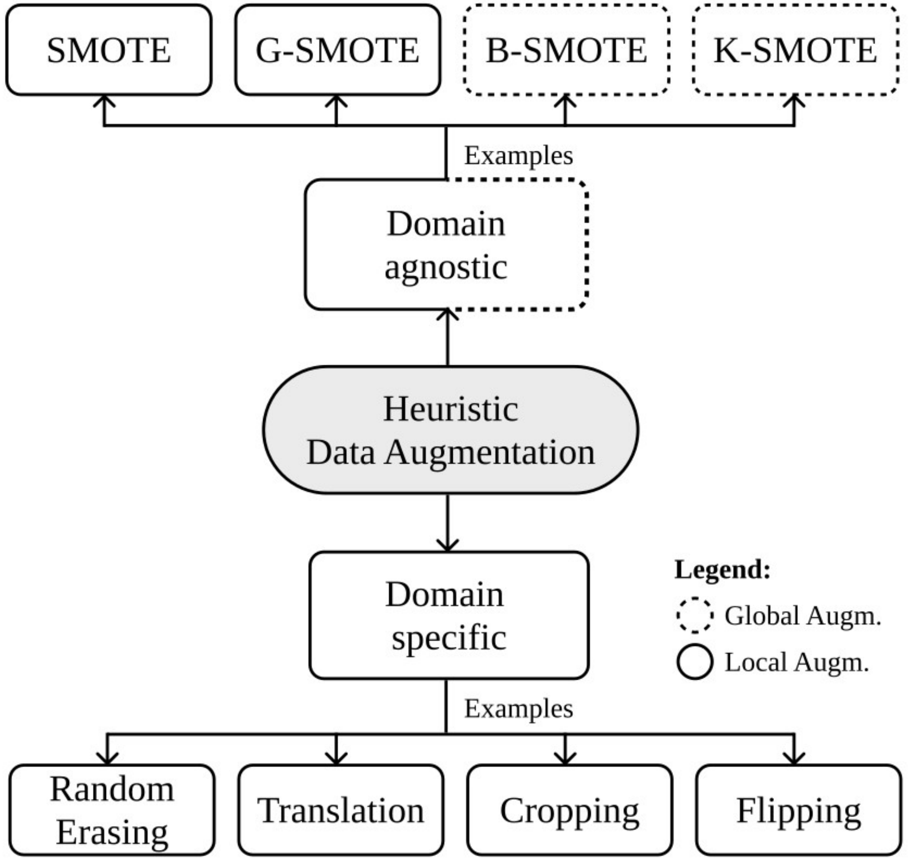


Figure 1.1.: Schema containing a general Heuristic Data Augmentation taxonomy.

Chapter 1 – Introduction

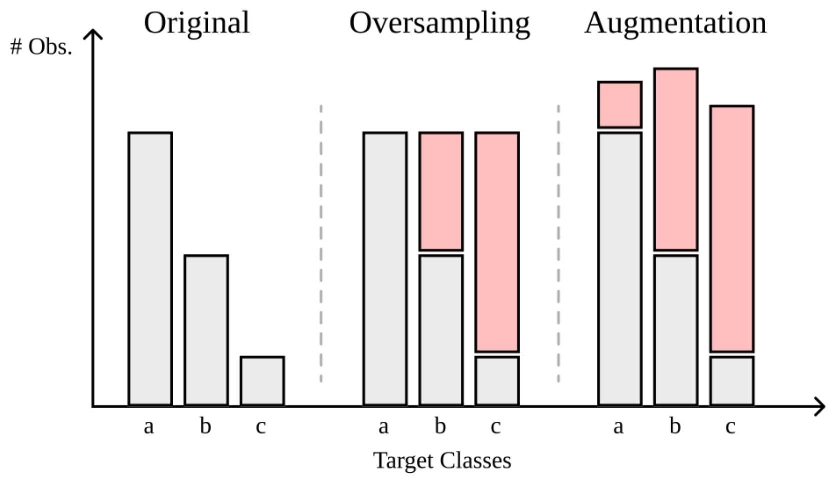


Figure 1.2.: Examples of data augmentation Strategies. The salmon-colored bars represent artificial data using the normal oversampling (center group) and an example of augmentation (right group) strategies.

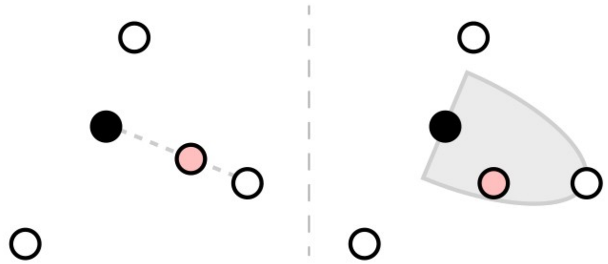


Figure 1.3.: Examples of data generation using SMOTE and G-SMOTE. In this example, both G-SMOTE's deformation and truncation parameters assume values around 0.5.

Chapter 1 – Introduction

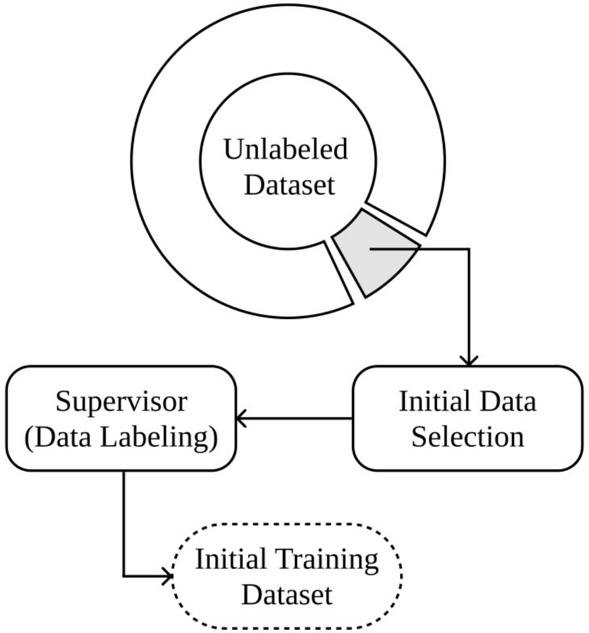


Figure 1.4.: Diagram depicting an AL initialization.

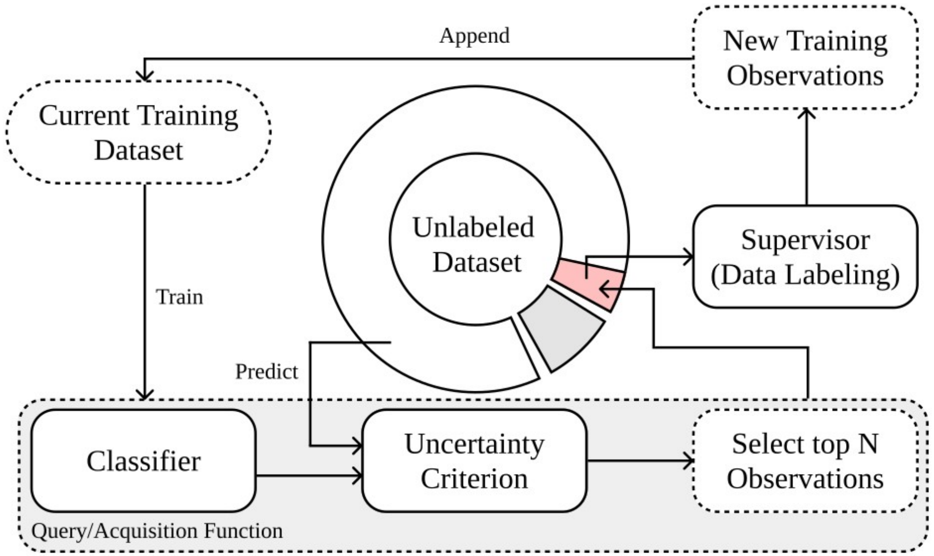


Figure 1.5.: Diagram depicting an AL iteration. In the first iteration, the training set collected during the initialization process becomes the "Current Training Dataset".

Chapter 2 – Literature review

Table 2.3.: Analysis of synthetic data generation mechanisms.

Type	Mechanism	Smoothness	Manifold	Priv.	Reg.	Ovs.	AL	Semi-SL	Self-SL
Perturbation	Random	✓	✓	×	×	✓	×	×	×
	Laplace	✓	✓	✓	×	×	×	×	×
	Gaussian	✓	✓	✓	✓	×	×	✓	✓
	Swap-noise	×	×	×	×	×	×	✓	✓
	Zero-out noise	×	×	×	×	×	×	×	✓
PDF	Gaussian Gen.	×	✓	✓	×	✓	×	×	×
	Gaussian Mix.	×	✓	✓	×	✓	×	×	×
	KDE	×	✓	×	×	✓	×	×	×
PGM	Bayesian Net.	×	×	✓	✓	×	×	×	×
	Gibbs	×	×	×	✓	✓	×	×	×
	Random Walk	×	×	×	×	✓	×	×	×
Linear	Between-class Int.	×	✓	×	✓	×	✓	✓	×
	Within-class Int.	✓	✓	×	✓	✓	✓	✓	×
	Extrapolation	✓	✓	×	✓	✓	×	×	×
	Hard Extra.	✓	✓	×	✓	✓	×	×	×
	Inter.+Extra.	✓	✓	×	×	✓	×	×	×
	Difference Transf.	✓	✓	×	✓	×	×	×	×
Geometric	Hypersphere	✓	✓	×	×	✓	✓	×	×
	Triangular	✓	✓	×	×	×	×	✓	×
	Hyperrectangle	×	✓	×	✓	×	×	×	×
Neural nets.	GAN	×	×	✓	✓	✓	✓	×	×
	AE	×	×	×	✓	✓	✓	✓	×
Others	Exponential M.	×	×	✓	×	×	×	×	×
	Reconstruction err.	×	×	×	×	✓	×	×	×

Chapter 2 – Literature review

ID	A	B	C
1	0.27	0.77	0.99
2	0.89	0.23	0.48
3	0.53	0.66	0.31
4	0.12	0.91	0.65
5	0.64	0.01	0.10

Swap-noise: $\epsilon = \begin{bmatrix} 0.53 & 0.77 & 0.10 \end{bmatrix} \longrightarrow x^s = \begin{bmatrix} 0.89 & 0.77 & 0.10 \end{bmatrix}$

Zero-out: $\epsilon = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \longrightarrow x^s = \begin{bmatrix} 0.89 & 0 & 0 \end{bmatrix}$

Gaussian: $\epsilon = \begin{bmatrix} 0.89 & 0.23 & 0.48 \\ -0.13 & +0.09 & +0.01 \end{bmatrix} \longrightarrow x^s = \begin{bmatrix} 0.89 & 0.32 & 0.49 \end{bmatrix}$

$x_2 = \begin{bmatrix} 0.89 & 0.23 & 0.48 \end{bmatrix}$ $m = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$

Figure 2.2.: Examples of synthetic observations generated with different masking approaches.

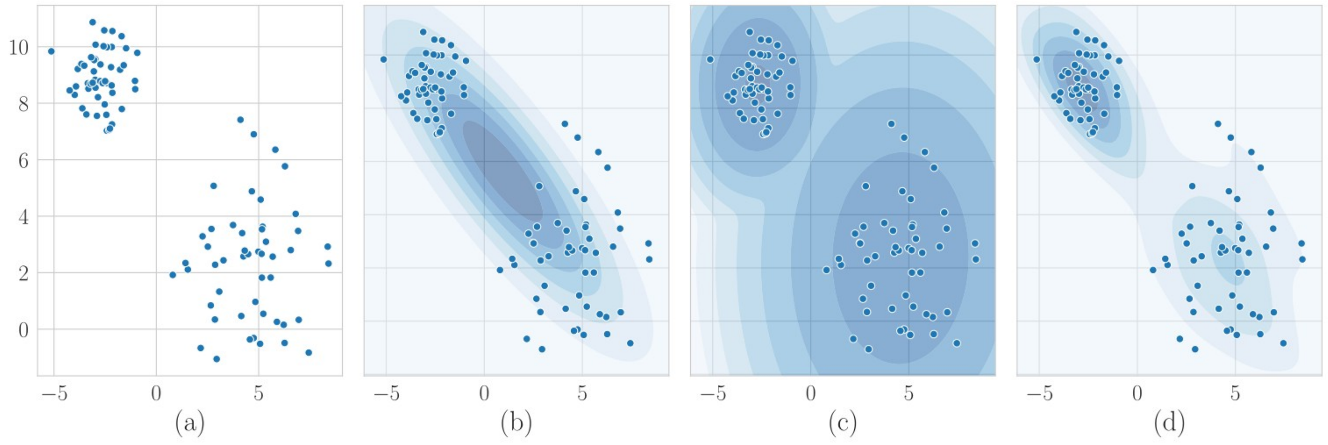


Figure 2.3.: Examples of PDF mechanisms fitted to a mock dataset. Legend: (a) Original dataset, (b) Gaussian generative model, (c) Gaussian Mixture Model and (d) Gaussian Kernel Density Estimation.

Chapter 2 – Literature review

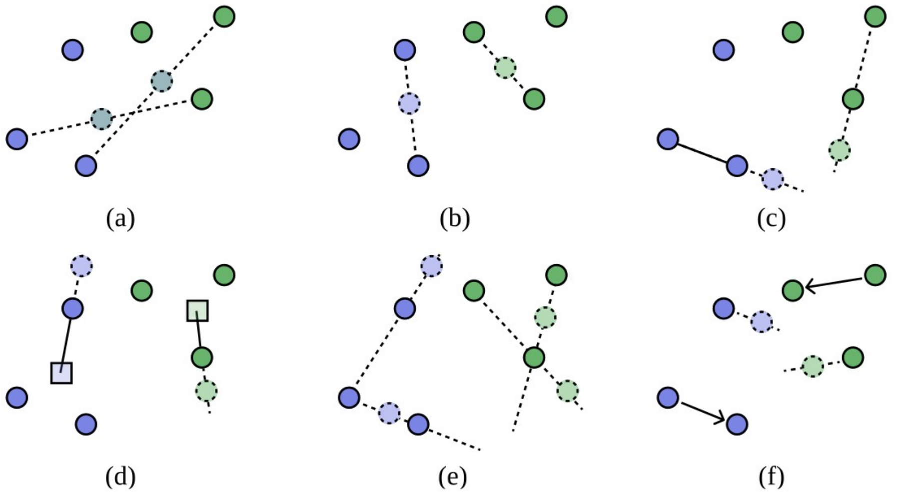


Figure 2.4.: Examples of linear transformation mechanisms. Legend: (a) Between-class interpolation, (b) Within-class interpolation, (c) Observation-based extrapolation, (d) Hard extrapolation, (e) Combination of interpolation and extrapolation and (f) Difference transform.

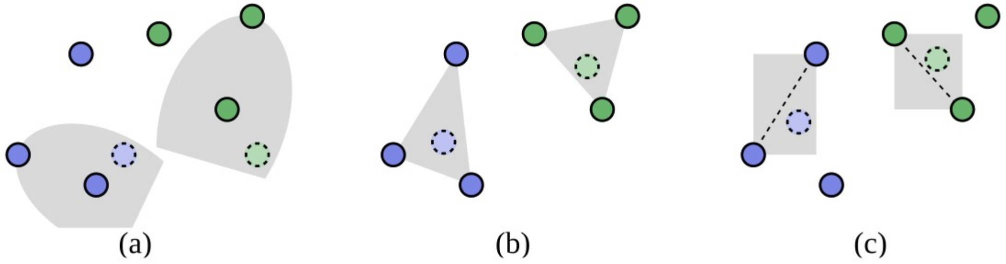


Figure 2.5.: Examples of geometric transformation mechanisms. Legend: (a) hypersphere mechanism, (b) triangular mechanism and (c) hyperrectangle mechanism.

Chapter 3 – G-SMOTENC

Table 3.1.: Description of the datasets collected after data preprocessing. The sampling strategy is similar across datasets. Legend: (IR) Imbalance Ratio

Dataset	Metric	Non-Metric	Obs.	Min. Obs.	Maj. Obs.	IR	Classes
Abalone	1	7	4139	15	689	45.93	18
Adult	8	6	5000	1268	3732	2.94	2
Adult (10)	8	6	5000	451	4549	10.09	2
Annealing	4	6	790	34	608	17.88	4
Census	24	7	5000	337	4663	13.84	2
Contraceptive	4	5	1473	333	629	1.89	3
Contraceptive (10)	4	5	1036	62	629	10.15	3
Contraceptive (20)	4	5	990	31	629	20.29	3
Contraceptive (31)	4	5	973	20	629	31.45	3
Contraceptive (41)	4	5	966	15	629	41.93	3
Covertype	2	10	5000	20	2449	122.45	7
Credit Approval	9	6	653	296	357	1.21	2
German Credit	13	7	1000	300	700	2.33	2
German Credit (10)	13	7	770	70	700	10.00	2
German Credit (20)	13	7	735	35	700	20.00	2
German Credit (30)	13	7	723	23	700	30.43	2
German Credit (41)	13	7	717	17	700	41.18	2
Heart Disease	5	5	740	22	357	16.23	5
Heart Disease (21)	5	5	735	17	357	21.00	5

Chapter 3 – G-SMOTENC

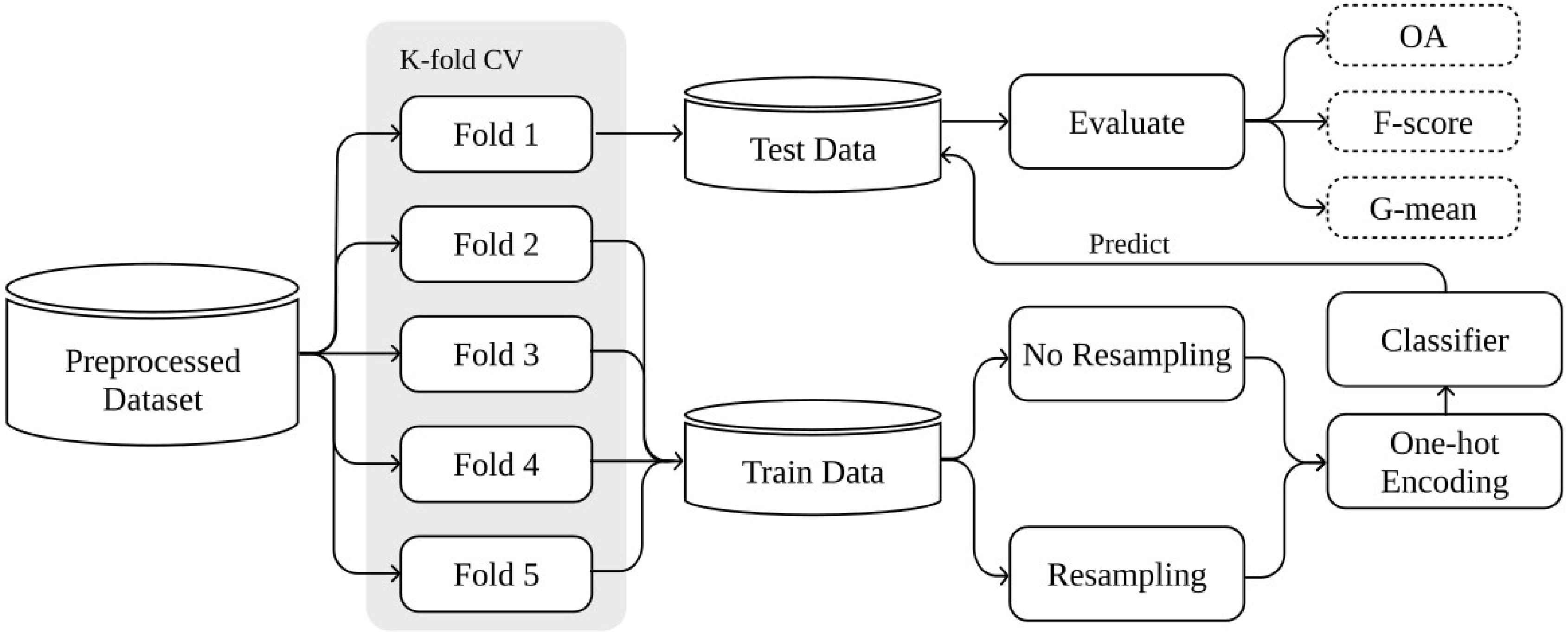


Figure 3.2.: Experimental procedure used in this study.

Chapter 3 – G-SMOTENC

Classifier	Metric	G-SMOTENC	NONE	SMOTENC	ROS	RUS	SMOTE-ENC
DT	OA	1.66 ± 0.13	**1.61 ± 0.27**	3.58 ± 0.20	4.68 ± 0.15	5.42 ± 0.27	4.05 ± 0.23
DT	F-Score	**1.32 ± 0.11**	3.84 ± 0.40	3.13 ± 0.20	4.32 ± 0.19	5.47 ± 0.23	2.92 ± 0.34
DT	G-Mean	**1.68 ± 0.24**	5.84 ± 0.09	2.82 ± 0.21	2.95 ± 0.32	4.26 ± 0.32	3.45 ± 0.30
KNN	OA	2.50 ± 0.17	**1.37 ± 0.28**	4.21 ± 0.25	3.34 ± 0.35	5.68 ± 0.22	3.89 ± 0.15
KNN	F-Score	**1.37 ± 0.16**	3.95 ± 0.35	3.11 ± 0.29	3.47 ± 0.36	5.53 ± 0.23	3.58 ± 0.23
KNN	G-Mean	**1.74 ± 0.17**	5.84 ± 0.12	2.89 ± 0.23	3.76 ± 0.33	3.00 ± 0.45	3.76 ± 0.23
LR	OA	2.74 ± 0.19	**1.37 ± 0.28**	3.08 ± 0.21	4.34 ± 0.30	5.74 ± 0.17	3.74 ± 0.28
LR	F-Score	**2.11 ± 0.24**	4.53 ± 0.35	2.37 ± 0.28	3.47 ± 0.32	5.21 ± 0.27	3.32 ± 0.38
LR	G-Mean	2.13 ± 0.26	6.00 ± 0.00	3.61 ± 0.21	**2.11 ± 0.23**	3.32 ± 0.40	3.84 ± 0.28
RF	OA	1.82 ± 0.11	**1.24 ± 0.09**	3.97 ± 0.16	4.32 ± 0.21	5.92 ± 0.06	3.74 ± 0.22
RF	F-Score	**1.32 ± 0.13**	5.05 ± 0.31	3.16 ± 0.22	3.05 ± 0.31	5.37 ± 0.14	3.05 ± 0.27
RF	G-Mean	**1.68 ± 0.22**	5.79 ± 0.21	3.26 ± 0.28	2.47 ± 0.30	3.89 ± 0.35	3.89 ± 0.19

Mean rankings over the different datasets (20), folds (5) and runs (3) used in the experiment.

Chapter 3 – G-SMOTENC

Table 3.4.: Mean scores over the different datasets, folds and runs used in the experiment

Classifier	Metric	G-SMOTENC	NONE	SMOTENC	ROS	RUS	SMOTE-ENC
DT	OA	0.74 ± 0.05	$\mathbf{0.75 \pm 0.04}$	0.68 ± 0.04	0.66 ± 0.04	0.58 ± 0.04	0.65 ± 0.04
DT	F-Score	$\mathbf{0.56 \pm 0.04}$	0.52 ± 0.04	0.54 ± 0.04	0.52 ± 0.04	0.48 ± 0.04	0.51 ± 0.04
DT	G-Mean	$\mathbf{0.69 \pm 0.03}$	0.60 ± 0.02	0.68 ± 0.03	0.67 ± 0.03	0.65 ± 0.03	0.66 ± 0.03
KNN	OA	0.69 ± 0.04	$\mathbf{0.73 \pm 0.05}$	0.67 ± 0.04	0.69 ± 0.05	0.57 ± 0.04	0.68 ± 0.05
KNN	F-Score	$\mathbf{0.53 \pm 0.04}$	0.50 ± 0.04	0.52 ± 0.04	0.52 ± 0.04	0.46 ± 0.04	0.51 ± 0.04
KNN	G-Mean	$\mathbf{0.66 \pm 0.03}$	0.58 ± 0.03	0.64 ± 0.03	0.62 ± 0.03	0.65 ± 0.03	0.63 ± 0.03
LR	OA	0.68 ± 0.05	$\mathbf{0.75 \pm 0.04}$	0.68 ± 0.05	0.66 ± 0.05	0.58 ± 0.04	0.67 ± 0.04
LR	F-Score	$\mathbf{0.54 \pm 0.04}$	0.52 ± 0.04	$\mathbf{0.54 \pm 0.04}$	0.53 ± 0.04	0.48 ± 0.04	0.52 ± 0.04
LR	G-Mean	$\mathbf{0.69 \pm 0.02}$	0.60 ± 0.03	0.68 ± 0.02	$\mathbf{0.69 \pm 0.03}$	0.67 ± 0.03	0.67 ± 0.03
RF	OA	0.74 ± 0.04	$\mathbf{0.76 \pm 0.04}$	0.69 ± 0.04	0.69 ± 0.04	0.59 ± 0.04	0.68 ± 0.05
RF	F-Score	$\mathbf{0.57 \pm 0.04}$	0.48 ± 0.04	0.55 ± 0.04	0.55 ± 0.04	0.49 ± 0.04	0.53 ± 0.04
RF	G-Mean	$\mathbf{0.70 \pm 0.02}$	0.57 ± 0.02	0.68 ± 0.03	0.69 ± 0.03	0.68 ± 0.03	0.68 ± 0.02

Chapter 3 – G-SMOTENC

Table 3.5.: Results for the Friedman test. Statistical significance is tested at a level of $\alpha = 0.05$. The null hypothesis is that there is no difference in the classification outcome across resamplers.

Classifier	Metric	p-value	Significance
DT	F-Score	2.2e-10	True
DT	G-Mean	1.2e-10	True
KNN	F-Score	2.3e-09	True
KNN	G-Mean	9.4e-10	True
LR	F-Score	2.1e-07	True
LR	G-Mean	9.7e-11	True
RF	F-Score	8.5e-12	True
RF	G-Mean	2.0e-10	True

Table 3.6.: Adjusted p-values using the Holm-Bonferroni test. Statistical significance is tested at a level of $\alpha = 0.05$. The null hypothesis is that the benchmark methods perform similarly to the control method (G-SMOTENC).

Classifier	Metric	NONE	SMOTENC	ROS	RUS	SMOTE-ENC
DT	F-Score	**1.5e-04**	**1.5e-04**	**7.3e-06**	**1.2e-06**	1.0e-01
DT	G-Mean	**5.6e-07**	**2.7e-03**	**2.8e-02**	**3.9e-04**	**2.3e-02**
KNN	F-Score	**6.4e-04**	**2.2e-04**	**7.2e-04**	**6.4e-04**	**5.9e-06**
KNN	G-Mean	**1.6e-05**	**9.6e-03**	**6.5e-03**	2.0e-01	**3.5e-03**
LR	F-Score	**4.0e-03**	6.1e-01	**9.2e-03**	**3.6e-04**	5.6e-02
LR	G-Mean	**1.6e-07**	**4.0e-04**	8.6e-01	2.4e-01	**4.7e-03**
RF	F-Score	**1.7e-06**	**2.4e-04**	**8.0e-03**	**1.7e-06**	**8.0e-03**
RF	G-Mean	**3.8e-06**	**8.8e-03**	2.5e-01	**2.3e-02**	**1.7e-03**

Chapter 3 – G-SMOTENC

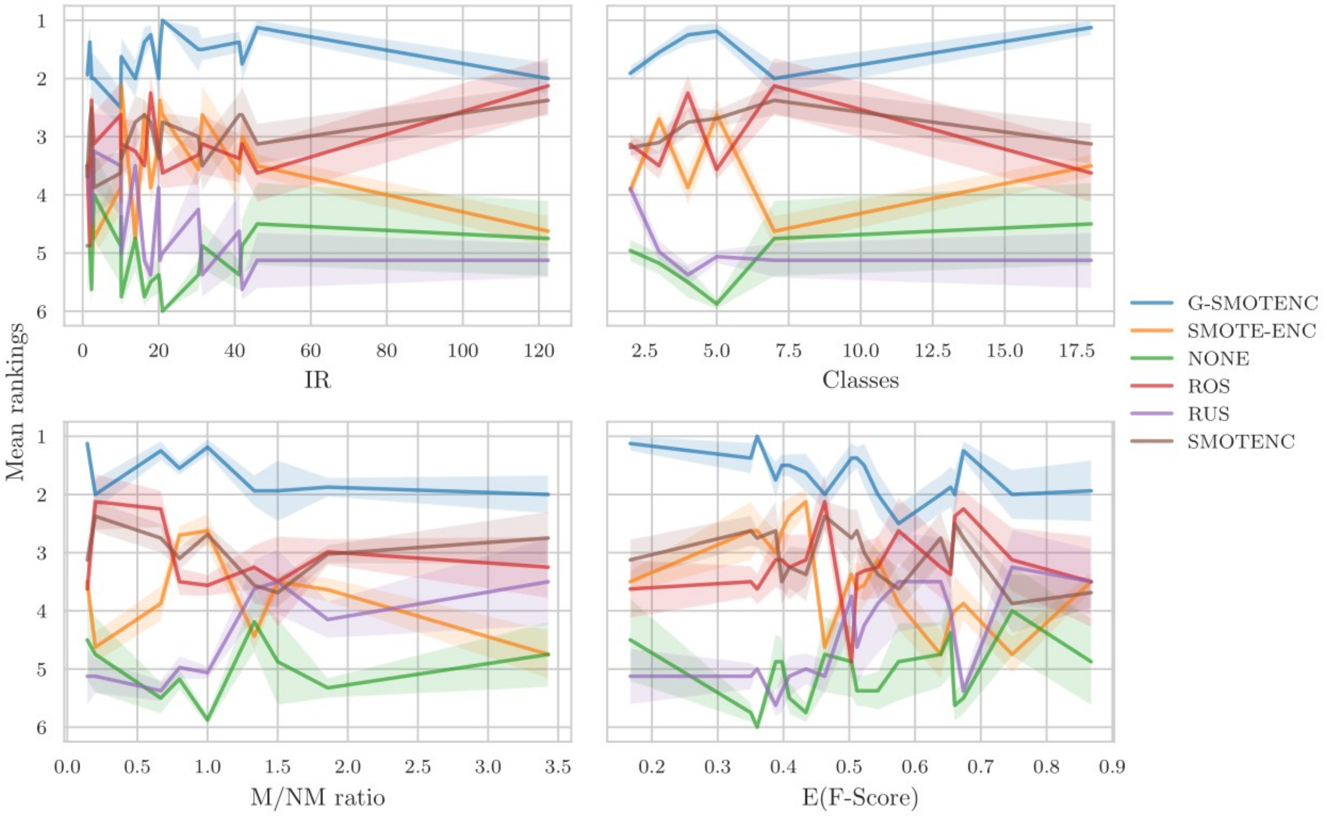


Figure 3.3.: Average ranking of oversamplers over different characteristics of the datasets used in the experiment. Legend: IR — Imbalance Ratio, Classes — Number of classes in the dataset, M/NM ratio — ratio between the number of metric and non-metric features, E(F-Score) — Mean F-Score of dataset across all combinations of classifiers and oversamplers.

Chapter 4 – K-Means SMOTE (LULC)

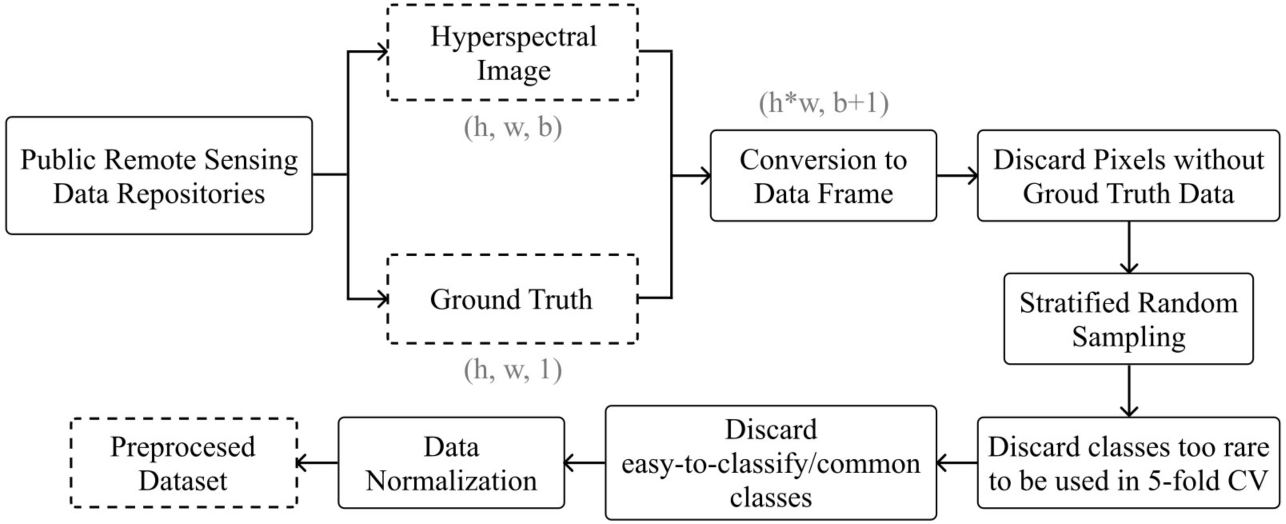


Figure 4.4.: Data collection and preprocessing pipeline.

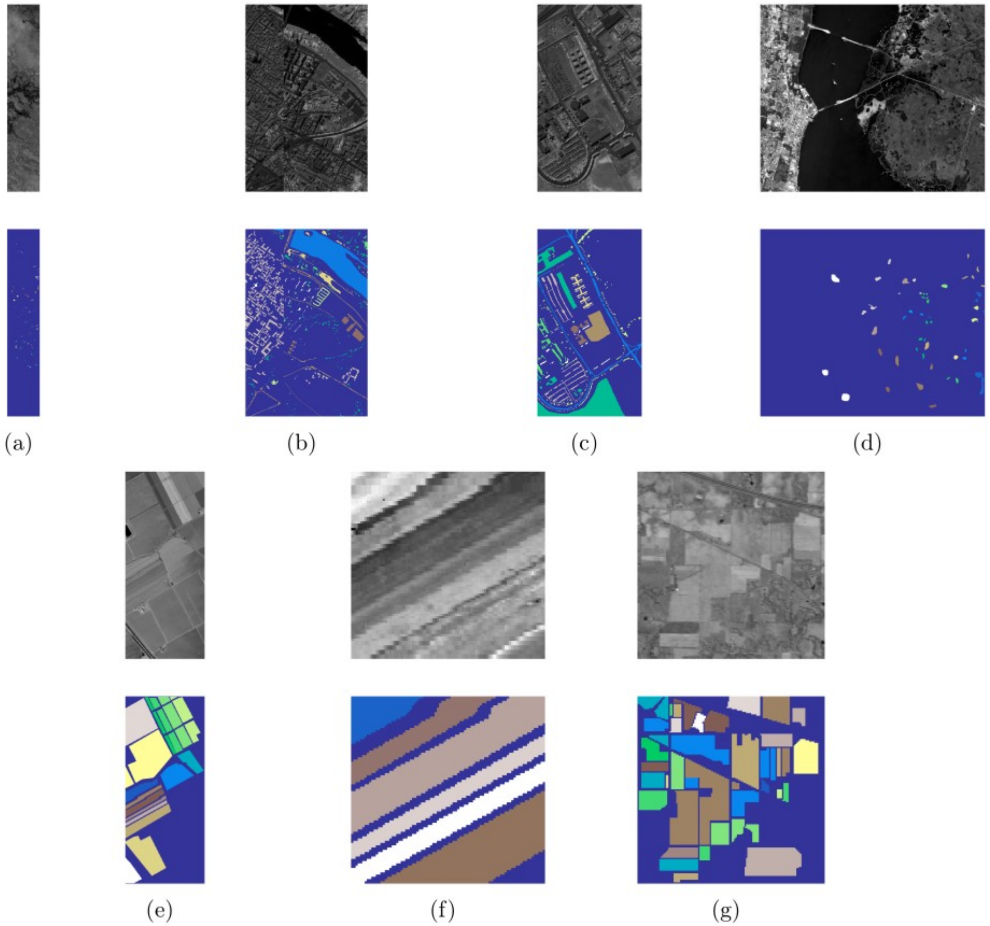


Figure 4.5.: Gray scale visualization of a band (top row) and ground truth (bottom row) of each scene used in this study. (a) Botswana, (b) Pavia Center, (c) Pavia University, (d) Kennedy Space Center, (e) Salinas, (f) Salinas A, (g) Indian Pines.

Chapter 4 – K-Means SMOTE (LULC)

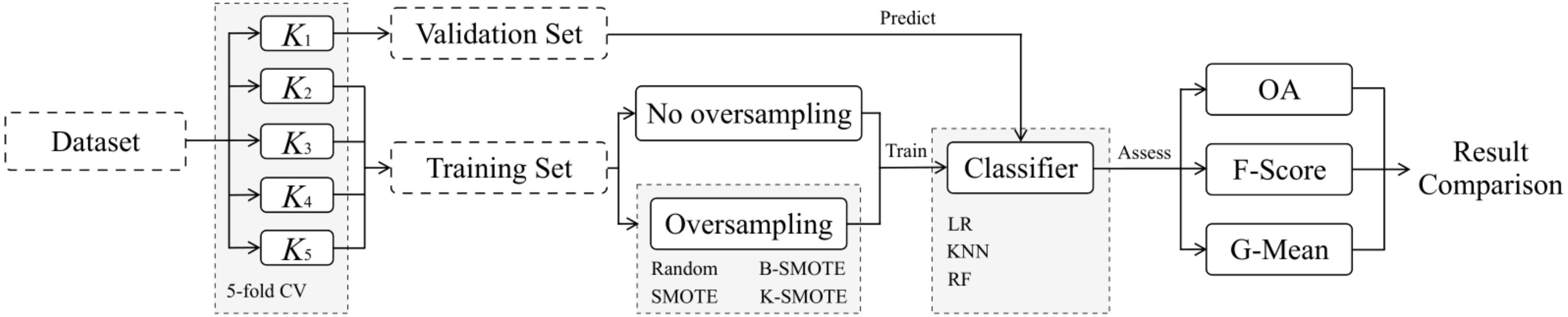
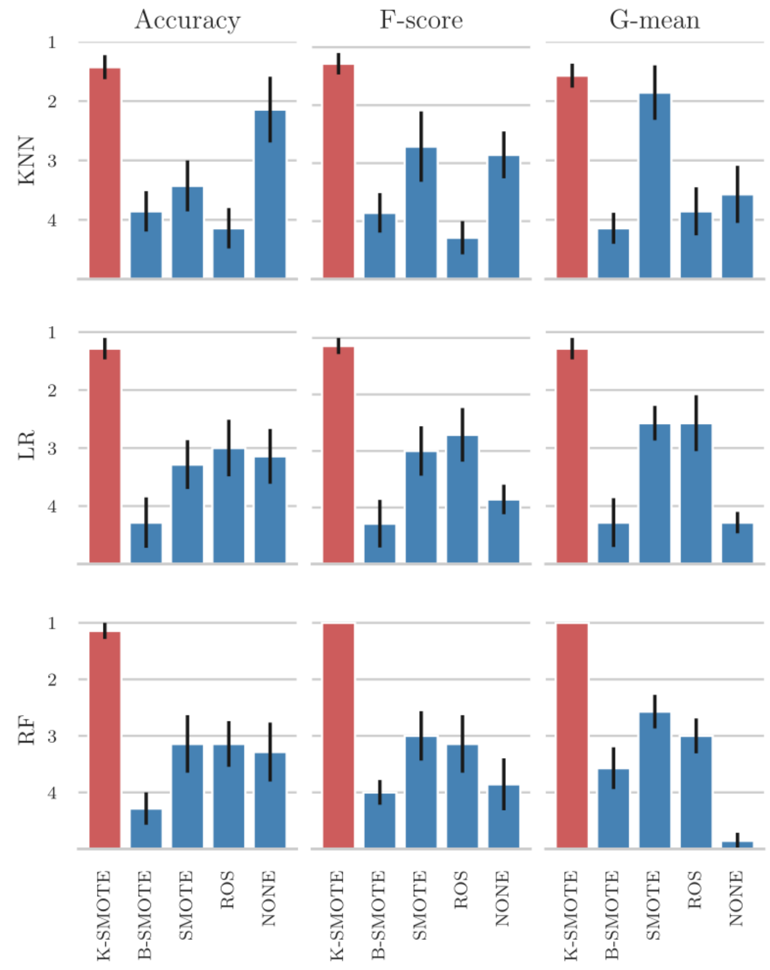


Figure 4.6.: Experimental procedure. The performance metrics are averaged over the 5 folds across each of the 3 different initializations of this procedure for a given combination of oversampler, classifier and hyperparameter definition.

Chapter 4 – K-Means SMOTE (LULC)

Dataset	Features	Instances	Min. Instances	Maj. Instances	IR	Classes
Botswana	145	288	20	41	2.05	11
Pavia Centre	102	3898	278	879	3.16	7
Kennedy Space Center	176	497	23	80	3.48	11
Salinas A	224	535	37	166	4.49	6
Pavia University	103	2392	89	679	7.63	8
Salinas	224	4236	91	719	7.9	15
Indian Pines	220	984	21	236	11.24	11

Description of the datasets used for this experiment



Mean ranking of oversamplers across datasets.

Chapter 4 – K-Means SMOTE (LULC)

Classifier	Metric	p-value	Significance
LR	Accuracy	9.8e-03	True
LR	F-score	2.3e-03	True
LR	G-mean	9.8e-04	True
KNN	Accuracy	4.3e-03	True
KNN	F-score	4.3e-03	True
KNN	G-mean	3.0e-03	True
RF	Accuracy	5.5e-03	True
RF	F-score	2.9e-03	True
RF	G-mean	1.8e-04	True

Table 4.4.: Results for Friedman test. Statistical significance is tested at a level of $\alpha = 0.05$. The null hypothesis is that there is no difference in the classification outcome across oversamplers.

Dataset	NONE	ROS	SMOTE	B-SMOTE
Botswana	**3.1e-02**	**3.9e-03**	**3.9e-03**	**3.9e-03**
Pavia Centre	**3.1e-02**	**3.9e-03**	**1.2e-02**	**3.9e-03**
Kennedy Space Center	**3.1e-02**	**3.9e-03**	**2.7e-02**	**3.9e-03**
Salinas A	**3.1e-02**	**3.9e-03**	**1.2e-02**	**3.9e-03**
Pavia University	**3.1e-02**	**3.9e-03**	**3.9e-03**	**3.9e-03**
Salinas	**3.1e-02**	5.5e-02	**2.7e-02**	**3.9e-03**
Indian Pines	**3.1e-02**	**3.9e-03**	**7.8e-03**	**3.9e-03**

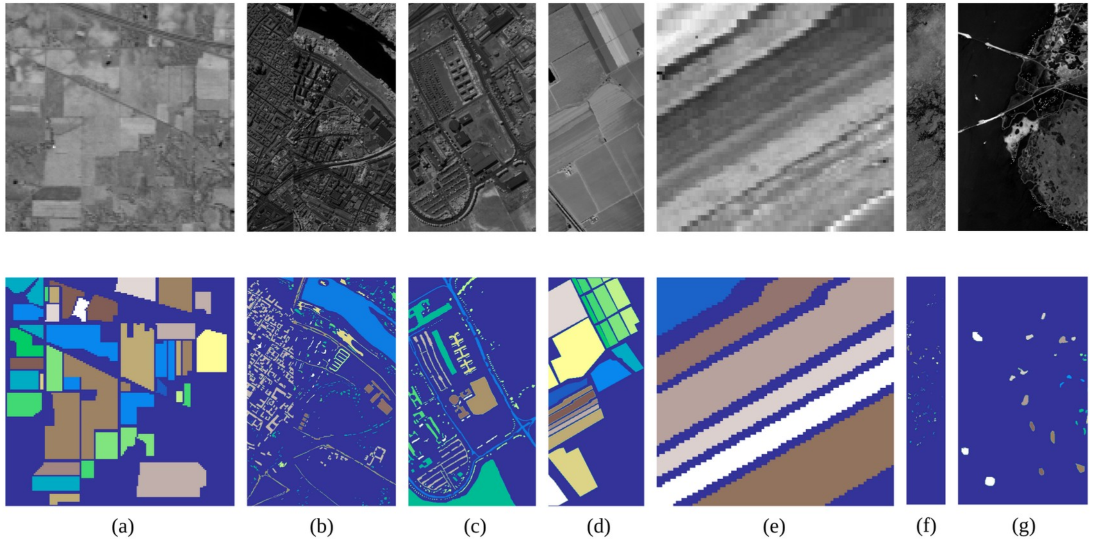
Table 4.5.: *p-values* of the Wilcoxon signed-rank test. Boldface values are statistically significant at a significance level of $\alpha = 0.05$.

Chapter 5 – Active Learning (LULC)

Dataset	Sensor	Location	Dimension	Bands	Res. (m)	Classes
Botswana	Hyperion	Okavango Delta	1476 x 256	145	30	14
Salinas A	AVIRIS	California, USA	86 x 83	224	3.7	6
Kennedy Space Center	AVIRIS	Florida, USA	512 x 614	176	18	16
Indian Pines	AVIRIS	NW Indiana, USA	145 x 145	220	20	16
Salinas	AVIRIS	California, USA	512 x 217	224	3.7	16
Pavia University	ROSIS	Pavia, Italy	610 x 610	103	1.3	9
Pavia Centre	ROSIS	Pavia, Italy	1096 x 1096	102	1.3	9

Table 5.1.: Description of the hyperspectral scenes used in this experiment. The column "Res. (m)" refers to the resolution of the sensors (in meters) that captured each of the scenes.

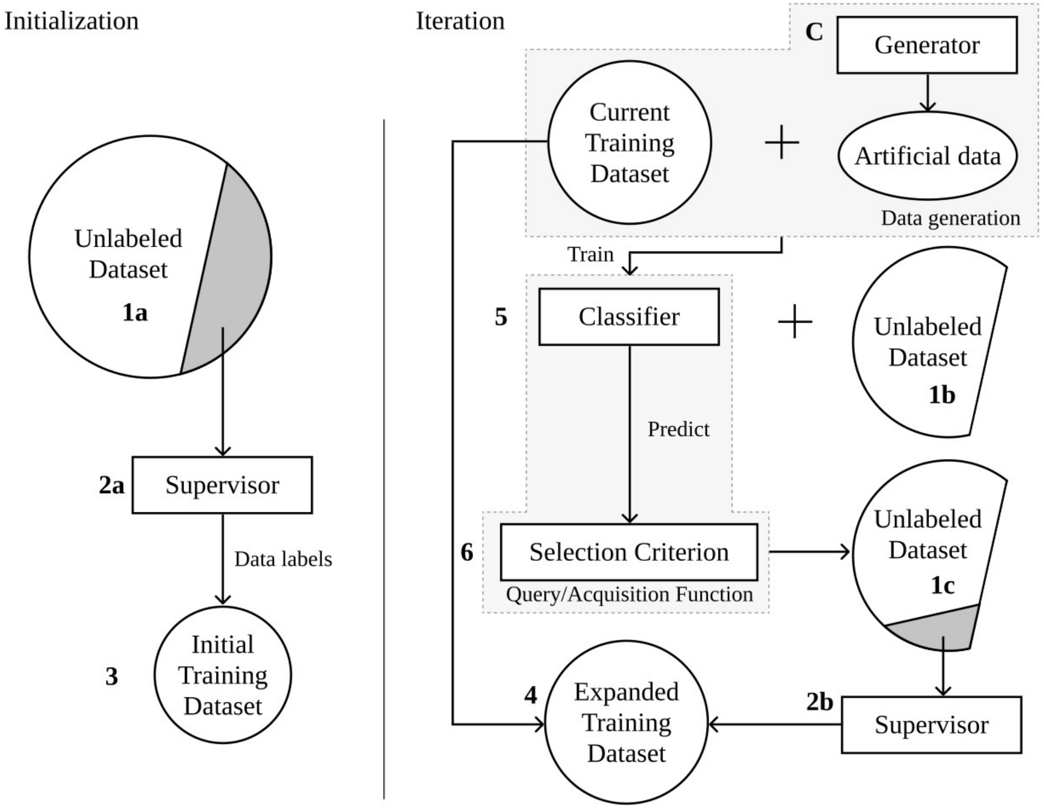
Chapter 5 – Active Learning (LULC)



Gray scale visualization of a band and ground truth of each scene used in this study.

Dataset	Features	Instances	Min. Instances	Maj. Instances	IR	Classes
Botswana	145	1500	89	154	1.73	12
Salinas A	224	1500	109	428	3.93	6
Kennedy Space Center	176	1500	47	272	5.79	12
Indian Pines	220	1500	31	366	11.81	12
Salinas	224	1500	25	312	12.48	16
Pavia University	103	1500	33	654	19.82	9
Pavia Centre	102	1500	27	668	24.74	9

Description of the datasets collected from each corresponding scene.



Proposed method

Chapter 5 – Active Learning (LULC)

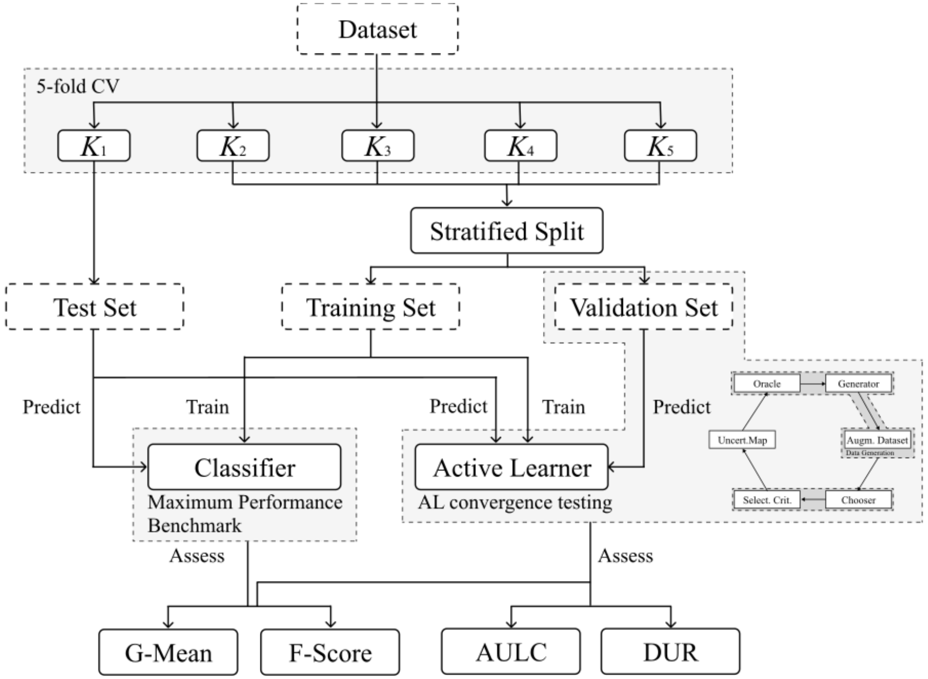


Figure 5.5.: Experimental procedure. The datasets extracted from hyperspectral scenes are split in 5 folds. 1 of those (*e.g.*, K_1) is used to test the optimal performance of AL algorithms and the classification without AL. The training set is used to iterate AL algorithms and train classifiers. The validation set is used to test the convergence of AL algorithms. The results are averaged over the 5 folds across each of the 3 different initializations of this procedure.

Chapter 5 – Active Learning (LULC)

Classifier	Evaluation Metric	MP	Standard	Proposed
KNN	F-score	0.838 ± 0.106	0.835 ± 0.115	**0.843 ± 0.105**
KNN	G-mean	0.907 ± 0.063	0.904 ± 0.069	**0.912 ± 0.061**
LR	F-score	**0.890 ± 0.084**	0.883 ± 0.096	0.887 ± 0.097
LR	G-mean	0.935 ± 0.052	0.931 ± 0.059	**0.938 ± 0.055**
RF	F-score	0.859 ± 0.083	0.866 ± 0.081	**0.869 ± 0.08**
RF	G-mean	0.918 ± 0.051	0.921 ± 0.051	**0.930 ± 0.043**

Table 5.7.: Optimal classification scores. The Maximum Performance (MP) classification scores are calculated using classifiers trained using the entire training set.

Dataset	p-value	Significance
Botswana	3.8e-03	True
Indian Pines	2.3e-04	True
Kennedy Space Center	1.3e-04	True
Pavia Centre	4.3e-03	True
Pavia University	4.6e-05	True
Salinas	4.6e-05	True
Salinas A	3.0e-03	True

Table 5.8.: Adjusted p-values using the Wilcoxon signed-rank method. Bold values are statistically significant at a level of $\alpha = 0.05$. The null hypothesis is that the performance of the proposed framework is similar to that of the original framework.

Chapter 6 – Active Learning (Synth. Data)

Table 6.2.: Description of the datasets collected after data preprocessing. The sampling strategy is similar across datasets. Legend: (IR) Imbalance Ratio

Dataset	Features	Instances	Minority instances	Majority instances	IR	Classes
Image Segmentation	14	1155	165	165	1.0	7
Mfeat Zernike	47	1994	198	200	1.01	10
Texture	40	1824	165	166	1.01	11
Waveform	40	1666	551	564	1.02	3
Pendigits	16	1832	176	191	1.09	10
Vehicle	18	846	199	218	1.1	4
Mice Protein	69	1073	105	150	1.43	8
Gas Drift	128	1987	234	430	1.84	6
Japanese Vowels	12	1992	156	323	2.07	9
Usps	256	1859	142	310	2.18	10
Gesture Segmentation	32	1974	200	590	2.95	5
Volkert	147	1943	45	427	9.49	10
Steel Plates	24	1941	55	673	12.24	7
Baseball	15	1320	57	1196	20.98	3
Wine Quality	11	1599	10	681	68.1	6

Chapter 6 – Active Learning (Synth. Data)

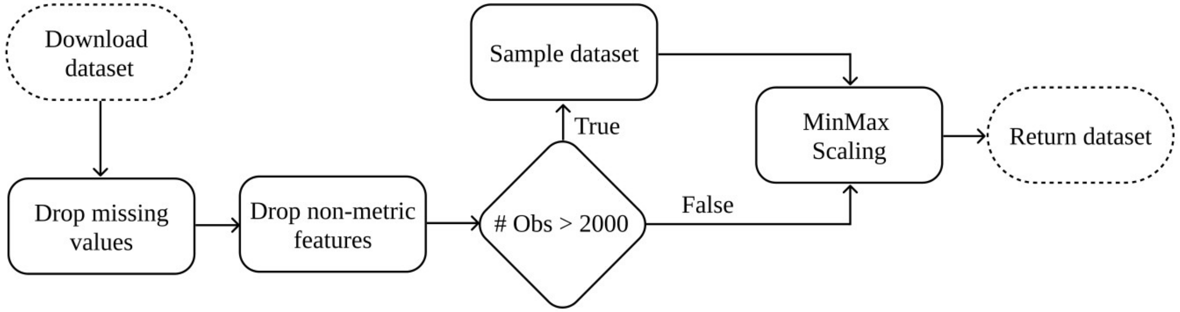


Figure 6.4.: Data preprocessing pipeline.

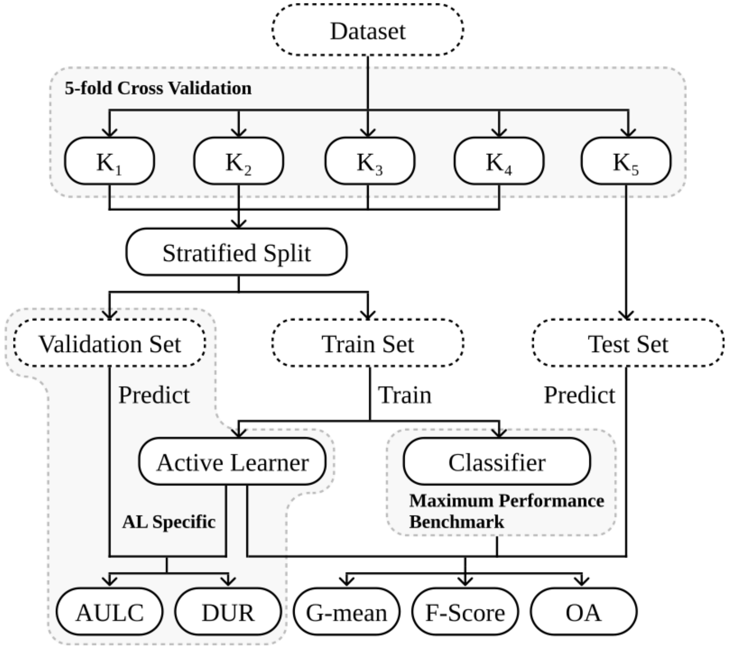


Figure 6.5.: Experimental procedure flowchart. The preprocessed datasets are split into five folds. One of the folds is used to test the best-found classifiers using AL and the classifiers trained using the entire training dataset (containing the remaining folds). The training set is used to run both the AL simulations as well as train the normal classifiers. The validation set is used to measure AL-specific performance metrics over each iteration. We use different subsets for overall classification performance and AL-specific performance to avoid data leakage.

Chapter 6 – Active Learning (Synth. Data)

Classifier	Evaluation Metric	Standard	Oversampling	Proposed
DT	Accuracy	2.13 ± 0.96	2.40 ± 0.49	**1.47 ± 0.62**
DT	F-score	2.47 ± 0.81	2.20 ± 0.40	**1.33 ± 0.70**
DT	G-mean	2.73 ± 0.57	1.93 ± 0.44	**1.33 ± 0.70**
KNN	Accuracy	2.07 ± 0.93	2.07 ± 0.68	**1.87 ± 0.81**
KNN	F-score	2.47 ± 0.81	1.87 ± 0.50	**1.67 ± 0.87**
KNN	G-mean	2.87 ± 0.34	**1.47 ± 0.50**	1.67 ± 0.70
LR	Accuracy	2.13 ± 0.88	2.20 ± 0.65	**1.67 ± 0.79**
LR	F-score	2.80 ± 0.40	1.87 ± 0.50	**1.33 ± 0.70**
LR	G-mean	2.80 ± 0.40	1.80 ± 0.54	**1.40 ± 0.71**
RF	Accuracy	2.27 ± 0.85	**1.87 ± 0.50**	**1.87 ± 0.96**
RF	F-score	2.73 ± 0.57	1.80 ± 0.54	**1.47 ± 0.72**
RF	G-mean	2.87 ± 0.34	**1.53 ± 0.50**	1.60 ± 0.71

Mean rankings of the AULC metric over the different datasets (15), folds (5), and runs (3) used in the experiment.

Chapter 6 – Active Learning (Synth. Data)

Table 6.7.: Optimal classification scores. The Maximum Performance (MP) classification scores are calculated using classifiers trained using the entire training set.

Classifier	Evaluation Metric	MP	Standard	Oversampling	Proposed
DT	Accuracy	**0.732 ± 0.155**	0.726 ± 0.157	0.721 ± 0.167	0.727 ± 0.168
DT	F-score	0.682 ± 0.194	0.679 ± 0.193	0.679 ± 0.197	**0.684 ± 0.200**
DT	G-mean	0.792 ± 0.138	0.791 ± 0.136	0.797 ± 0.134	**0.800 ± 0.137**
KNN	Accuracy	**0.801 ± 0.164**	0.799 ± 0.168	0.784 ± 0.183	0.789 ± 0.183
KNN	F-score	0.742 ± 0.224	0.744 ± 0.223	0.741 ± 0.223	**0.746 ± 0.224**
KNN	G-mean	0.827 ± 0.160	0.829 ± 0.158	0.839 ± 0.146	**0.840 ± 0.147**
LR	Accuracy	0.778 ± 0.157	**0.791 ± 0.158**	0.764 ± 0.184	0.773 ± 0.185
LR	F-score	0.693 ± 0.243	0.717 ± 0.241	0.718 ± 0.222	**0.727 ± 0.226**
LR	G-mean	0.796 ± 0.171	0.814 ± 0.165	0.839 ± 0.130	**0.842 ± 0.137**
RF	Accuracy	0.827 ± 0.145	**0.832 ± 0.148**	0.827 ± 0.154	0.829 ± 0.153
RF	F-score	0.767 ± 0.215	0.775 ± 0.216	0.781 ± 0.204	**0.784 ± 0.204**
RF	G-mean	0.844 ± 0.148	0.849 ± 0.149	0.863 ± 0.131	**0.865 ± 0.131**

Chapter 6 – Active Learning (Synth. Data)

Table 6.8.: Friedman test results. Statistical significance is tested at a level of $\alpha = 0.05$. The null hypothesis is that there is no difference in the classification outcome across oversamplers.

Classifier	Evaluation Metric	p-value	Significance
DT	Accuracy	1.1e-15	True
DT	F-score	2.4e-31	True
DT	G-mean	2.3e-23	True
KNN	Accuracy	5.9e-20	True
KNN	F-score	8.8e-69	True
KNN	G-mean	8.8e-52	True
LR	Accuracy	1.1e-30	True
LR	F-score	4.0e-98	True
LR	G-mean	2.3e-83	True
RF	Accuracy	2.8e-26	True
RF	F-score	1.8e-88	True
RF	G-mean	1.8e-61	True

Table 6.9.: Adjusted p-values using the Wilcoxon signed-rank method. Bold values are statistically significant at a level of $\alpha = 0.05$. The null hypothesis is that the performance of the proposed framework is similar to that of the oversampling or standard framework.

Dataset	Oversampling	Standard
Baseball	5.0e-01	3.4e-01
Gas Drift	**3.7e-26**	**4.6e-57**
Gesture Segmentation	**1.3e-02**	**8.7e-04**
Image Segmentation	**9.6e-18**	**2.1e-44**
Japanese Vowels	**2.4e-09**	**1.6e-32**
Mfeat Zernike	**1.2e-12**	**9.5e-40**
Mice Protein	**6.5e-32**	**1.5e-61**
Pendigits	**5.0e-18**	**2.3e-45**
Steel Plates	**3.4e-04**	**1.3e-08**
Texture	**1.5e-22**	**6.7e-57**
Usps	3.8e-01	**2.1e-29**
Vehicle	**7.4e-11**	**7.9e-13**
Volkert	2.5e-01	**1.3e-02**
Waveform	**8.9e-08**	**2.6e-02**
Wine Quality	**3.8e-05**	**6.1e-03**

Chapter 6 – Active Learning (Synth. Data)

Table 6.10.: Adjusted p-values using the Holm-Bonferroni method. Bold values are statistically significant at a level of $\alpha = 0.05$. The null hypothesis is that the Oversampling or Proposed method does not perform better than the control method (Standard AL framework).

Classifier	Evaluation Metric	Oversampling	Proposed
DT	Accuracy	7.7e-01	**1.1e-04**
DT	F-score	6.3e-02	**2.0e-06**
DT	G-mean	**1.0e-08**	**2.9e-12**
KNN	Accuracy	**1.0e-02**	8.5e-01
KNN	F-score	**7.1e-07**	**8.3e-13**
KNN	G-mean	**1.9e-11**	**1.0e-12**
LR	Accuracy	**3.2e-02**	8.3e-01
LR	F-score	**1.5e-09**	**5.8e-17**
LR	G-mean	**1.9e-13**	**5.6e-16**
RF	Accuracy	4.3e-01	4.3e-01
RF	F-score	**1.4e-11**	**1.1e-12**
RF	G-mean	**1.5e-10**	**1.2e-10**