# New Developments for Addressing Class Imbalance Issue in Classification Tasks

Ph.D. Thesis Defense

Surani Matharaarachchi

November, 08 2024

University of Manitoba

# Outline

# Introduction

- The rapid advancement of science and technology has resulted in increasingly complex datasets
- Predictive Modeling
- Make data-driven decisions
- Challenges in Predictive Modeling: Class Imbalance Issue
  - Abnormal instances
  - Curse of dimensionality

# Introduction

- The rapid advancement of science and technology has resulted in increasingly complex datasets
- Predictive Modeling
- Make data-driven decisions
- Challenges in Predictive Modeling: Class Imbalance Issue
  - Abnormal instances
  - Curse of dimensionality

# Class Imbalance Issue

- Occurs when the number of instances in different classes is significantly disproportionate.
- Examples:
  - Fraud Detection
  - Spam Detection
  - Medical Diagnosis
  - Churn Prediction
- Issues:
  - Leads to biased models
  - Decreases predictive accuracy
- Abnormal Instances

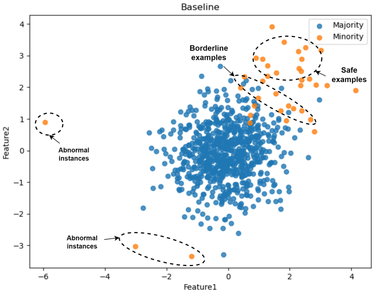



Figure: Class imbalance with outliers in minority class

# Class Imbalance Issue

- Occurs when the number of instances in different classes is significantly disproportionate.
- Examples:
  - Fraud Detection
  - Spam Detection
  - Medical Diagnosis
  - Churn Prediction
- Issues:
  - Leads to biased models
  - Decreases predictive accuracy
- Abnormal Instances

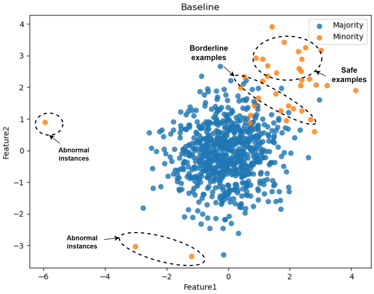



Figure: Class imbalance with outliers in minority class

# Class Imbalance Issue

- Occurs when the number of instances in different classes is significantly disproportionate.
- Examples:
  - Fraud Detection
  - Spam Detection
  - Medical Diagnosis
  - Churn Prediction
- Issues:
  - Leads to biased models
  - Decreases predictive accuracy
- Abnormal Instances

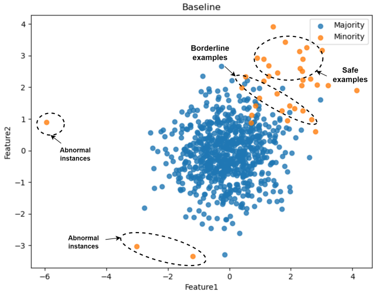



Figure: Class imbalance with outliers in minority class

# Class Imbalance Issue

- Occurs when the number of instances in different classes is significantly disproportionate.
- Examples:
  - Fraud Detection
  - Spam Detection
  - Medical Diagnosis
  - Churn Prediction
- Issues:
  - Leads to biased models
  - Decreases predictive accuracy
- Abnormal Instances

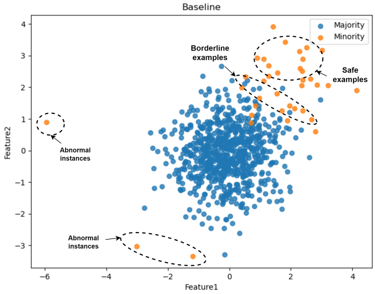



Figure: Class imbalance with outliers in minority class

# Objectives

- Develop novel techniques for addressing class imbalance in classification tasks.
  - To investigate the impact of outliers within the minority class using popular existing methods.
  - To propose innovative strategies capable of mitigating the adverse effects of outliers on class imbalance data.
  - To further extend our approaches to address high-dimensionality issue.
  - To offer empirical evidence, supported by simulated and experimental results, that demonstrates the effectiveness of these proposed solutions in enhancing classification performance.

# Thesis Contribution

- Challenges of Imbalanced Data: Identifying Long COVID Patients

1. Discovering Long COVID Symptom Patterns: Association Rule Mining and Sentiment Analysis in Social Media Tweets (Published) [5]
2. Long COVID Prediction in Manitoba Using Clinical Notes Data: A Machine Learning Approach (In Review) [6]

- Advancements for Imbalance Data Classification

3. Enhancing SMOTE for Imbalanced Data with Abnormal Minority Instances Just Published!!! [9]
4. Deep-ExtSMOTE: Integrating Autoencoders for Advanced Mitigation of Class Imbalance in High-Dimensional Data Classification (In Review) [4]

## Third Manuscript

Enhancing SMOTE for Imbalanced Data with Abnormal Minority Instances [9]

Machine Learning with Applications 18 (2024) 100597

Contents lists available at ScienceDirect

Machine Learning with Applications

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

### Enhancing SMOTE for imbalanced data with abnormal minority instances

Surani Matharaarachchi [a,*], Mike Domaratzki [b], Saman Muthukumarana [a]

[a] Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[b] Department of Computer Science, Western University, London, ON, N6A 5B7, Canada

ARTICLE INFO

Dataset link: http://archive.ics.uci.edu/ml

*Keywords:*
Class imbalance
Abnormal instances
Inverse distance
SMOTE
Bayesian Gaussian mixture models
Dirichlet distribution

ABSTRACT

Imbalanced datasets are frequent in machine learning, where certain classes are markedly underrepresented compared to others. This imbalance often results in sub-optimal model performance, as classifiers tend to favour the majority class. A significant challenge arises when abnormal instances, such as outliers, exist within the minority class, diminishing the effectiveness of traditional re-sampling methods like the Synthetic Minority Over-sampling Technique (SMOTE). This manuscript addresses this critical issue by introducing four SMOTE extensions: Distance ExtSMOTE, Dirichlet ExtSMOTE, FCRP SMOTE, and BGMM SMOTE. These methods leverage a weighted average of neighbouring instances to enhance the quality of synthetic samples and mitigate the impact of outliers. Comprehensive experiments conducted on diverse simulated and real-world imbalanced datasets demonstrate that the proposed methods improve classification performance compared to the original SMOTE and its most competitive variants. Notably, we demonstrate that Dirichlet ExtSMOTE outperforms most other proposed and existing SMOTE variants in terms of achieving better F1 score, MCC, and PR-AUC. Our results underscore the effectiveness of these advanced SMOTE extensions in tackling class imbalance, particularly in the presence of abnormal instances, offering robust solutions for real-world applications.

# Synthetic Minority Oversampling Technique (SMOTE)

- Resampling
- Balancing the Dataset:
  - Create new samples for the minority class.
- Technique:
  - Interpolate between randomly chosen minority class samples and their nearest neighbors.
  - $p_{new} = p_0 + \alpha(p_3 - p_0)$

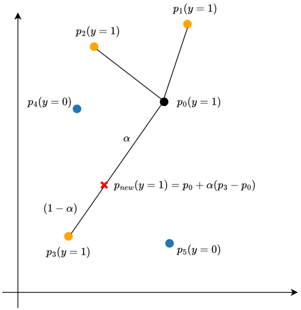



Figure: SMOTE data generation

# Limitation with SMOTE

- Challenged by outliers within the minority class.

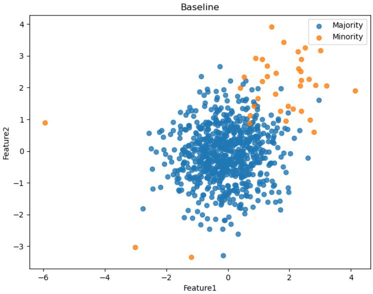



Figure: Original Data

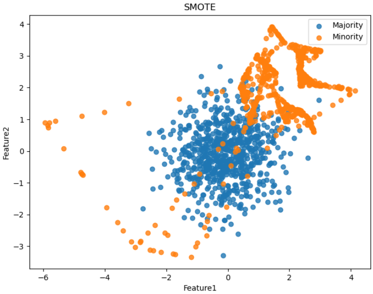



Figure: Re-sampled data with SMOTE

# Proposed Solution

- Technique:
  - Use a weighted average of neighbouring instances.
  - $p_{new} = \frac{\sum_{j=1}^{k}(w_j \times p_j)}{\sum_{j=1}^{k} w_j}, j = 1, \ldots, k$
  - Improve robustness against outliers and noisy data.
  - Learn from a more extensive set of nearest neighbours.
- Challenge:
  - Selecting suitable weights to enhance resilience to outliers and noisy data.

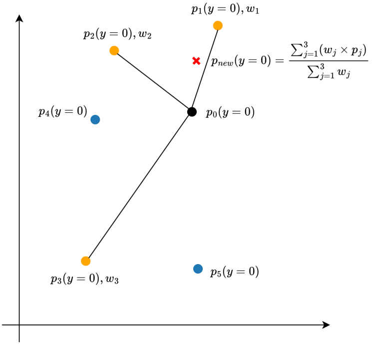



Figure: Proposed method data generation

# Proposed Solution

- Technique:
  - Use a weighted average of neighbouring instances.
  - $p_{new} = \frac{\sum_{j=1}^{k}(w_j \times p_j)}{\sum_{j=1}^{k} w_j}, j = 1, \ldots, k$
  - Improve robustness against outliers and noisy data.
  - Learn from a more extensive set of nearest neighbours.
- Challenge:
  - Selecting suitable weights to enhance resilience to outliers and noisy data.

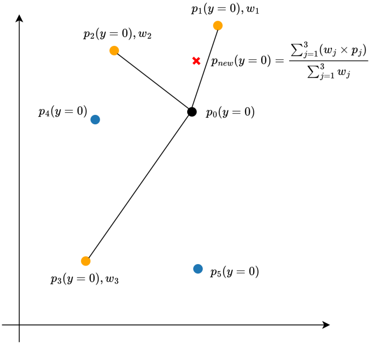



Figure: Proposed method data generation

# How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

# How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

# How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

# How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

# How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

# 1. Distance extSMOTE

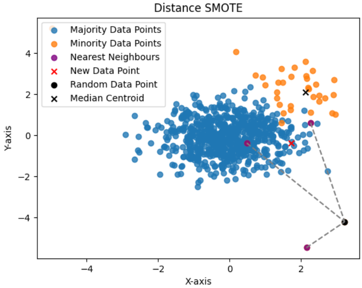
- $d_j \in \mathbb{R}$ is the Euclidean distance between the median centroid of the minority class and the nearest neighbours
- $w_j = d_{j,norm}^{-1}$ = Normalized inverse distance

**Algorithm** Distance ExtSMOTE

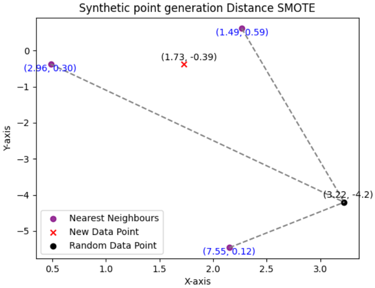
**Require:** $X \in \mathbb{R}^{n \times p}$ the features, $Y \in \{0,1\}^n$ the binary class label outputs.
**Require:** $k \in \mathbb{N}$ the number of neighbors to select for the $k$-Nearest Neighbors.
**Ensure:** Generated data $X_{new} \in \mathbb{R}^{q \times p}$ and $Y_{new} \in \{0,1\}^q$ with $q$ points created.

```
Denote by S_1 the number of points labelled as the minority class and S_0 the number of points labelled as the majority class.
Initialize X_new and Y_new as empty vectors.
Obtain the median centroid (μ) of the minority class.
while S_1 < S_0 do
    Filter D = {X_i | Y_i = 1}, the set of points labeled as minority class 1.
    Randomly choose r ∈ D and find the indices of its k nearest neighbors, r_1, ..., r_k.
    Consider the inverse distances, from μ, to each nearest neighbour as weights, w_j = d_j^{-1}
    x^new ← Σ(w_j × x_{r_j}) / Σ w_j  for all j from 1 to k.
    y^new ← 1
    S_1 = S_1 + 1
    Append x^new to X_new, append y^new to Y_new
end while
return X_new, Y_new
```

# 1. Distance extSMOTE




(a) This scenario occurs when an outlier is chosen as a neighbouring point.




(b) The values within parentheses indicate $(d_j, w_j)$.

Figure: An example of creating a sample - Distance extSMOTE

# 2. Dirichlet extSMOTE

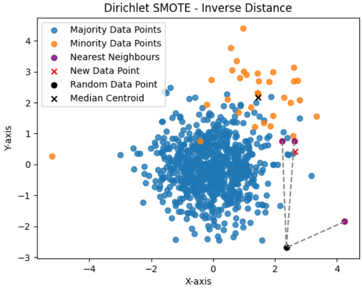
- The pdf of the Dirichlet distribution for a point $\boldsymbol{p}$ on the simplex:

$$w_j = P(\boldsymbol{p}|\boldsymbol{\alpha}) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_K) \stackrel{\text{def}}{=} \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \quad (1)$$

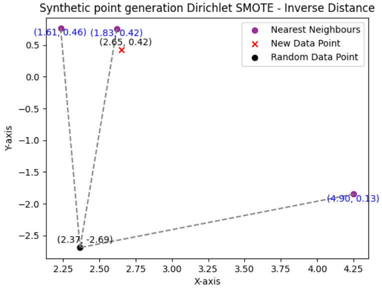
**Algorithm** Dirichlet ExtSMOTE

```
if Type is 'Inverse distance (D)' then
    Calculate the distances, D = [d_1, ..., d_k] from μ to each nearest neighbour and obtain the reciprocal of each distance D^{-1} = [1/d_1, ..., 1/d_k].
  Then α = D^{-1} × m
else if Type is 'Uniform Vector (UV)' then
    Generate a vector α = 1_k × m, where 1_k = [1, ..., 1]
else if Type is 'Uniform Distribution (UD)' then
    Generate vector U of size k from uniform(0, 1) distribution, then α = U × m.
end if
Use α as parameters to the Dirichlet Distribution and generate random weights w_j ~ Dir(α)
x^new ← Σ w_j x_{r_j} for all j from 1 to k, as Σ w_j = 1
y^new ← 1
S_1 = S_1 + 1
Append x^new to X_new, append y^new to Y_new
return X_new, Y_new
```

# 2. Dirichlet extSMOTE (Inverse Distance)



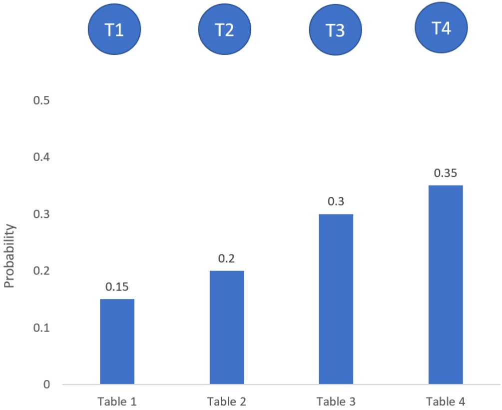

(a) This scenario occurs when an outlier is chosen as a neighbouring point.




(b) The values within parentheses indicate $(d_j, w_j)$.

Figure: An example of creating a sample - Dirichlet extSMOTE

# 3. FCRP SMOTE




- Showcasing the weight selection of FCRP SMOTE using the Chinese restaurant process concept with finite number of tables with a parameter value $\alpha = 0.1$

# 3. FCRP SMOTE

## **Algorithm** FCRP SMOTE

**Require:** $X \in \mathbb{R}^{n \times p}$ the features, $Y \in \{0,1\}^n$ the binary class label outputs.
**Require:** $k \in \mathbb{N}$ the number of neighbors to select for the $k$-Nearest Neighbors.
**Require:** $\alpha \in \mathbb{R}$, scalar parameter to update preferences.
**Ensure:** Generated data $X_{new} \in \mathbb{R}^{q \times p}$ and $Y_{new} \in \{0,1\}^q$ with $q$ the number of points created.

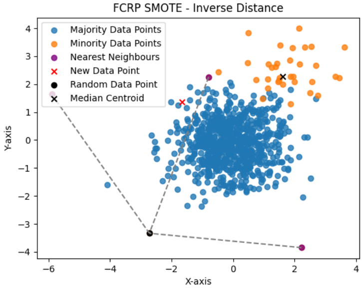
Denote by $S_1$ the number of points labelled as the minority class and $S_0$ the number of points labelled as the majority class.
Initialize $X_{new}$ and $Y_{new}$ as empty vectors.
Filter $\mathcal{D} = X_i | Y_i = 1$, the set of points labeled as minority class 1 and obtain the median centroid $(c_m)$ of the minority cluster.
**while** $S1 < S_0$ **do**
    Randomly choose $r \in \mathcal{D}$ and find the indices of its $k$ nearest neighbors, $\{r_1, \ldots, r_k\}$.
    Consider the normalized inverse distances, from $c_m$, to each nearest neighbour as initial preferences, $P = \boldsymbol{D}_{norm}^{-1}$ and choose first nearest neighbour with probability $p_i$, $i$ from $1, \ldots, k$.
    **for** N-1 **do**
        Choose the next nearest neighbour with the following updated probabilities $q_i$,

$$q_i = \begin{cases} \frac{p_i + \alpha}{1+\alpha}, & \text{for previously chosen neighbour} \\ \frac{p_i}{1+\alpha}, & \text{for other neighbours} \end{cases}$$

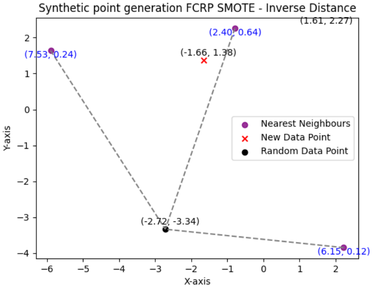
        $p_i = q_i$
    **end for**
    Obtain the final preferences for each nerighbour $p_i$ as the weights $w_j$.
    $x^{new} \leftarrow \sum (w_j \times x_{r_j})$ for all $j$ from 1 to $k$ and $y^{new} \leftarrow 1$
    $S_1 = S_1 + 1$
    Append $x^{new}$ to $X_{new}$, append $y^{new}$ to $Y_{new}$
**end while**
return $X_{new}, Y_{new}$

# 3. FCRP SMOTE

- Initial preferences $= d_{norm}^{-1}$
- $w_j$ = Final allocation probabilities




(a) This scenario occurs when an outlier is chosen as a neighbouring point.




(b) The values within parentheses indicate $(d_j, w_j)$.

Figure: An example of creating a sample - FCRP SMOTE

# 4. BGMM SMOTE

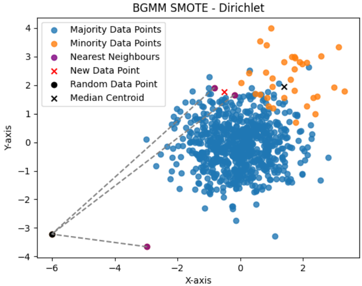
Bayesian Gaussian Mixture Models (BGMM)

- A probabilistic model used for clustering
- Cluster Assignment

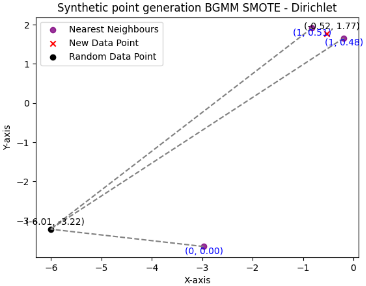
1. Expectation Maximization:
   - Expectation (E-step): For each data point, the model calculates the probability of the point belonging to each cluster
   - Maximization (M-step): Update the parameters of the model by maximizing the expected log-likelihood
2. Cluster Assignment: Probabilistically assigns data points to clusters based on the calculated probabilities.
3. Soft Assignments: This does not definitively allocate a point to a single cluster.

# 4. BGMM SMOTE

- $c_j$ = Cluster assignment of the $j^{th}$ nearest neighbour
- $w_j$ = Normalized cluster probability of the cluster which the median centroid belongs




(a) This scenario occurs when an outlier is chosen as a neighboring point.




(b) The values within parentheses indicate $(c_j, w_j)$.

Figure: An example of creating a sample - BGMM SMOTE (D)

# Synthetic Data Generation

- $\boldsymbol{X}_{minority-outliers} \sim \mathcal{N}_{2\times 2}(\boldsymbol{\mu}^{(1)}_{2\times 1}, \boldsymbol{\Sigma}^{(1)}_{2\times 2})$
- $\boldsymbol{X}_{majority} \sim \mathcal{N}_{2\times 2}(\boldsymbol{\mu}^{(2)}_{2\times 1}, \boldsymbol{\Sigma}^{(2)}_{2\times 2})$
- $\boldsymbol{X}_{outliers} \sim \mathrm{Uniform}([-10, 10]^2)$

$$\boldsymbol{\mu}^{(1)}_{2\times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}_{2\times 1}, \boldsymbol{\Sigma}^{(1)}_{2\times 2} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\mu}^{(2)}_{2\times 1} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times 1}, \boldsymbol{\Sigma}^{(2)}_{2\times 2} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}_{2\times 2}$$

# Synthetic Data Generation

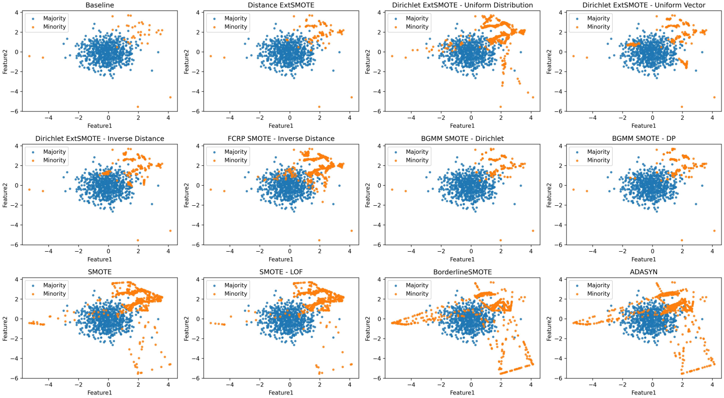



Figure: Comparison of resampled data

# Simulation Results (Noisy Moons)

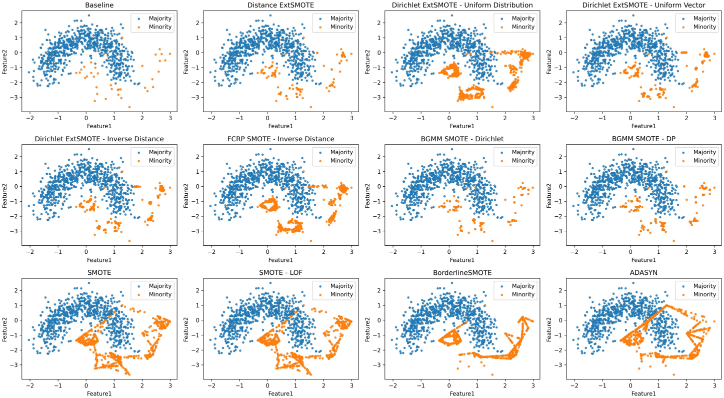



Figure: Comparison of resampled data

# Simulation Results (Noisy Circles)

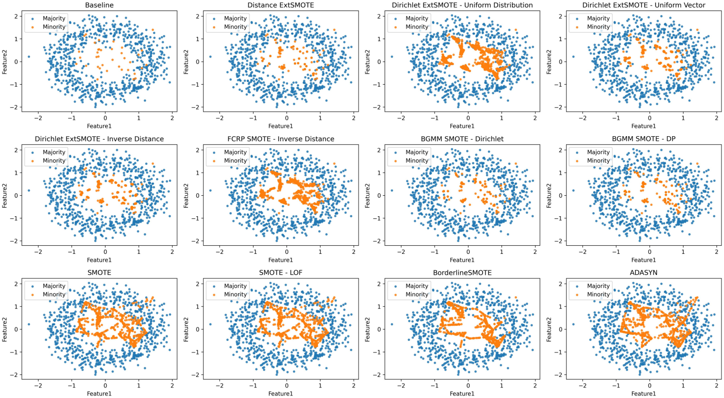



Figure: Comparison of resampled data

# Synthetic Data Generation

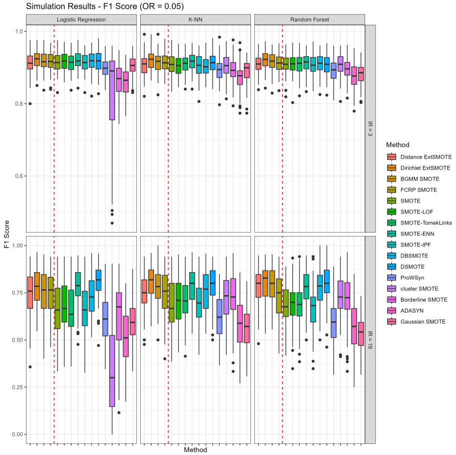



Figure: F1 Scores for 100 simulated datasets with 5-fold cross validation

# Application Data

Table: Characteristics of the binary class datasets used in the computational study.

| No | Dataset | Instances | Features | Minority class | Majority class | %Minority | %Majority | IR | Presence of LOF Outliers |
|---|---|---|---|---|---|---|---|---|---|
| 1 | yeast6 | 1484 | 8 | EXC | Remaining classes | 2.36 | 97.64 | 41.40 | Yes |
| 2 | yeast5 | 1484 | 8 | EXC, ERL | Remaining classes | 2.70 | 97.30 | 36.10 | Yes |
| 3 | yeast-1289vs7 | 947 | 8 | VAC | NUC, CYT, ERL, POX | 3.17 | 96.83 | 30.57 | Yes |
| 4 | yeast4 | 1484 | 8 | ME2 | Remaining classes | 3.44 | 96.56 | 28.10 | Yes |
| 5 | yeast-2vs8 | 483 | 8 | POX | CYT | 4.14 | 95.86 | 23.15 | Yes |
| 6 | glass12357vs6 | 214 | 9 | 6 | Remaining classes | 4.21 | 95.79 | 22.78 | Yes |
| 7 | yeast-1458vs7 | 693 | 8 | VAC | NUC, ME3, ME2, POX | 4.33 | 95.67 | 22.10 | Yes |
| 8 | oil | 937 | 49 | minority | majority | 4.38 | 95.62 | 21.85 | No |
| 9 | abalone9_18 | 731 | 7 | 9, 18 | Remaining classes | 5.75 | 94.25 | 16.40 | Yes |
| 10 | glass12367vs5 | 214 | 9 | 5 | Remaining classes | 6.07 | 93.93 | 15.46 | Yes |
| 11 | thyroid_sick | 3772 | 52 | sick | healthy | 6.12 | 93.88 | 15.33 | Yes |
| 12 | yeast-1vs7 | 459 | 8 | VAC | NUC | 6.54 | 93.46 | 14.30 | Yes |
| 13 | us_crime | 1994 | 100 | >0.65 | <=0.65 | 7.52 | 92.48 | 12.29 | Yes |
| 14 | glass12vs5 | 159 | 9 | 5 | 1, 2 | 8.18 | 91.82 | 11.23 | Yes |
| 15 | spectrometer | 531 | 93 | >=44 | <44 | 8.47 | 91.53 | 10.80 | Yes |
| 16 | landsat_satellite | 6435 | 36 | 2 | Remaining classes | 9.73 | 90.27 | 9.28 | Yes |
| 17 | mfeatmor0 | 2000 | 6 | 0, 1 | Remaining classes | 10.00 | 90.00 | 9.00 | Yes |
| 18 | yeast3 | 1484 | 8 | ME3 | Remaining classes | 10.98 | 89.02 | 8.10 | Yes |
| 19 | mfeatmor01 | 2000 | 6 | 0 | Remaining classes | 20.00 | 80.00 | 4.00 | Yes |
| 20 | glass123vs567 | 214 | 9 | 5, 6, 7 | Remaining classes | 23.83 | 76.17 | 3.20 | Yes |
| 21 | parkinsons | 195 | 22 | 1 | 0 | 24.62 | 75.38 | 3.06 | Yes |
| 22 | habermans_survival | 306 | 3 | 2 | 1 | 26.47 | 73.53 | 2.78 | Yes |
| 23 | glass23567vs1 | 214 | 9 | 1 | Remaining classes | 32.71 | 67.29 | 2.06 | Yes |
| 24 | breast_cancer | 569 | 30 | M | B | 37.26 | 62.74 | 1.68 | Yes |
| 25 | banknote | 1372 | 4 | 1 | Remaining classes | 44.46 | 55.54 | 1.25 | Yes |

# Application Results

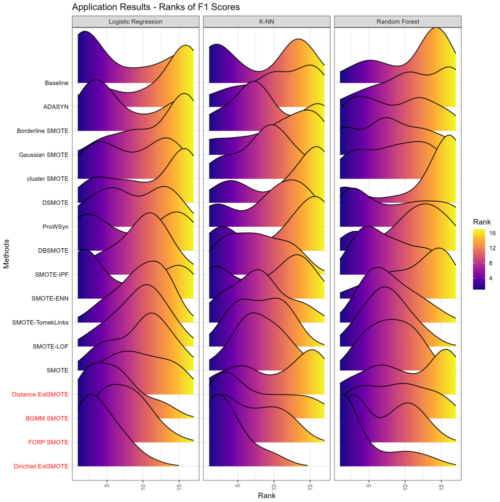



Figure: F1 Score Ranks for the datasets with 100 $\times$ 5-fold cross validation

## Fourth Manuscript

Deep-ExtSMOTE: Integrating Autoencoders for Advanced Mitigation of Class Imbalance in High-Dimensional Data Classification [4]

# High-Dimensional Data

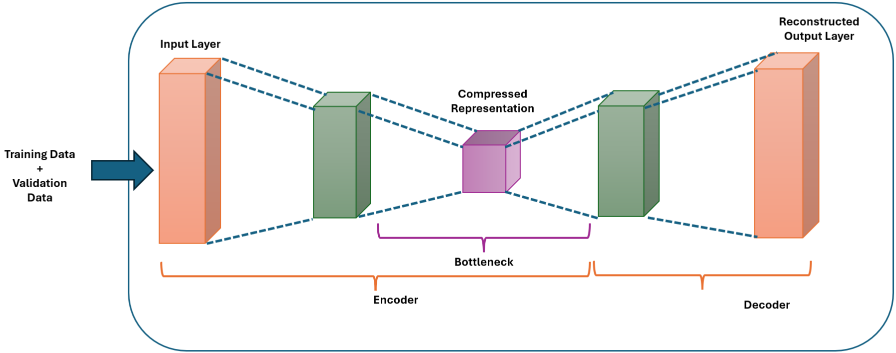
- Curse of Dimensionality
  - A large number of features relative to the available data, “large p, small n” problem [3].
  - Challenges:
    - Data Sparsity
    - Increased Model Complexity and Overfitting
    - Computational Challenges
- Feature Reduction
  - A critical strategy to address the challenges of high dimensionality in class imbalance [2, 7, 8].

# High-Dimensional Data

- Curse of Dimensionality
  - A large number of features relative to the available data, “large p, small n” problem [3].
  - Challenges:
    - Data Sparsity
    - Increased Model Complexity and Overfitting
    - Computational Challenges
- Feature Reduction
  - A critical strategy to address the challenges of high dimensionality in class imbalance [2, 7, 8].

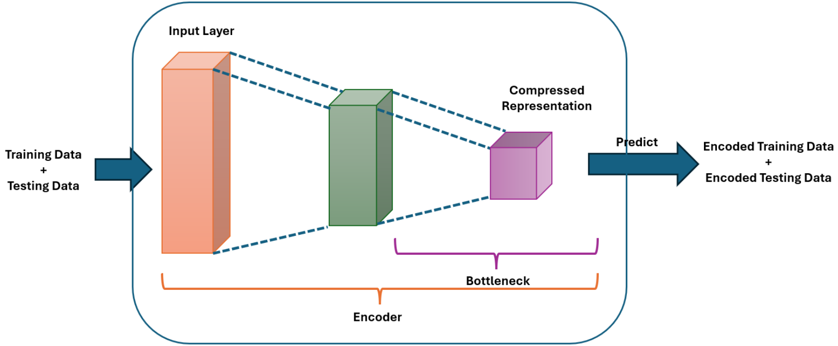
# 5. Deep-ExtSMOTE

- Autoencoder + Dirichlet ExtSMOTE
- Step 1: Train the Autoencoder



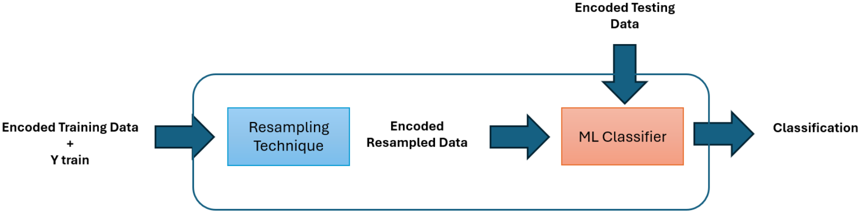
# 5. Deep-ExtSMOTE

- Step 2: Extract Encoded Representation



# 5. Deep-ExtSMOTE

- Step 3: Resampling and Classification



# Simulation Results

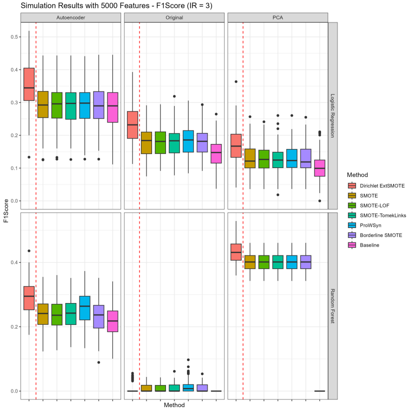



Figure: F1-Score distribution for 100 trials using simulated datasets with 1000 samples and 5000 features (2000 informative), with an imbalance ratio (IR) of 3.

# Application Results

- Application 1: Isolet (Continuous Binary Classification)
  - Dataset includes 617 continuous features, representing processed characteristics of the audio signals.
  - Scenario 1: Original Isolet Dataset
    - Dataset includes 7797 samples, resulting in a feature-to-sample ratio of approximately 0.0791.
  - Scenario 2: Reduced Isolet Dataset
    - Selected a subset of 1000 samples from the original 7797 samples. This adjustment resulted in a feature-to-sample ratio of 0.617.

# Application Results

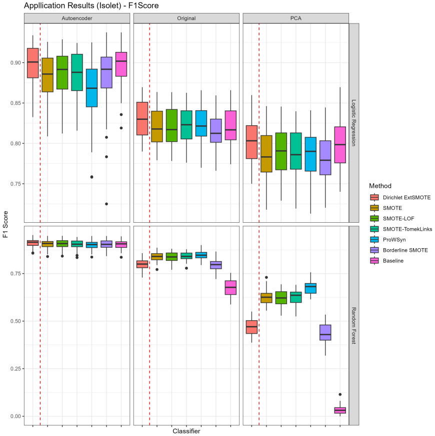



Figure: F1 Scores for the Isolet dataset across 50 training and test splits.

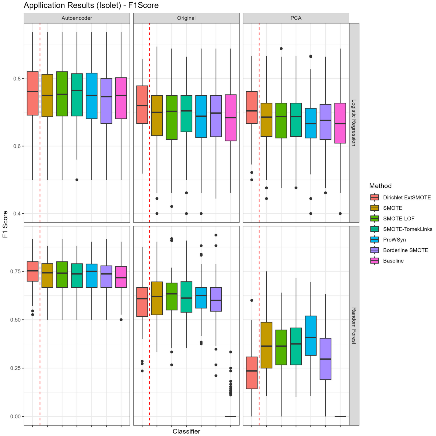



Figure: F1 Scores for the reduced Isolet dataset across 50 training and test splits.

# Application Results

- Application 2: Chile (Categorical Binary Classification)
  - Predict the yield of 204 Chile pepper genotypes from multi-environment trials in New Mexico, USA.
  - Conduct experiment by starting with 2,500 features and increasing the number of features to 7,500.
  - Feature-to-sample ratio ranging from approximately 12.25 to 37.7.

# Application Results

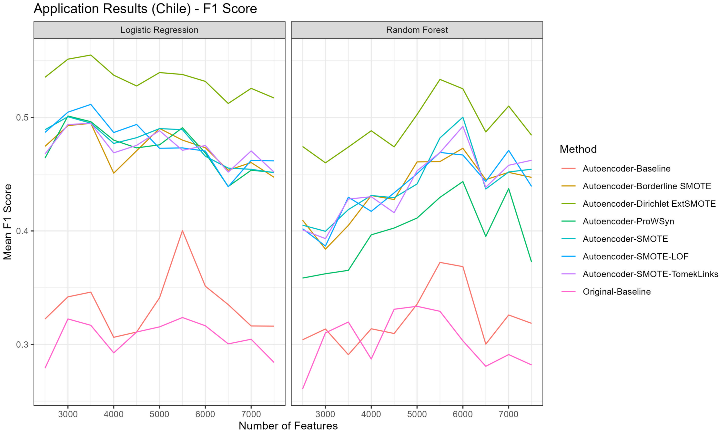



Figure: F1 score comparison with varying feature numbers.

# Application Results

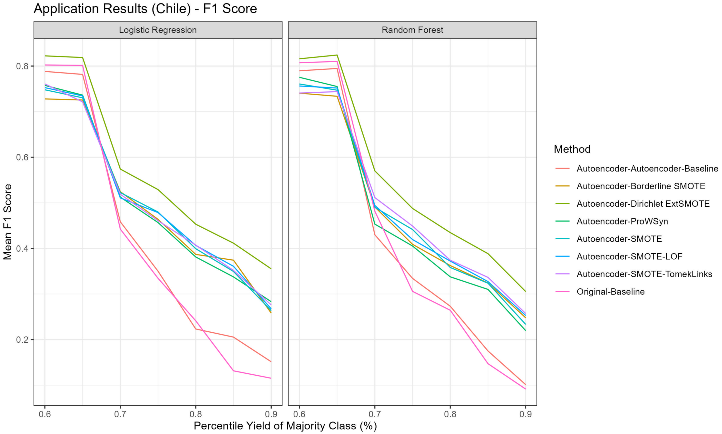



Figure: F1 score comparison with varying imbalance ratios.

# Conclusion

- Class imbalance is a significant problem in classification.
- Novel methods advancing imbalanced classification within machine learning.
- Effectively incorporate measures to minimize outlier effects and curse of dimensionality.
- Create more **accurate and reliable predictive models.**
- Across diverse domains, including fraud detection, medical diagnosis, and churn prediction.
- All the computing were done using Python on Digital Research Alliance of Canada computing cluster.

# References

[1] Bej, S., N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer (2021). Loras: an oversampling approach for imbalanced datasets. *Machine learning 110*(2), 279–301.

[2] Garzon, M. (2022). *Dimensionality reduction in data science*. Cham, Switzerland: Springer.

[3] Huynh, P.-H., V. H. Nguyen, and T.-N. Do (2020). Improvements in the large p, small n classification issue. *SN computer science 1*(4), 207–.

[4] Matharaarachchi, S., M. Domaratzki, , and S. Muthukumarana (2024). Deep-ExtSMOTE: Integrating autoencoders for advanced mitigation of class imbalance in high-dimensional data classification. *Journal of Data Science (In Review)*.

[5] Matharaarachchi, S., M. Domaratzki, A. Katz, and S. Muthukumarana (2022). Discovering long covid symptom patterns: Association rule mining and sentiment analysis in social media tweets. *JMIR formative research 6*(9), e37984–e37984.

[6] Matharaarachchi, S., M. Domaratzki, A. Katz, and S. Muthukumarana (2024). Long covid prediction in manitoba using clinical notes data: A machine learning approach. *Intelligence-Based Medicine (In Review)*.

[7] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2021). Assessing feature selection method performance with class imbalance data. *Machine learning with applications 6*, 100170–.

[8] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2022). Minimizing features while maintaining performance in data classification problems. *PeerJ. Computer science 8*, e1081–e1081.

[9] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*.

# Acknowledgment

I would like to express my special thanks of gratitude to

- my supervisors, Dr. Saman Muthukumarana and Dr. Mike Domaratzki, for their excellent guidance.
- Dr. Alan Katz and Dr. Max Turgeon for providing constructive feedback.
- Dr. Colin Garroway, for chairing the session today.
- my external examiner, Dr. Matthew Pratola, for taking the time to review my thesis and provide valuable feedback.
- the Manitoba Centre for Health Policy (MCHP) for providing the data.
- the University of Manitoba Graduate Fellowship and the Department of Statistics for funding and resources.
- my family and friends for their continuous support.

# Thank You!

Contact: matharas@myumanitoba.ca