ISLAMIC UNIVERSITY OF TECHNOLOGY(IUT)

LEARNING FROM IMBALANCED DATA

Author:

MD. SALMAN MOHOSHEU 190021218

MD. ABDULLAH AL NOMAN 190021236

AL-AMIN 190021210

Supervisor:

Mr. Asif Newaz

LECTURER

A thesis submitted as a Partial Fulfillment of the Degree of

Bachelor of Science (B.Sc.)

In the

Department of Electrical and Electronic Engineering (EEE)

LEARNING FROM IMBALANCED DATA

Approved by:

Lecturer Mr. Asif Newaz
Supervisor

Department of Electrical and Electronic Engineering

Islamic University of Technology (IUT),

Boardbazar, Gazipur-1704.

Date:

Declaration of Authorship

This is to certify that the work presented in this Thesis entitled, “Learning from Imbalanced Data”, is the outcome of the research carried out under the supervision of Mr. Asif Newaz, Lecturer, Islamic University of Technology.

Signatures of the Candidates

MD. SALMAN MOHOSHEU

MD. ABDULLAH AL NOMAN

AL-AMIN

Acknowledgment

First and foremost, we would like to extol The Almighty Allah (SWT), the Most Gracious, and the Most Merciful, for bestowing upon us the health, wisdom, and perseverance required to bring this thesis to fruition. His divine blessings have been our guiding light throughout this endeavor.

Our deepest gratitude extends to our beloved parents for their support, both financial and emotional. Their constant motivation, sacrifices, and boundless love have been the bedrock of our journey, and we owe them an immeasurable debt of thanks.

We also wish to convey our profound appreciation to our esteemed thesis supervisor, Mr. Asif Newaz, Lecturer, Department of Electrical and Electronic Engineering. His invaluable counsel, steadfast support, and insightful guidance have been indispensable to the successful completion of this thesis. His expertise, encouragement, and knowledge have profoundly enriched our academic experience, and we are truly honored to have had him as our mentor.

Furthermore, we extend our heartfelt thanks to our peers and friends for their encouragement and assistance throughout this journey. Their constructive feedback and camaraderie have been of immense value. Special thanks to Taufikur Rahman Fuad for his help and support in this research.

Lastly, we are deeply grateful to all the Department of Electrical and Electronic Engineering faculty members and staff for cultivating an environment conducive to learning and research. Their dedication to fostering academic excellence has greatly contributed to our intellectual growth and development.

Publications

Journal:

Asif Newaz, Md Salman Mohosheu, Md. Abdullah Al Noman. “Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques”

https://doi.org/10.1016/j.imu.2023.101361

Conference:

Md. Salman Mohosheu, MD. Abdullah al Noman, Asif Newaz, Al-Amin, Taskeed Jabid. “A Comprehensive Evaluation of Sampling Techniques in Addressing Class Imbalance Across Diverse Datasets”

DOI: 10.1109/ICEEICT62016.2024.10534464

Codes and additional resources:

https://github.com/salman-84/Learning-from-imbalanced-data

Abstract

Class imbalance is a common challenge in real-world datasets. In critical applications such as medical diagnosis, intrusion detection, fault detection, and disease identification. In most of these cases, the positive examples are very rare. For this, machine learning models often get biased towards to negative class and identify any unseen samples as negative class examples. This imbalance mostly favors the majority class, resulting in poor prediction performance for the minority class. This thesis thoroughly evaluates various state-of-the-art methods for addressing class imbalance over 100+ datasets with different imbalance ratios. A through experimental analysis have been done to find out the patterns of the outcomes. By experimenting with numerous sampling strategies, including under-sampling, over-sampling, and hybrid approaches, this study highlights the strengths and weaknesses of each technique. Additionally, we explored the impact of class overlap, a condition where instances of different classes share similar features, further complicating predictive modeling. The findings underscore the necessity of combining sampling methods with cost-sensitive learning to improve prediction accuracy and generalization. The research introduces novel hybrid approaches that optimize the balance between majority and minority classes, demonstrating significant improvements in performance. These advancements contribute valuable insights and methodologies for future research and practical applications in handling imbalanced data.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ML	machine learning
IR	imbalance ratio
SMOTE	Synthetic Minority Over-sampling Technique
BRF	Balanced Random Forest
RUS	Random Undersampling
iBRF	enhanced Balanced Random Forest classifier
RF	Random Forest
ADASYN	Adaptive Synthetic Sampling technique
TP	True Positive
TN	True Negative
FN	False Negative
FP	False Positive
F-1 score	F-score of 1(harmonic mean of the precision and recall)
G-MEAN	geometric mean
ROC	Receiver Operating Characteristic
AUC-ROC	Area under the curve of Receiver Operating Characteristic
TPR	true positive rate
FPR	false positive rate
SMOTE-IPF	SMOTE with Instance Pre-selection Function
SMOBD	SMOTE with Borderline Distance
G-SMOTE	Geometric SMOTE
MCC	Matthews Correlation Coefficient

CNN	Condensed Nearest Neighbor
ENN	Edited Nearest Neighbors
OSS	One-Sided Selection
IHT	Instance Hardness Threshold
NC	Neighborhood Cleaning
IBA	Instance-Based Algorithm
GAN	generative adversarial networks
OBE	Outcome Based Education

Introduction

The majority of real-world datasets contain some degree of imbalance. The data may be very distorted, particularly in some important applications. It becomes very difficult for the machine learning classifier to learn from the skewed data sets. So, to create a trustworthy prediction framework, the appropriate measures must be taken to solve the class imbalance issue [[1].

Predictive modeling frequently faces class imbalance, especially in the medical field. The majority and minority classes in many real-world applications, such as the prediction of complications from myocardial infarction, and Intrusion Detection System are notably imbalanced in datasets. The minority class, which is usually the class of more interest in different settings, experiences subpar prediction performance as a result of this imbalance, which biases machine learning (ML) algorithms toward the majority class. Many methods have been developed to address this issue, but their success varies across different datasets and levels of imbalance. This thesis aims to thoroughly evaluate the performance of various advanced methods on different imbalanced datasets, highlighting the strengths and weaknesses of each approach. An uneven distribution of classes, where the minority class is considerably underrepresented in comparison to the majority class, is a common feature of imbalanced datasets. To address the effectiveness of various imbalance handling strategies, the study entailed a great deal of experimentation on more than 100 datasets with different imbalance ratios. The datasets cover a wide range of imbalance ratios, from slightly skewed distributions to highly skewed ones. For example, in certain datasets, the minority class makes up about 30–40% of the total samples, resulting in a relatively low imbalance ratio. The minority class may make up between 10 and 20 percent of the samples in other datasets. Prediction performance can be enhanced and the class imbalance problem efficiently addressed by sampling strategies. But the main factors that determine success are the resampling technique used and other intrinsic properties of the data, like class overlap, noisy sample presence, imbalance ratio (IR), etc.[2]

1.1 Class Imbalance and Class Overlap:

1.1.1 Class Imbalance:

Class imbalance affects many real-world applications and is a major difficulty in predictive modeling. It happens when there are disproportionately more instances of one class (the majority class) in the dataset than in the other class (the minority class). The machine learning algorithms become biased in favor of the majority class because of this data imbalance. The minority class is frequently of greater relevance in many domains, such as fraud detection, disease diagnosis, intrusion detection, and customer churn prediction, because it represents the essential occurrences that must be accurately detected. For example, in the minority class of medical diagnosis, there may be examples of uncommon but critical illnesses that, although underrepresented in the dataset, require proper detection. Reducing class disparity is essential to creating accurate prediction

models. Commonly employed methods include under-sampling, which lowers the number of majority class samples, and oversampling, which raises the number of minority class samples. To enhance model performance, hybrid approaches that combine these techniques with cost-sensitive learning—giving larger penalties for incorrectly categorizing instances of minority classes—are also used. Recently, Generative adversarial network is also used to produce examples of minority classes [3]. By balancing the class distribution and improving the model's capacity to correctly anticipate the minority class, these techniques contribute to the development of more reliable and efficient prediction systems in a variety of domains.

$$Imbalance\ ratio\ (IR)) = \frac{number\ of\ majority\ class\ samples}{number\ of\ minority\ class\ samples} \quad (1.1)$$

1.1.2 Class Overlap:

Predictive modeling is further complicated by the essential problem of class overlap in imbalanced datasets. When instances of the majority class and the minority class have similar traits, it is called class overlap, and this makes it harder for ML models to distinguish between the two. Predictive model performance can be severely harmed by this overlap since it complicates the decision boundaries and increases the rate of misclassification. Because the ML models with imbalanced datasets are already biased towards the majority class as a result of the imbalance, the problem of class overlap is made more difficult. This bias makes it more likely that occurrences of the minority class will be mistakenly identified as belonging to the majority class when classes overlap [4].

This overlap raises the possibility of incorrectly identifying occurrences of the minority class. To address this, several techniques are used, including cost-sensitive learning to assign higher penalties for misclassifying minority classes, advanced sampling techniques like SMOTE to generate better representative samples, and ensemble methods like Balanced Random Forest to integrate multiple weak learners for better handling of overlapping classes. These solutions improve overall performance in many real-world applications by effectively addressing class overlap and enhancing the model's capacity to forecast minority class instances.

Class overlap has not been mathematically well characterized, and no standard measurement exists. This research adapts a measure from Garcia et al. [5] to calculate the overlap degree concerning the minority class area, addressing potential biases caused by class imbalance.

$$Degree\ of\ Overlap\ (\%) = \frac{overlaping\ area}{minority\ class\ area} * 100 \quad (1.2)$$

1.2 Motivation

Imbalanced data is a common and important problem in the fields of machine learning and data mining. Datasets that exhibit a considerable overrepresentation in one class relative to other classes are considered imbalanced. Predictive model performance and dependability are severely hampered by this mismatch. Here, we experimented some ways to tackle the issue of learning from imbalanced data.

Firstly, in Real-World Applications, ensuring effective learning from imbalanced data is essential for deploying robust and reliable machine learning models in these domains, where the minority class often represents the most critical outcomes. Imbalanced data is common in many real-world applications, such as fraud detection, where fraudulent transactions are much rarer than legitimate ones, and medical diagnosis, where certain diseases are significantly less common than others.

Secondly, Performance Degradation of Standalone Algorithms, when used to imbalanced datasets, traditional machine learning techniques often presume a balanced class distribution and hence perform poorly. Low prediction accuracy for the minority class is caused by this bias towards the dominant class. This might result in potentially disastrous outcomes in situations such as medical diagnosis, when uncommon but critical illnesses go unidentified. Creating strategies to address class imbalance is crucial to raising these models' overall effectiveness and dependability.

Thirdly, Societal and Ethical Implications, ignoring class disparity can have serious negative effects on society and ethics, particularly when it comes to applications that deal with justice and equity. Minority groups may, for example, be underrepresented in loan approval procedures, which can produce biased results and reinforce prejudice. By addressing imbalanced data, models become more egalitarian and fairer, advancing social justice and lowering systematic prejudice.

Fourthly, Advancements in Machine Learning Techniques and the problem of imbalanced data have led to significant progress in machine learning. Scholars have devised an array of tactics, including cost-sensitive learning, resampling techniques (oversampling the minority class or under-sampling the majority class), and specialized algorithms intended to better manage imbalance. These developments not only enhance model performance on imbalanced datasets but also add significant new techniques and perspectives to the machine learning community at large.

Another one is Improved Decision-Making. Precise forecasts for the underrepresented group are frequently essential for efficient decision-making across diverse fields. For example, finding uncommon failure occurrences in predictive maintenance can save expensive downtime and increase operational effectiveness. Accurately identifying uncommon diseases in the medical field can result in early treatments and improved patient outcomes. Through an emphasis on learning from imbalanced data, we may improve our ability to make decisions in a variety of contexts.

1.3 Research Objectives:

This study's main goal is to thoroughly assess and improve the application of diverse sampling strategies in order to address the problem of class imbalance in predictive modeling. In particular, we approached to,

I) Examine the Effects of Sampling Methodologies:

Perform a thorough experimental investigation to determine the advantages and disadvantages of both widely used and cutting-edge sampling strategies. To ensure thorough examination, a variety of 100 datasets with different imbalance ratios are used. Examine more than 35 distinct sample strategies, both established and new, to determine how effective they are in various imbalanced situations

II) Identify and address common issues related to class imbalance: Reduce the bias that machine learning models have in favor of the majority class, which frequently results in less-than-ideal performance.

Inadequate Generalization: To increase prediction models' performance on data that hasn't been observed yet, strengthen their generalization skills.

Increased False Negatives: Lower the false negative rate, which is essential in applications like fraud detection and medical diagnosis where the minority class is of greater concern.

Misclassification Costs: Use cost-sensitive learning and sophisticated sampling strategies to reduce the significant costs connected with misclassifications, particularly for the minority class.

III) Create Novel Approaches:

Create innovative procedures that combine the advantages of several sampling strategies based on the results of our thorough investigation. The objective of these novel methods is to achieve the best possible balance between under-sampling the majority class and oversampling the minority class. Provide hybrid approaches that combine cost-sensitive learning with sampling to improve prediction performance and lessen the drawbacks of applying these strategies independently.

IV) Provide Novel Approaches for Upcoming Studies:

Offer perspectives and theories to direct future studies in the area of imbalanced learning. This involves proposing fresh approaches to class overlap, feature engineering, and ensemble method integration. To further enhance the handling of imbalanced datasets, promote the investigation of novel sampling strategies and their applications in many fields.

1.4 Contributions

1.4.1 Extensive Experimental Analysis:

Our thesis makes a significant contribution in the form of a thorough experimental study that aims to reveal the advantages and disadvantages of both widely used and cutting-edge sampling strategies. This research encompasses more than 100 datasets that illustrate many real-world settings, including healthcare, banking, and network security. The datasets display a broad range of imbalance ratios, from slightly skewed distributions to highly skewed ones. We do a thorough analysis and comparison of over 34 distinct sampling strategies, encompassing both state-of-the-art and popular techniques such as SMOTE and ADASYN. We use all the chosen methods for each dataset, training machine learning models and evaluating their efficacy with cross-validation to guarantee robustness. Metrics like accuracy, recall, F1-score, and AUC-ROC are used to gauge the models' effectiveness. This thorough assessment offers insightful information on the advantages and disadvantages of every strategy, illustrating the situations in which particular approaches work well or poorly. By applying these ideas, practitioners may optimize model performance by making well-informed judgments about which strategies to use in particular settings. Additionally, our results advance the field by pointing out shortcomings in existing approaches and providing guidance for future studies aimed at creating more potent remedies for imbalanced data processing.

1.4.2 Development of New Methodologies:

Building on the knowledge gathered from our thorough investigation, we have created new approaches intended to address the difficulties presented by imbalanced data, improving the functionality of machine learning models in these situations. These methods improve upon the drawbacks of previous approaches by including a number of novel strategies and improvements. For instance, we may incorporate cost-sensitive learning strategies that give classes varying weights according to their significance, sophisticated resampling techniques that carefully choose samples to better balance the dataset, and hybrid approaches that combine several techniques to take advantage of their advantages. To make sure these new approaches are reliable, adaptable, and have good generalization, they have been put through a thorough testing process on a variety of datasets with varying imbalance ratios. Our objective is to offer useful technologies that are easily implemented in real-world scenarios where precisely identifying the minority class is essential, such as fraud detection, medical diagnostics, and anomaly detection. Our methodologies bridge the gap between theoretical research and practical application, thereby improving decision-making and prediction accuracy in critical areas while simultaneously advancing the state of research and providing immediate benefits to practitioners dealing with imbalanced data.

1.4.3 Suggesting New Directions in Research:

Our thesis not only proposes new theories and suggests directions for future study, but it also goes further into the field of imbalanced data. Our aim is to broaden the theoretical framework related to the comprehension and handling of imbalanced datasets in addition to introducing novel methodologies. Through the application of our research findings, we provide new ideas and methods that might revolutionize the field of data science and machine learning. These realizations are meant to act as stimulants for more research and inquiry rather than being restricted to the pages of our thesis. Our goal is to encourage scholars to investigate previously undiscovered or understudied areas by setting out into new territory. We anticipate that by working together, we will be able to create fresh approaches and strategies that will successfully tackle the enduring problems caused by class disparity. Our ultimate goal is to make a significant contribution to the field's continued growth, encouraging an innovative culture and advancing machine learning and data science.

Chapter 2

Literature Review:

The majority of real-world datasets have some degree of imbalance. The data may be very distorted, particularly in some important applications. To create a trustworthy prediction framework, appropriate actions must be taken to address the issue of class imbalance [1].An efficient way to address the imbalanced classification issue is to resample the data to balance the uneven class distribution before training the model. Oversampling and under-sampling are two major categories under which various resampling approaches have been proposed throughout the years [6]Under-sampling removes instances of the majority class from the data, whereas oversampling creates new minority-class samples. Two non-heuristic techniques are random under-sampling (RUS) and random oversampling (ROS), wherein the majority-class instances are randomly deleted to reduce the number of samples and the minority-class instances are duplicated to increase the number of samples, respectively. Several alternative heuristic methods have been devised to produce novel synthetic samples instead of merely replicating them [7]. To address the issue of class imbalance, scholars have developed a variety of strategies across time. This covers many SMOTE technique iterations [8]. The blending of cost-sensitive learning and sampling methods [9], the hybridization of various sampling strategies [10], and the integration of evolutionary algorithms with sampling techniques [11]. Gyorgy Kovac illustrated 85 SMOTE-variants' performance on a variety of imbalanced datasets [12]. After analyzing thus many permutations, the author came to the conclusion that there was no discernible change in performance. Although numerous adaptations of the original algorithm were developed as a result of the SMOTE approach's success, the oversampling technique is typically insufficient to provide the desired performance [13].Hybridization appears to be a more promising concept, and various methods for it have been devised. For example, Xu et al. presented a hybridization between M-SMOTE and ENN for medically imbalanced data in and compared the performance of various approaches using the MCC score [14]. It is feasible to hybridize other techniques in different ways, but more research is needed in this area. A significant problem with the sampling approaches is that they lead to an increase in variance and a loss of generalizability in the models [15].In this context, using ensemble learning makes sense, and researchers have created fresh ways to apply ensemble techniques to issues with class imbalance. As stated in[10], an ensemble strategy called "Random Balance" was proposed by Dıez-Pastor et al. to provide more diversity. It uses the SMOTE algorithm to randomly balance various subsets of the ensemble. The technique was further expanded by the authors to include multiclass imbalanced circumstances [16]. To enhance the performance of the ensemble techniques, Ribeiro et al. presented a multi-objective optimization strategy in [17]. They improved the model's performance by testing it on an actual anomaly detection problem. Yang et al. [18] presented a hybrid ensemble classifier in a different work that blends cost-sensitive learning with density-based under-sampling. Both bagging and boosting frameworks can be used with the sampling strategies. However, as the actual results from this study show, simply merging the approaches with the ensemble methodologies does not increase the prediction performance. Prior to making it easier to distinguish between the two classes, the sampling strategy must be able to generate appropriately resampled data. To improve generalization, these sampling strategies can be further integrated into the framework for ensemble learning. Multiple weak learners, each trained on a distinct subset of the data, make up ensemble algorithms. To create a more reliable classifier, the predictions from several weak learners—

typically decision trees (DT)—are added together. These ensemble methods frequently yield better results than a single DT or other classifiers and have lower bias and volatility. One such ensemble method that employs RUS to resample every bootstrap subset of the data before training the DTs is the Balanced Random Forest (BRF) classifier [19]. Chawla et al.[8] were the ones who first presented the SMOTE method. Over time, this heuristic sampling method has grown in popularity. The algorithm has been widely applied in numerous fields since its creation. A comprehensive overview of the many SMOTE algorithm enhancements created during the past 10 years was given by Fernandez et al. [8]. The writers did not, however, offer a critical evaluation of these methods' effectiveness. To perform their research, Blagus et al. [20] tested the SMOTE algorithm in high-dimensional environments. The authors looked at the SMOTE algorithm's applicability to high-dimensional data from both an empirical and theoretical standpoint. They also looked into how SMOTE-based oversampling impacts various classification systems. According to the authors' research, the SMOTE method works well in low-dimensional environments but not in high-dimensionally imbalanced data. Before using SMOTE, they advised using feature selection approaches to reduce the dimension size. In their investigation, Kovacs et al. investigated the performance of several SMOTE extensions by classifying them according to the techniques used in [12], [21]. Although they were able to pinpoint some of the most effective methods, their research was restricted to SMOTE extensions alone. In a lithological classification situation using geophysical data, Nugroho et al. [22] examined several imbalanced learning strategies, including class-weight tuning and hybrid sampling. They discovered that oversampling methods produced more trustworthy outcomes than others. An experimental design was developed by Prati et al.[23] to assess the effectiveness of four distinct sampling strategies on 22 imbalanced datasets. They also worked with one of the most widely used cost-sensitive algorithms, MetaCost. The results of our investigation are consistent with their study, which found no discernible benefit of SMOTE over its variations. A thorough analysis of the methods applied in multi-label imbalanced classification problems was given by Tarekegn et al. [24].In their study, they did not conduct any experimental analysis to evaluate the state-of-the-art approaches they highlighted for the task.

Methodology

3.1 Overview:

Predictive modeling frequently faces class imbalance, which presents serious difficulties for vital applications including risk assessment, fraud detection, and medical diagnostics. A dataset that is imbalanced exhibits a disproportionate distribution of classes, which can lead to a number of serious issues. Machine learning algorithms have a tendency to become biased in favor of the majority class, producing models that forecast the majority class more often than the minority class, which is usually of greater interest. This bias results in poor generalization performance, especially when it comes to detecting occurrences of minority classes, which has serious real-world ramifications like missed medical diagnoses or undiscovered fraud. Imbalanced datasets also raise the risk of false negatives, which are particularly troublesome in industries like healthcare since they mistakenly classify cases of the minority class as instances of the majority class. Traditional algorithms do not sufficiently account for these varied misclassification costs, which results in inefficient decision-making. The costs associated with misclassifications are frequently larger for the minority class. The methods of oversampling, under-sampling, and cost-sensitive learning that are currently in use to address class imbalance have drawbacks. Oversampling can cause overfitting, under-sampling can cause the loss of important information, and cost-sensitive learning might not be able to adequately handle the complexity of imbalanced datasets found in the real world. In order to tackle these issues, this thesis carried out a thorough experimental examination of more than 35 cutting-edge sampling strategies on a wide range of 100 datasets with different imbalance ratios. The objective is to determine the advantages and disadvantages of these approaches and create new approaches that combine their benefits. This will result in solid solutions for enhancing model performance when a class imbalance exists and advancing the cause of more precise and dependable predictive modeling in a range of important applications.

3.2 Classification Algorithms:

3.2.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised learning algorithms used for classification and regression. SVM works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. This involves using support vectors, which is equidistant from both positive and negative examples. Besides, the kernel trick allows SVM to handle non-linear boundaries by transforming data into higher-dimensional spaces. SVM is particularly effective in high-dimensional spaces.

3.2.2 Random Forests (RF)

Random Forests (RF) is a robust ensemble learning algorithms used for classification and regression tasks. It constructs multiple decision trees using bootstrap samples of the training data and aggregates their predictions to improve accuracy and reduce overfitting. By randomly selecting subsets of features for splitting at each node, RF introduces additional randomness, enhancing model generalization. RF is used as a base model in almost every case for its robustness.

3.3 Evaluation Metrics:

3.3.1 Accuracy:

Accuracy measures the percentage of correct predictions among all examples. It is calculated by dividing the number of correct predictions by the total number of predictions. Accuracy is useful for balanced datasets with evenly represented classes but may not fully represent model performance on imbalanced datasets. While straightforward, accuracy is a starting point for evaluating a model's overall predictive performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{3.1}$$

3.3.2 Precision:

Precision measures the percentage of correctly predicted positives among all instances predicted as positive. It is calculated by dividing the number of true positives by the sum of true and false positives. Precision is crucial in scenarios where false positives carry significant financial risk, as it highlights the model's ability to avoid them. High precision indicates a reliable model with a low false positive rate. However, for a complete evaluation of a model's performance, precision should be considered alongside other metrics like recall.

$$\text{Precision} = \frac{TP}{TP+FN} \tag{3.2}$$

3.3.3 Specificity:

Specificity measures a model's ability to correctly identify negative examples in binary classification tasks. It is calculated by dividing the number of true negatives by the sum of true negatives and false positives. Specificity indicates the percentage of accurately identified negative cases and works with sensitivity (recall) to evaluate a model's effectiveness. High specificity signifies a low false positive rate, which is vital in scenarios where correctly recognizing negative instances is crucial. For a complete understanding of a model's performance, specificity should be considered alongside other metrics.

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (3.3)$$

3.3.4 Recall (Sensitivity):

Recall, also known as sensitivity or the true positive rate, measures a model's ability to identify all relevant instances of a given class. It calculates the percentage of actual positive cases correctly predicted as positive. Recall is crucial when missing positive cases has serious consequences. It is computed by dividing the number of true positives by the sum of true positives and false negatives. High recall indicates a low false negative rate, showing the model's effectiveness in capturing positive examples. For a comprehensive evaluation of model performance, recall should be considered alongside other metrics like accuracy.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3.4)$$

3.3.5 F1- SCORE:

The F1 score provides a balanced measure of a model's performance by combining precision and recall into a single metric. It is the harmonic mean of precision and recall, useful for imbalanced datasets or when false positives and false negatives carry equal costs. A high F1 score indicates the model effectively balances precision and recall, demonstrating robust performance across multiple parameters.

$$\text{F1} = \frac{2 \times precision \times recall}{precision+recall} \quad (3.5)$$

3.3.6 G-MEAN:

The G-Mean measures a model's performance by calculating the geometric mean of sensitivity (true positive rate) and specificity (true negative rate). It is particularly useful for imbalanced datasets, as it equally weighs both metrics. A high G-Mean indicates the model effectively classifies both positive and negative examples, providing a single number that represents balanced performance across classes.

$$\text{G-mean} = \sqrt{\text{Sensitivity x Specificity}} \quad (3.7)$$

3.3.7 ROC-AUC Score:

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance measurement for classification problems at various threshold settings. It quantifies the overall ability of the model to distinguish between positive and negative classes, with values ranging from

0 to 1, where a higher value indicates better performance. An AUC of 1 represents perfect classification, while an AUC of 0.5 suggests no discriminative power, equivalent to random guessing.

$$\text{TPR}=\frac{TP}{TP+FN} \qquad (3.8)$$

$$\text{FPR}=\frac{FP}{FP+TN} \qquad (3.9)$$

3.3.8 MCC:

The Matthews Correlation Coefficient (MCC) assesses binary classification performance, especially with imbalanced datasets. It considers true positives, true negatives, false positives, and false negatives, producing a value between -1 and 1. An MCC of 1 indicates perfect prediction, 0 indicates random guessing, and -1 indicates total disagreement. MCC is a comprehensive measure, accounting for all confusion matrix elements and remaining stable with imbalanced data. The definition of MCC in mathematics is:

$$\text{MCC}=\frac{TP\,.\,TN-FP\,.\,FN}{\sqrt{(\text{TP}+\text{FP}).\ (TP+FN).\ (TN+FP).\ (TN+FN)}} \qquad (3.10)$$

3.4 Handling Imbalanced Dataset

Imbalanced data handling techniques can broadly be divided into two categories: Data Level Modification and Algorithmic Level Modification. Within Data Level Modification, strategies like oversampling, under-sampling, and hybrid sampling are examined. These techniques directly manipulate the dataset, ensuring a balanced representation of both majority and minority classes, thus fostering equitable learning. Concurrently, Algorithmic Level Modification strategies, including a cost-sensitive learning This approach not only mitigates the adverse effects of class imbalance but also enhance model performance. By experimenting with these methodologies, our thesis aims to contribute substantially to the development of machine learning models, particularly in domains where accurate classification of minority class instances is paramount.

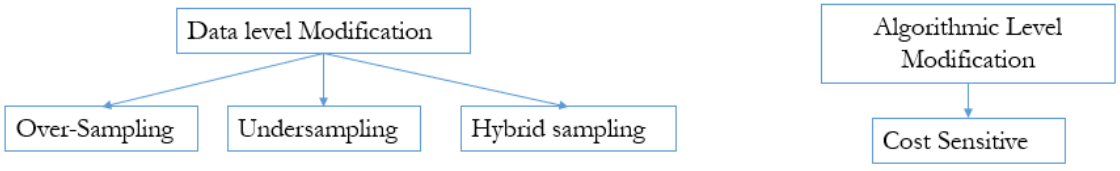

Fig. 1. Broad categories of imbalanced data handling techniques.

3.4.1 Data Level Modification

Oversampling:

Oversampling is a technique used in machine learning to address class imbalance, where one class has significantly fewer instances than the other. It involves increasing the number of instances in the minority class to balance the class distribution, thereby improving model performance by providing more representative training data.

Several oversampling methods have been developed to achieve this goal, each with its unique approach:

Random Oversampling (ROS): ROS duplicates randomly selected instances from the minority class to increase its representation in the dataset.

Synthetic Minority Oversampling Technique (SMOTE): SMOTE generates synthetic instances for the minority class by interpolating between existing minority class samples and their nearest neighbors in feature space. This method aims to create new, realistic minority class instances rather than simply duplicating existing ones.

Adaptive Synthetic Minority Oversampling Technique (ADASYN): ADASYN is an extension of SMOTE that adjusts the density of synthetic samples based on the local distribution of minority class instances. It focuses more on generating synthetic instances in regions where the class imbalance is more pronounced.

Borderline-SMOTE: Borderline-SMOTE is a variation of SMOTE that only generates synthetic instances near the decision boundary between classes, aiming to improve the generalization capability of the model.

Polynom-fit-SMOTE: This method fits a polynomial surface to the minority class instances and generates synthetic samples along this surface to better capture the underlying distribution of the minority class.

ProWSyn: ProWSyn combines both oversampling and under-sampling techniques by applying SMOTE to the minority class and simultaneously under-sampling the majority class, resulting in a more balanced dataset.

SMOTE-IPF: SMOTE with Instance Pre-selection Function (IPF) selects informative instances from the minority class to oversample, enhancing the quality of synthetic samples generated by SMOTE.

SMOBD: SMOTE with Borderline Distance (SMOBD) selects minority class instances based on their proximity to the decision boundary and then applies SMOTE to generate synthetic samples, focusing on regions crucial for classification.

G-SMOTE: Geometric SMOTE (G-SMOTE) adjusts the number of synthetic samples generated for each minority class instance based on its local geometric characteristics, providing a more adaptive oversampling approach.

Assembled-SMOTE: Assembled-SMOTE combines multiple base oversampling algorithms, such as SMOTE and ADASYN, to improve overall performance.

In our study, we conducted a comprehensive evaluation of various oversampling methods, including ROS, ADASYN, SMOTE, Borderline-SMOTE, Polynom-fit-SMOTE, ProWSyn, SMOTE-IPF, SMOBD, G-SMOTE, and Assembled-SMOTE. Utilizing metrics such as the Matthews Correlation Coefficient (MCC) and other complementary measures, we systematically assessed the performance of these techniques across diverse datasets and domains.

Our findings revealed significant variations in the effectiveness of oversampling methods, with certain techniques demonstrating superior performance in specific contexts. ROS, ADASYN, SMOTE, and their variations emerged as the most promising oversampling strategies, showcasing their versatility and efficacy in addressing class imbalance. Moreover, our study elucidated the nuanced impact of oversampling techniques on model performance, providing valuable insights for practitioners seeking to optimize machine learning models in imbalanced settings.

Under-sampling:

Under-sampling is a technique used in machine learning to address class imbalance by reducing the number of instances in the majority class. By randomly removing samples from the majority class, under-sampling aims to create a more balanced distribution of classes in the dataset, thus mitigating the dominance of the majority class during model training.

Several under-sampling methods have been developed to achieve this objective, each with its unique approach:

Random Under-sampling: Randomly selects a subset of instances from the majority class to match the size of the minority class. While simple to implement, random under-sampling may lead to the loss of valuable information and potentially important instances from the majority class.

NearMiss: NearMiss is a family of under-sampling techniques that select instances from the majority class based on their proximity to instances in the minority class. NearMiss variants include NearMiss-1, which selects instances from the majority class with the smallest average distance to the nearest k minority class instances, and NearMiss-2, which focuses on selecting instances from the majority class that are farthest from the minority class instances.

CNN: CNN is an iterative under-sampling technique that starts with a small subset of majority class instances and gradually adds instances from the majority class that are correctly classified by a nearest neighbor classifier trained on the minority class instances.

Tomek Links: Tomek Links are pairs of instances from different classes that are closest to each other. Under-sampling by removing Tomek Links eliminates instances that are ambiguous or noisy, thereby improving the separation between classes.

ENN: ENN is an iterative under-sampling technique that removes majority class instances whose class label differs from the majority class label of their k nearest neighbors. This process aims to remove noisy or misclassified majority class instances.

OSS: OSS combines the CNN and ENN algorithms to iteratively select a subset of majority class instances that are well-separated from the minority class.

IHT: IHT selects instances from the majority class that are misclassified with high certainty by a classifier trained on the original imbalanced dataset. This approach aims to retain instances that are more challenging to classify correctly.

These under-sampling techniques offer different trade-offs in terms of preserving information from the majority class, maintaining class distribution, and computational complexity. The choice of under-sampling method depends on the specific characteristics of the dataset and the goals of the machine learning task.

Hybrid Sampling:

Among the hybridized techniques used in practice are:

SMOTE-ENN: SMOTE-ENN combines the oversampling capabilities of SMOTE with the instance editing functionality of ENN. It first applies SMOTE to generate synthetic instances for the minority class and then uses ENN to remove noisy or misclassified instances from both the majority and minority classes.

SMOTE-Tomek: SMOTE-Tomek combines SMOTE oversampling with Tomek Links under-sampling. After generating synthetic instances with SMOTE, Tomek Links are used to identify and remove pairs of instances (Tomek Links) that are nearest neighbors but belong to different classes, thereby improving the separation between classes.

SMOTE-CNN: SMOTE-CNN combines SMOTE oversampling with CNN under-sampling. It first applies SMOTE to augment the minority class and then uses CNN to iteratively select a subset of majority class instances that are well-separated from the minority class.

SMOTE-NC: SMOTE-NC is a hybrid approach that combines SMOTE oversampling with a neighborhood cleaning step. After generating synthetic instances with SMOTE, instances are removed if they are misclassified by a k-nearest neighbor classifier trained on the original imbalanced dataset.

3.4.2 Algorithmic Level Modification

Cost-Sensitive Learning:

Cost-sensitive learning is a technique used to address class imbalance by modifying the learning algorithm to incorporate the costs associated with misclassifying minority class instances. Unlike traditional learning algorithms that treat misclassification errors uniformly across classes, cost-sensitive learning assigns higher penalties to errors made on the minority class, thereby emphasizing the importance of correctly classifying minority instances.

Additionally, frameworks like imbalanced-learn provide specialized tools and techniques for cost-sensitive learning in Python. Imbalanced-learn offers various resampling algorithms, ensemble methods, and metrics tailored for imbalanced datasets, facilitating the implementation of cost-sensitive learning approaches within machine learning pipelines.

Our findings highlighted the efficacy of cost-sensitive learning in improving model performance on imbalanced datasets, particularly when combined with appropriate resampling techniques and evaluation metrics. By incorporating cost-sensitive learning into our analysis, we provided valuable guidance for practitioners seeking to develop robust and accurate models in imbalanced learning scenarios.

3.4.3 Ensemble Techniques

Ensemble techniques offer a powerful strategy for addressing class imbalance by combining multiple base learners to create a more robust and accurate model. These techniques leverage the diversity of individual models to compensate for weaknesses and biases, ultimately improving overall performance on imbalanced datasets.

Several ensemble methods have been developed, each with its unique approach to leveraging multiple models:

Balanced Random Forest: Balanced Random Forest is an extension of Random Forest that incorporates techniques to address class imbalance. It adjusts the training process to give more weight to minority class instances, thereby improving their representation in the ensemble model.

EasyEnsemble: EasyEnsemble is a specific ensemble method designed for imbalanced datasets. It creates multiple balanced subsets of the original dataset by under-sampling the majority class and then trains individual base learners on each subset. By focusing on balanced subsets, EasyEnsemble can effectively mitigate the effects of class imbalance.

OverBoost: OverBoost is an ensemble learning technique that enhances classifier performance on imbalanced datasets by emphasizing the importance of data sampling. It increases the weights of misclassified minority class instances during the boosting process. This targeted resampling ensures that the classifier gives more attention to the minority class, reducing bias towards the majority class.

3.4.4 GAN

Generative adversarial networks (GANs) involve a dynamic adversarial process between two neural networks, the generator and the discriminator. The discriminator's main job is to examine data samples and determine which ones are genuine and which are made up by the generator. On the other hand, the generator creates random noise to create data samples, such pictures, in order to produce outputs that appear realistic. The generator refines its tactics to provide increasingly convincing results as the discriminator improves its capacity to distinguish between real and produced data through recurrent training. The two networks have a symbiotic relationship under this competitive learning framework, where the discriminator's judgment drives the generator's evolution to produce more authentic data, and the generator's advancement tests the discriminator's discriminatory abilities. GANs have become a highly versatile tool across multiple domains, exhibiting outstanding performance in applications from data synthesis, which makes it easier to generate synthetic data to supplement small training datasets, to domain adaptation, where models trained in one domain are effectively transferred to another. Moreover, GANs have made a substantial contribution to image production by making it possible to produce high-fidelity pictures for a variety of uses in entertainment, design, and the arts. Their capacity to provide accurate data has created new opportunities for creativity and innovation, with the potential to have a revolutionary effect on a variety of sectors. GANs have several difficulties despite their amazing success, such as mode collapse, stability problems during training, and the possibility of producing biased or unwanted results. As GANs continue to advance and find more complex uses, researchers are actively working to discover solutions to these problems. In conclusion, GANs are a revolutionary approach to deep learning, fostering innovation and progress in a variety of domains thanks to their framework for competitive learning and the mutually beneficial interaction between the discriminator and the generator.

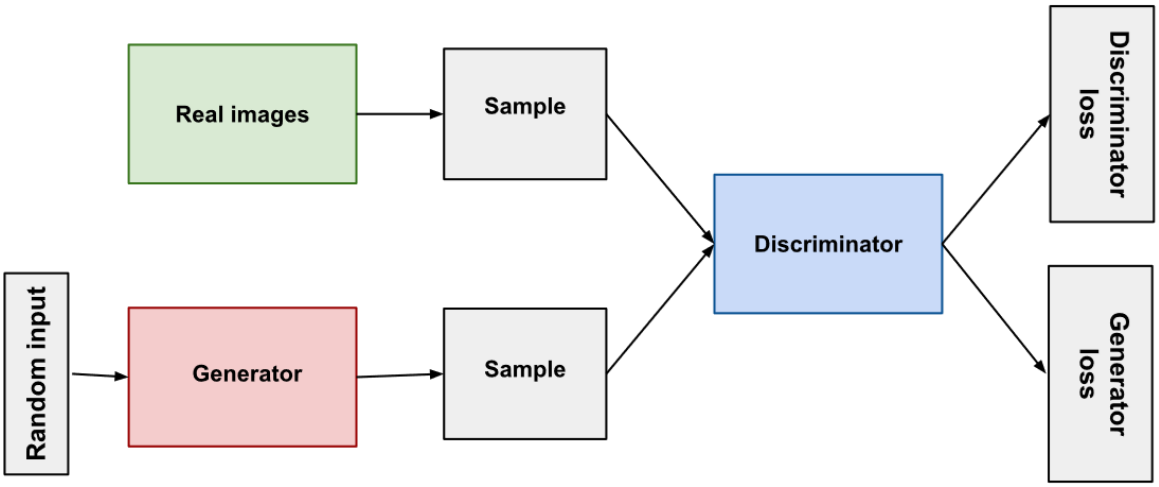

Fig. 2. Working procedure of Generative Adversarial Network (GAN)

Performance analysis of different imbalanced data handling techniques:

4.1 Overview:

In this section, the performance analysis of different imbalanced data handling techniques is discussed. One of the limitations of imbalanced data handling research is that the experiments are done with only a handful of datasets. However, in this research, the techniques are tested on a wide range of datasets with varied imbalanced ratios. In total 84 imbalanced datasets with different sample sizes, dimensions, overlapping, and class imbalance ratios (IR) are used in this experiment. By analyzing the results obtained, we attempt to address questions like how these algorithms perform with different degrees of imbalance, which techniques constantly provide better results, what hinders the performance of these approaches in high imbalance scenarios, and what are more suitable approaches for handling such scenarios.

4.2 Experimental Design:

Working with imbalanced data requires some careful handling. To avoid any data leakage, the data was first split into training and testing sets. Only the training set was resampled. The testing set was untouched and only used for validation purposes. A stratified 5-fold cross-validation scheme was undertaken, and the average of the results from five different testing folds was considered. The experiments were conducted on 84 imbalanced datasets collected from the KEEL dataset repository. A list of the datasets utilized in this study is provided in the associated repository of this paper. The IR of the datasets varied from 1.8 to 129. Five different measures are calculated to evaluate the performance: MCC, G-mean, ROC-AUC, sensitivity, and specificity. Support Vector Machine (SVM) and Random Forest (RF) classifiers are utilized as the learning algorithm. The outline of the experimental design is presented in Fig. 2. We have grouped the datasets into 3 IR categories: low imbalance (IR<10), mid imbalance (IR=10 to 30), and high imbalance (IR>30). The IR value ranges are chosen arbitrarily for ease of discussion. The default parameters of the Scikit-Learn, Imblearn, and smote_variants libraries were utilized in training the models.

To compare the performance of different approaches, we consider the MCC metric in particular. Other metrics such as accuracy, sensitivity, specificity or F1-score get biased towards one class and do not represent the entire scenario. MCC score takes into consideration all 4 confusion matrix parameters and only provides a high score when the classifier is performing well in all 4 categories– True positive, True negative, False positive, and False negative. Other metrics such as G-mean or ROC-AUC are based on sensitivity and specificity values. While these metrics are quite useful in imbalanced scenarios, they do not represent the overall classification performance. A

classifier may produce a very high sensitivity score (correctly classifying only a smaller number of minority class samples) but a comparatively lower specificity score (misclassifying a larger number of majority class samples). However, the g-mean score will still be high as the measure is the geometric mean of sensitivity and specificity. It does not take into account the total number of misclassifications made by the model, unlike MCC. MCC score will drop if too many misclassifications are made. Our goal is to assess the overall classification performance of the model and therefore, the MCC score is considered for evaluation.

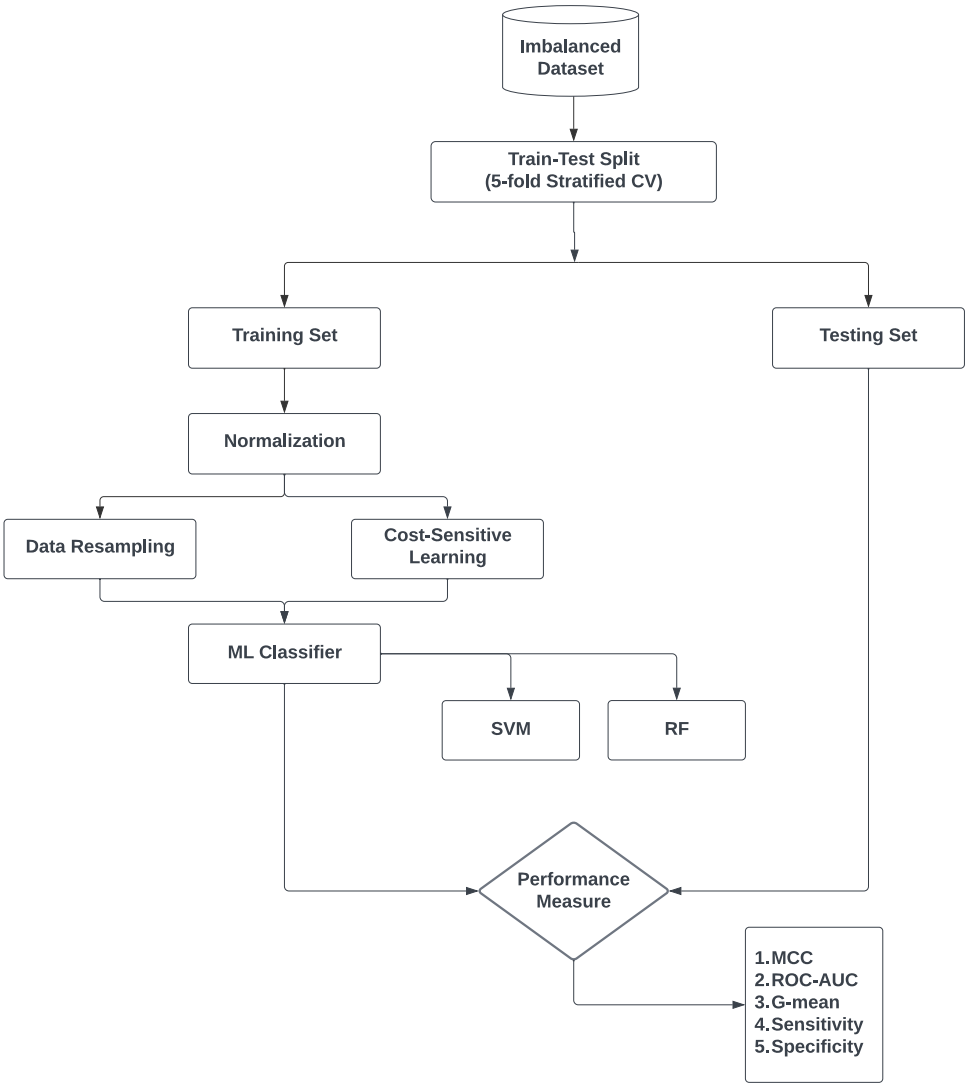

Fig. 3. The outline of the experimental design

4.3 Performance analysis of the oversampling techniques:

Oversampling techniques generate new data points for the minority class, making it easier for the machine learning algorithm to classify the test samples. In this study, we used 13 oversampling techniques including smote and its variants. The performance comparison of the oversampling techniques is shown in **Fig.4**. All the oversampling techniques showed performance improvements in different imbalanced scenarios. However, smote and its different variants e.g., LEE, CCR, LVQ-SMOTE, SMOBD, etc. showed similar performances. LEE scored the highest average MCC score of 69.6% for the datasets having an IR of less than 10. It performed quite well in highly imbalanced datasets as well. Polynom-fit-SMOTE produced the highest average MCC score of 59.6% for the datasets having IR within the range of 10 to 30 and 49.3% for IR more than 30. While a drop in performance in high imbalance cases is noticeable, it is noteworthy that among all the different sampling techniques, the OS algorithms provided the best performance.

The reason behind the drop in performance can be related to the fact that when the IR is high, a substantial number of new samples have to be generated to balance the dataset. When so many new samples are generated from only a handful of examples, noisy samples are generated that might not actually belong to the minority class, resulting in a poor MCC score. Another limitation of these OS techniques is that most of them do not attempt to reduce overlapping. While increasing the number of minority-class samples definitely improves the performance, we hypothesize that it can be further enhanced by reducing overlapping through additional means.

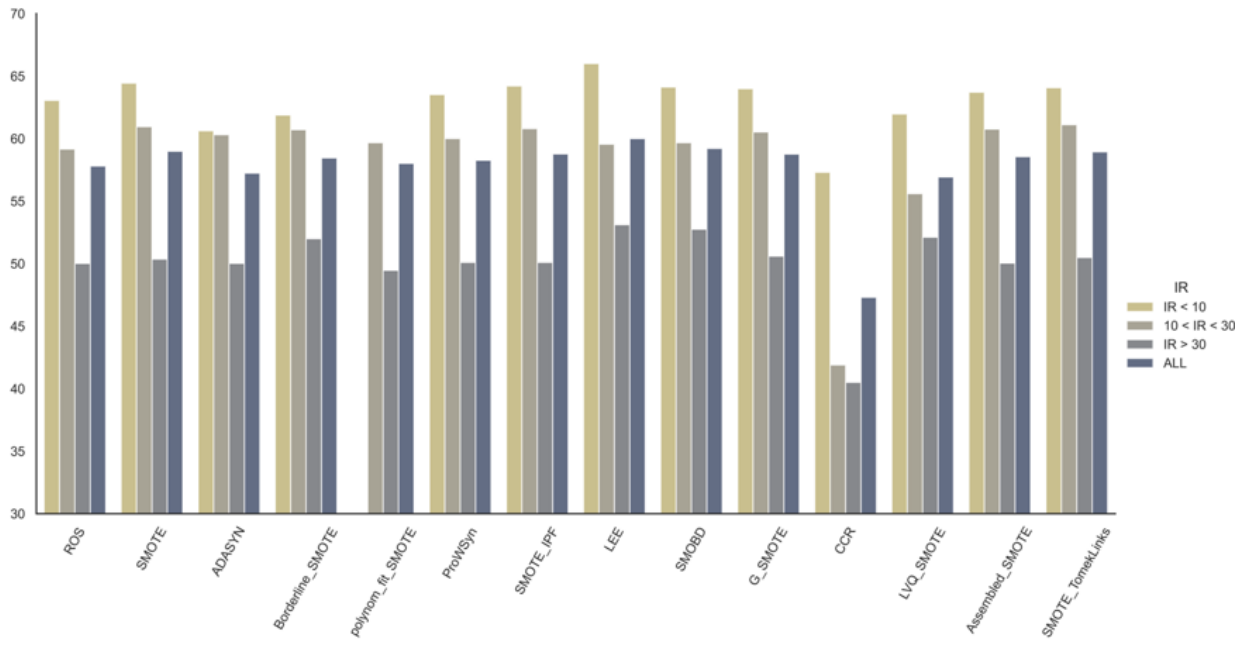

Fig. 4. Performance comparison of the oversampling techniques.

4.4 Performance analysis of the under-sampling techniques:

Under-sampling techniques are also very popular in handling imbalanced scenarios. A total of 7 US approaches were experimented in this study. While there is an overall improvement in performance with the application of these techniques, some critical observations were made. These techniques work well when the IR is lower. However, all the techniques showed a drastic decline in performance when the IR increased. Neighborhood Cleaning Rule (NC) showed the best performance among the under-samplers. NC scored 68.25% in terms of MCC for the datasets having an IR of up to 10. The average MCC score of NC drops to 54.65% when the IR is in-between 10 to 30. The performance further drops by 12% when the IR becomes more than 30. A similar drop in performance is observed in other US algorithms as well. A detailed performance comparison of the under-sampling techniques is provided in **Fig. 5.**

One of the key reasons behind this can be traced back to the fact that US techniques remove samples from the data to reduce overlapping and IR. 5 of these US techniques focus on reducing overlapping, while only RUS and Near-miss attempt to balance the class distribution. Now, when the IR is lower, these overlapping-based techniques successfully manage to alleviate the scenario and thus provide high prediction performance. However, when the IR is increased, these methods fail to adequately balance the data. While they may reduce overlapping to some extent, the data remains quite imbalanced and the bias persists towards the majority class. On the other hand, to balance the distribution, RUS or Near-miss algorithms remove samples which leads to information loss. With higher imbalances, significant loss of information can occur. This makes the system quite unreliable in making accurate predictions. Moreover, these US techniques do not increase the presence of minority-class samples in the data. Consequently, the classifier fails to correctly identify the rare samples in the data. This explains why the performance of the US techniques is comparatively much lower than the OS approaches.

Based on the experimental results, it can be concluded that while US strategies can effectively handle lower-class class-imbalance scenarios, the techniques are completely inappropriate and should be avoided when the data is highly imbalanced.

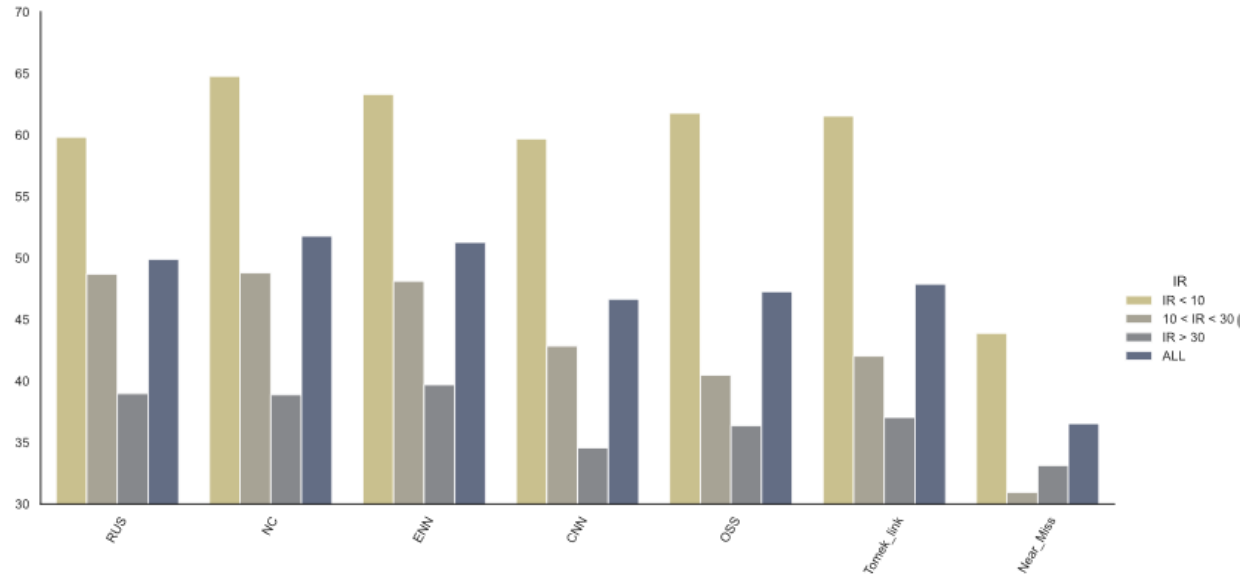

Fig. 5. Performance comparison of the under-sampling techniques

4.5 Performance analysis of the hybrid sampling techniques:

The hybrid sampling approach combines both OS and US strategies to provide a more balanced sampling method. Two popular hybrid sampling approaches, SMOTE-ENN and SMOTE-Tomek, were tested in this study. Both approaches provided significantly better results than their under-sampling counterparts. However, they did not provide much improvement compared to the SMOTE algorithm. Their performance was comparable with the other OS techniques. This type of hybridization has good prospects, especially in the higher imbalanced scenarios. Despite its potential advantages, it is noteworthy that the exploration of hybrid methods remains relatively limited in the existing literature. Hybridization can address both the overlapping and IR problems simultaneously and provide better results. Further investigation and empirical studies are suggested to comprehensively assess its efficacy and applicability in diverse datasets. In this experiment, we used two popular hybrid sampling techniques available in the imbalanced-learn library. The performance comparison of the hybrid sampling techniques is provided in **Fig. 6.**

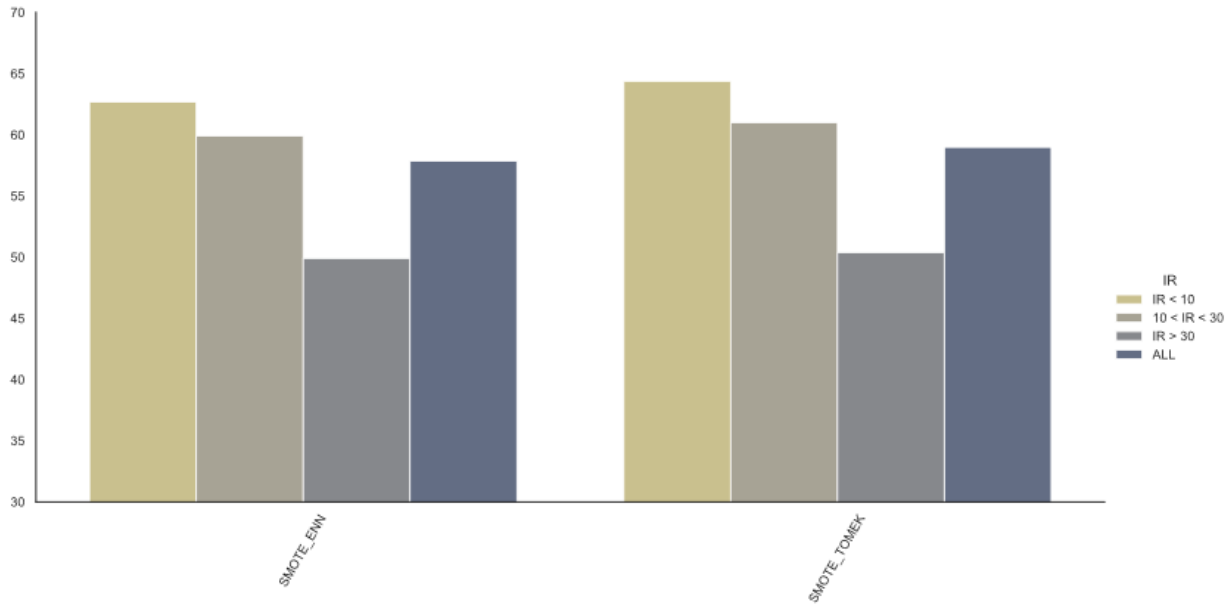

Fig. 6. Performance comparison of the hybrid sampling techniques

4.6 Performance analysis of the ensemble algorithms:

Ensemble algorithms combine multiple weak learners to provide more robust prediction performance. In this experiment, we tested 7 popular ensemble algorithms used in imbalanced learning. Average mcc scores obtained from the ensemble algorithms are provided in details in Table 1. The base learner for the ensembles is the Decision Tree. SMOTE-Bagging and Over-Bagging are two of the OS-based bagging ensemble techniques. RUSBoost and Over-Boost are two of the boosting-based ensemble approaches. While the goal of the ensemble approach is to improve performance, surprisingly, that kind of improvement was not achieved in imbalanced cases. Most of the ensembles could not even outperform the US techniques and their performance was considerably poor in highly imbalanced datasets. The results from the ensemble algorithms are reported in Table 3. Among these algorithms, the Over-boost technique performed the best. It scored 63.55% MCC when the IR < 10. The performance decreased to 58.50% as the IR increased up to 30. Significant degradation in performance is observed as the IR further increases. For the datasets having an IR of more than 30, the approach scored only 43.72%. The performance comparison of the ensemble algorithms is shown in **Fig. 7.**

Bagging ensembles work based on bootstrapping that creates random subsets of data. Weak learners are trained individually on these subsets and predictions are later aggregated. However, this does not mitigate the imbalance issue. So, the bootstrap subsets are resampled using RUS or SMOTE to attain balance. However, the original issues associated with these sampling techniques persist. Creating an ensemble incorporating the RUS approach reduces the chance of information loss, resulting in a slightly better performance from BRF or BB over RUS. However, as the RUS algorithm does not fix the class overlapping issue, it persists in the ensemble approaches, resulting in poorer performance compared to other US techniques. From the experimental results, it is evident that directly preprocessing the entire data with SMOTE or its variants provides better

results than merging them with the ensemble learning frameworks, especially for the higher imbalanced cases. To obtain better results, the underlying sampling techniques first need to mitigate the issues of imbalanced learning.

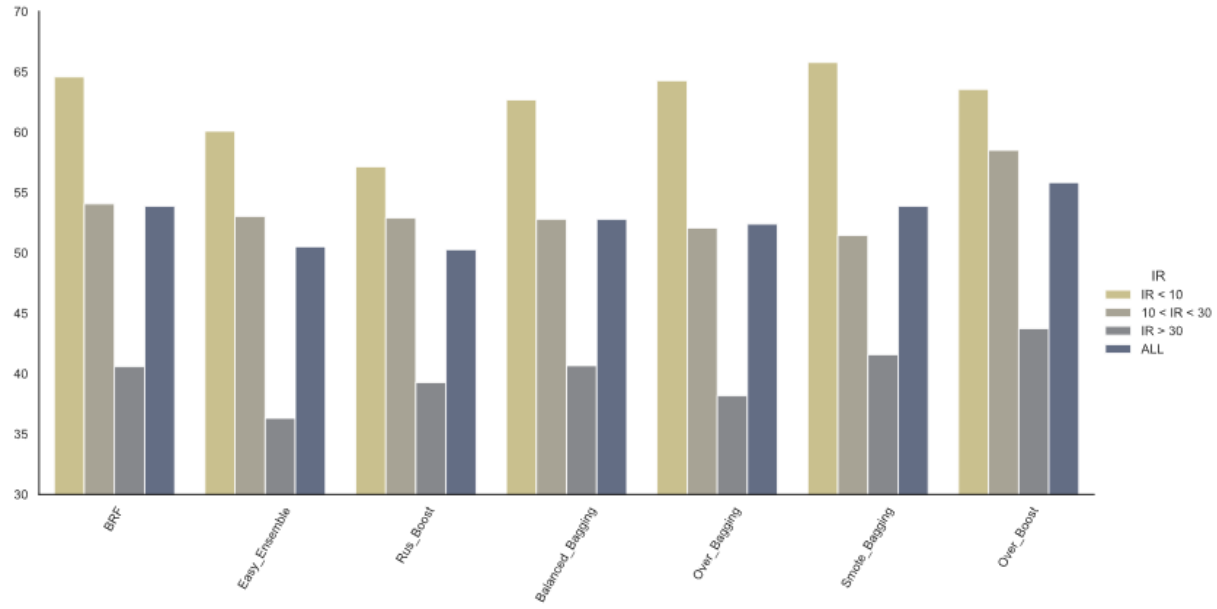

Fig. 7. Performance comparison of the ensemble algorithms

Table 1: Average mcc scores obtained from the ensemble algorithms (in percentage)

Method	Average MCC scores			
	IR < 10	10 ≤ IR < 30	IR ≥ 30	ALL
BRF	64.5724	54.0610	40.5768	53.8633
Balanced-Bagging (BB)	62.6736	52.8011	40.6634	52.7790
Over-Bagging	64.2763	52.0729	38.1633	52.3850
SMOTE-Bagging	**65.7740**	51.4461	41.5806	53.8749
RUSBoost	57.1440	52.8859	39.2745	50.2768

Over-Boost	63.5572	**58.5065**	**43.7280**	**55.8358**
Easy Ensemble	60.0863	53.0360	36.3134	50.5205

4.7 Performance analysis of cost-sensitive learning:

Cost-sensitive learning is an algorithm-level modification that increases the misclassification cost for the minority samples. As the penalty for misclassifying the minority class is greater, the algorithm will try not to misclassify any minority sample to reduce the overall cost. In this study, different costs were chosen for each dataset, and it was kept equal to the IR of that specific dataset. Cost-sensitive RF scored an average MCC score of 53.8% for all the datasets. For high imbalanced cases (IR>30), the MCC score was about 40%, which is comparable to other US or ensemble approaches. However, the OS techniques far outperform cost-sensitive learning. Cost-sensitive learning can be hybridized with oversampling techniques which is discussed in detail in the next chapter.

4.8 Performance analysis of GAN sampling techniques:

Generative adversarial network (GAN) is very popular for generating new data whenever the test samples are not sufficient. GAN can also be used in generating minority class samples in imbalanced scenarios. In this experiment, we used 4 types of GAN to produce new minority class samples. TVAE performed best whenever the imbalance ratio was less. However, with the increase of the imbalance ratio the performance of the GAN sampling techniques gradually declined. TVAE achieved an MCC score of 46% on average for the datasets. A detailed performance comparison of the GAN sampling technique is provided in **Fig. 8.**

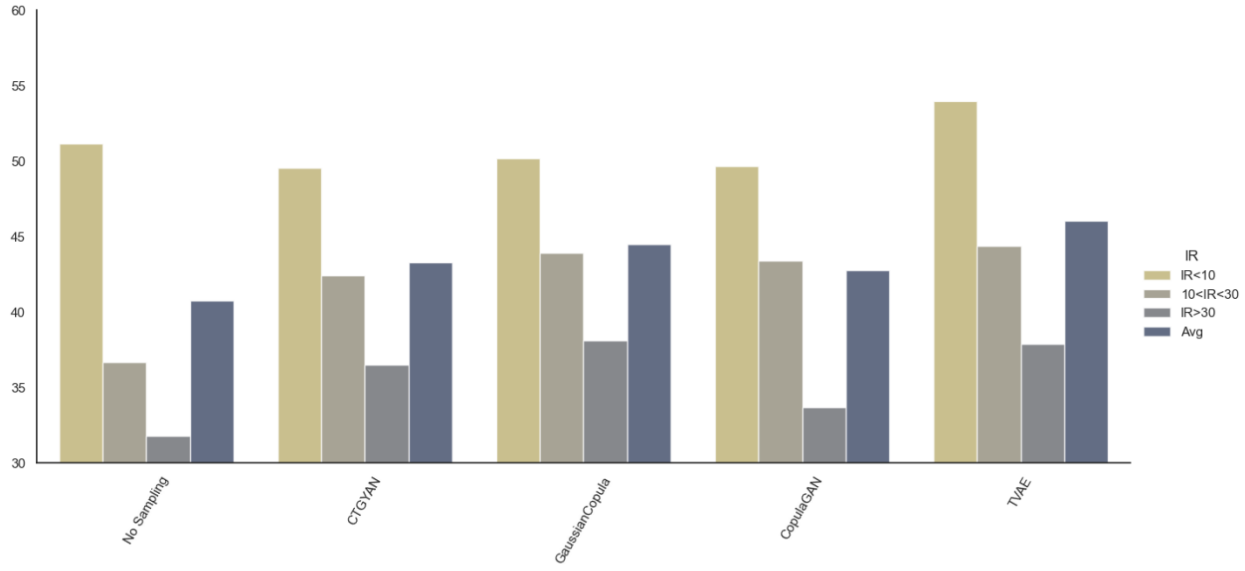

Fig. 8. Performance comparison of the GAN sampling techniques.

4.9 Performance comparison of all the techniques:

A total of 34 imbalanced data-handling methods were explored in this study. Among the experimented algorithms, LVQ-SMOTE, a variant of SMOTE, achieved the highest average MCC score of 59.14% for all datasets. Most of the techniques performed well when the data is less skewed (IR<10). However, as the IR increases, the performance of all techniques starts to decrease. This decline in performance is the highest in the under-sampling strategies. With larger imbalances in data, ensemble approaches also start to fall short in performance. In high imbalanced cases (IR>30), the performance from these approaches (US or ensemble or cost-sensitive learning) was significantly lower, making them unsuitable for such data. This fluctuation in performance is the lowest among the oversampling techniques.

Among all the techniques used, under-sampling approaches scored the lowest average MCC score. The best-performing under-sampling algorithm, NC, scored only 51.77%, which is significantly lower than the oversampling approaches. Based on the scores obtained from all experiments, it can be concluded that SMOTE variants are much more effective in handling datasets with high IR compared to other imbalanced data handling approaches. A comparison among the best-performing techniques in 5 different categories is provided in **Fig. 9.**

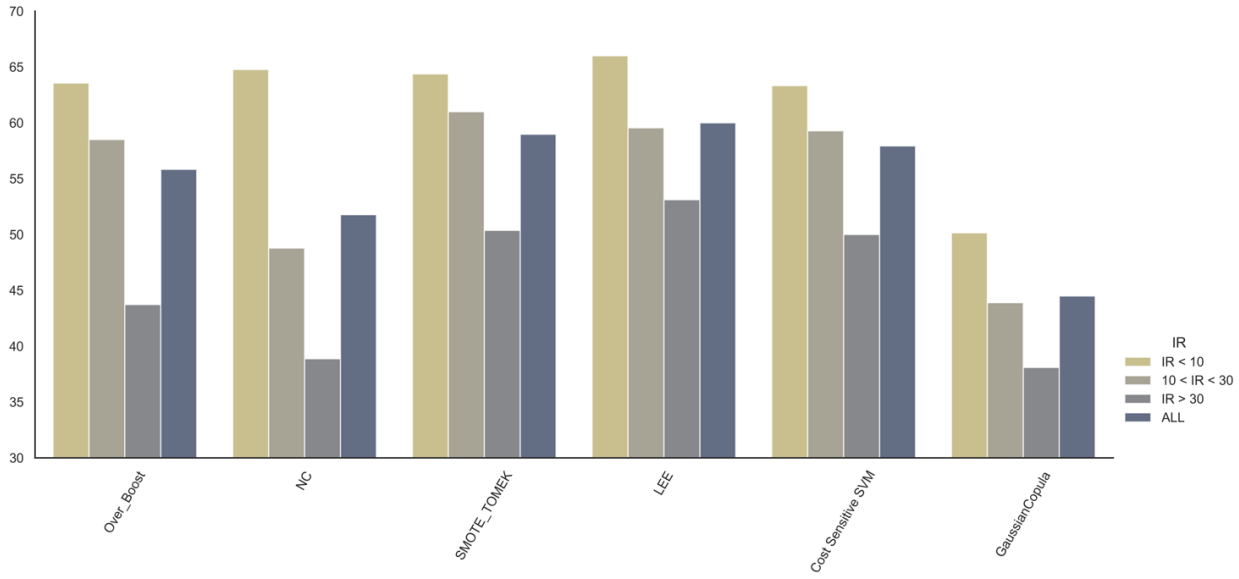

Fig. 9. Performance comparison of all the sampling techniques.

Chapter 5

Hybridization of sampling and cost-sensitive learning:

5.1 Overview:

Sampling is one of the most popular techniques to handle class imbalance problems. Creating new minority class samples or reducing the majority class data points helps to balance the dataset. This process eventually helps the machine learning classifier to learn the data pattern properly. Cost-sensitive learning is another popular method to handle the class imbalance issue. In cost-sensitive learning, the misclassification cost of the minority class is increased so that the classifier does not misclassify any minority data points to reduce overall cost. The common practice in the imbalanced data handling technique is to use either data sampling or cost-sensitive learning. How hybridization of these two techniques can produce better results than that of using them separately.

5.2 Proposed Methodology:

Data sampling and cost-sensitive learning are two distinct strategies used to address the issue of imbalanced learning. In data sampling, the dataset is balanced by either generating additional minority class samples or removing some majority class samples. Cost-sensitive learning, on the other hand, assigns a higher penalty to misclassifications of the minority class to reduce bias. Both strategies have their advantages and drawbacks. For example, eliminating too many majority class samples can result in the loss of significant information, while creating excessive minority class samples may reduce generalization ability. Cost-sensitive learning alone is not always effective because the data remains imbalanced, thus still susceptible to bias. Typically, these methods are used independently to manage imbalanced data. However, we propose a novel method that combines both approaches to achieve superior performance. The key idea is that a balanced integration of these methods can mitigate their individual limitations. This can be accomplished by initially reducing the imbalance ratio through sampling, followed by applying a cost-sensitive classifier with a moderate penalty for the minority class. Overproduction of minority class samples can result in synthetic samples that do not accurately represent the minority class. Conversely, removing too many majority class samples can lead to the loss of important data. A higher imbalance ratio necessitates a greater number of samples to be generated or removed, which increases the risk of overfitting. In our proposed method, the data remains somewhat imbalanced, thereby reducing the number of samples that need to be generated or removed. With a lower imbalance ratio, assigning a moderate penalty can sufficiently counteract the bias toward the majority class. Therefore, an optimal balance between these two approaches can lead to improved performance. Achieving this balance requires fine-tuning two parameters: the sampling ratio (α) and the weight factor (ω). The step-by-step procedure to develop such a hybrid model is presented below –

i. At first, we found out the imbalance ratio of the dataset. Imbalance ratio is very crucial in the later sampling part. The amount of sampling mostly depends on the ratio between presence of positive and negative class examples.

$$IR = \frac{number\ of\ samples\ in\ the\ majority\ class}{number\ of\ samples\ in\ the\ minority\ class}$$

ii. The processed dataset is then split into training and testing folds, with only the training fold undergoing resampling using the SMOTE algorithm to prevent data leakage. The model performance is evaluated on the testing fold. Smote has more than hundreds of variants. Various SMOTE algorithm variations can also be employed for resampling the training set.

iii. iii. The data was retained somewhat skewed during the resampling process rather than being completely balanced. The SMOTE implementation of the "Imbalanced-learn" package may also be adjusted with the α parameter. The right value for α cannot be predicted in advance because it relies on the facts. To determine which α would work best for the MI dataset, a grid-search technique was used. XGBoost is used as the base classifier. Parameter tuning can be done using the 'scale_pos_weight' parameter, where the sampling ratio and weight can be adjusted.

5.3 Result and Discussion:

The above-mentioned hybridized approach is used in a myocardial infarction dataset collected from the UCI machine learning repository [25]. External validation is often done to ensure the robustness of the predictive modeling. In this experiment, we also validated the modeling with 36 other imbalanced datasets, which were collected from UCI and KEEL repositories. We adopted a 10-fold repeated stratified cross-validation method to avoid any potential data leaking. For evaluation purposes, we mainly used ROC-AUC, Geometric mean score and the MCC scores to compare the performance of the classifiers.

5.3.1 Implementation and performance evaluation of the proposed strategy

The presence of class imbalance in the data results in an accuracy measure biased toward the majority class, as indicated by high specificity but low sensitivity scores in the algorithms. To mitigate this performance bias, this study employs a combination of data sampling and cost-sensitive learning techniques. Initially, the SMOTE algorithm is used to generate synthetic minority class samples, partially balancing the data. The XGBoost classifier is then trained on this resampled data. Rather than using a standard error-driven evaluation process, a cost-driven approach is adopted, modifying the algorithm to act as a cost-sensitive classifier by assigning higher costs to the minority class samples. This adjustment penalizes misclassifications of minority class instances more severely, thereby shifting the decision boundary away from the majority class and reducing bias.

A key challenge in resampling is determining the appropriate sampling ratio to control the number of synthetic samples generated, as too many can lead to overfitting. Similarly, selecting an appropriate weight for the minority class in cost-sensitive learning is crucial, as too small a weight fails to reduce majority class bias, while too large a weight can cause bias toward the minority class. Thus, finding the optimal weight requires careful consideration and cannot be predetermined. A grid-search approach is employed to determine the best values for the sampling ratio and class weights. A sampling ratio of 0.9 for the SMOTE algorithm and a weight value of 5 for the minority class yielded the best results. The results, presented in Table 2, demonstrate that combining these approaches enhances prediction performance and generalization by optimizing the number of synthetic samples generated and the class weights assigned.

Table 2: Performance comaparison of different sampling techniques with the XGBoost classifier (in percentage)

Metrics	SMOTE	ADASYN	RUS	Tomek-link	ENN	Weighted XGBoost	**Proposed**
Accuracy	91.1431	91.4009	81.9014	90.8850	90.4317	91.5311	**91.9843**
Sensitivity	56.3922	57.6688	**77.5485**	52.9694	58.1498	59.8520	65.0323
Specificity	97.2623	97.3384	82.6616	**97.5665**	96.1216	97.1103	96.73
G-mean	73.7770	74.6062	**79.8933**	71.7280	74.6397	76.0671	79.0185
ROC-AUC	76.8272	77.5036	80.1050	75.2680	77.1357	78.4811	**80.8812**
Precision	78.5531	78.9781	44.0476	**79.4797**	73.2514	78.6094	77.9264
MCC	61.6813	62.7711	48.7644	60.0629	59.7997	63.8604	**66.5321**

5.3.2 Performance comparison of the proposed approach with other approaches used in imbalanced learning

The proposed approach was evaluated against several popular techniques for imbalanced learning, including ADASYN, Random Under Sampling (RUS), Edited Nearest Neighbors (ENN), and Tomek-links based under-sampling. Additionally, comparisons were made using SMOTE and weighted XGBoost classifiers independently, with XGBoost serving as the base algorithm for all methods. ADASYN adaptively generates synthetic samples based on density distribution, creating more samples for harder-to-learn minority class instances. RUS balances the data by randomly removing majority class samples, while ENN uses the KNN algorithm to selectively remove misclassified majority class samples. Tomek-links under-sampling removes noisy boundary samples by identifying and eliminating majority class instances in Tomek-link pairs. Performance

measures from these techniques are detailed in Table 2, and Fig. 10 illustrates a comparative analysis.

As shown in Table 2, our proposed approach outperforms other techniques in metrics such as ROC-AUC, MCC, and overall accuracy. Although RUS achieves the highest sensitivity (77.55%), it suffers from the lowest specificity (82.66%), MCC (48.76%), and precision (44.05%), indicating a bias toward the minority class due to excessive removal of majority class instances. Conversely, Tomek-links achieves the highest specificity (97.57%) but the lowest sensitivity (52.97%). Composite metrics like ROC-AUC and MCC are more appropriate for assessing imbalanced learning performance, with our approach achieving the highest ROC-AUC (80.88%) and MCC (66.53%). Our hybrid approach also shows significant performance improvements compared to using SMOTE or weighted XGBoost independently, with MCC scores of 66.53% versus 61.68% and 63.86%, respectively.

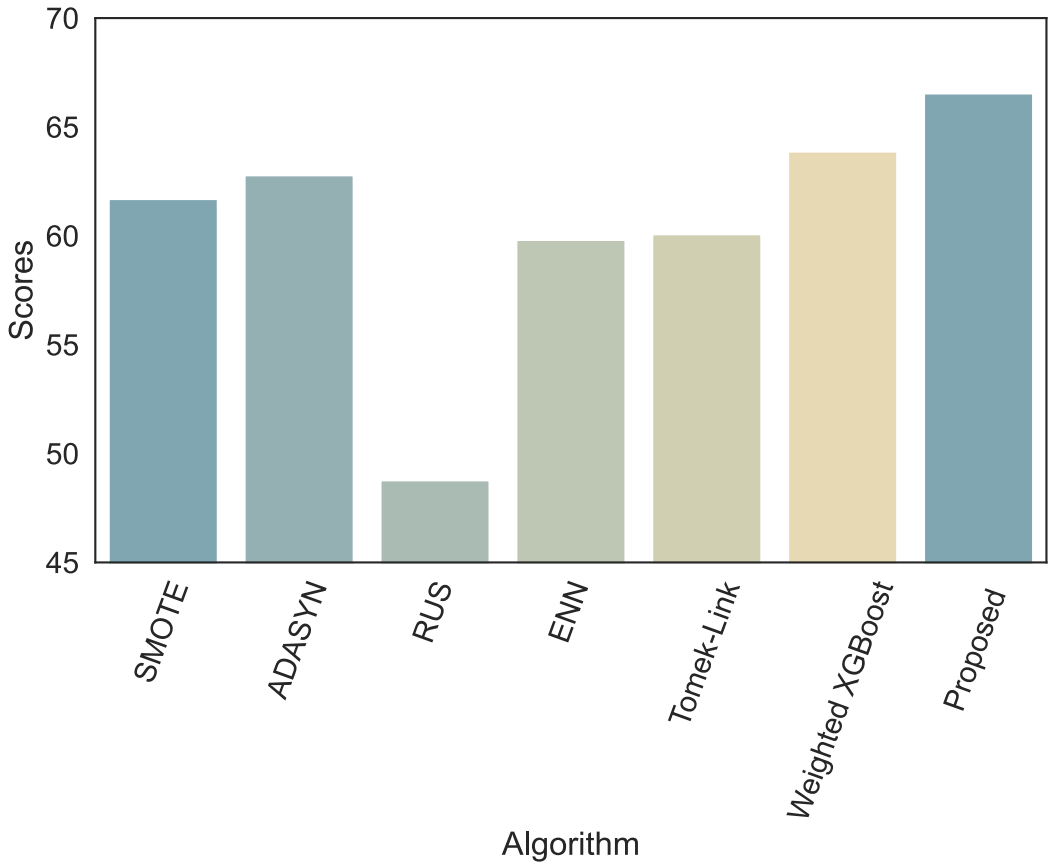

Fig. 10. Performance analysis of other sampling methods in MCC

5.3.3 Performance analysis of the proposed approach in other imbalanced datasets

For external validation, we evaluated the performance of our proposed approach in 36 other imbalanced datasets collected from the KEEL repository [26]. A summary of the datasets is provided in Table 3. The performance metrics obtained were compared with other widely used techniques in imbalanced learning, using the XGBoost classifier as the base algorithm and employing a 10-fold stratified cross-validation strategy for model evaluation. The data was normalized before training, and sampling was performed only on the training folds to prevent data leakage, with performance calculated on the testing folds. The average results from the 10 different testing folds were considered, following a pipeline similar to that illustrated in Figure 1. Detailed performance measures for each dataset are provided in the associated GitHub repository, including the average scores across 36 imbalanced datasets, as shown in Figure 11.

Figure 11 demonstrates that the proposed approach achieves the highest average ROC-AUC, g-mean, and MCC scores among all techniques. The hybrid method significantly outperforms the independent use of SMOTE or weighted classifiers. The performance of SMOTE and ADASYN was comparable, while the RUS algorithm achieved the highest sensitivity but had the lowest specificity and MCC scores, indicating an imbalanced prediction performance. Our proposed approach achieved superior scores, validating its effectiveness in addressing class imbalance.

Table 3: A summary of imbalanced datasets used for external validation

	Name	No. of samples	No. of positive samples	Number of features	Imbalance Ratio
1	wisconsin	683	239	9	1.86
2	vehicle2	846	218	18	2.88
3	vehicle1	846	217	18	2.9
4	vehicle3	846	212	18	2.99
5	vehicle0	846	199	18	3.25
6	new_thyroid1	215	35	5	5.14
7	ecoli2	336	52	7	5.46
8	glass6	214	29	9	6.38
9	yeast3	1484	163	10	8.1
10	ecoli3	336	35	7	8.6

11	yeast-2_vs_4	514	51	8	9.08
12	yeast-0-2-5-6_vs_3-7-8-9	1004	99	10	9.14
13	vowel	988	90	13	9.98
14	led7digit-0-2-4-6-7-89_vs_1	443	37	7	10.97
15	glass2	214	17	9	11.59
16	ecoli-0-1-4-7_vs_5-6	332	25	6	12.28
17	glass4	214	13	9	15.46
18	ecoli4	336	20	7	15.8
19	page-blocks-1-3_vs_4	472	28	10	15.86
20	abalone	731	42	8	16.4
21	yeast-1-4-5-8_vs_7	693	30	10	22.1
22	yeast	1484	51	10	28.1
23	yeast-1-2-8-9_vs_7	947	30	10	30.57
24	yeast5	1484	44	10	32.73
25	winequality-red-8_vs_6	656	18	11	35.44
26	abalone_17_vs_7_8_9_10	2338	58	8	39.31
27	winequality-white-3_vs_7	900	20	11	44
28	winequality-red-8_vs_6-7	855	18	11	46.5
29	Kddcup land_vs_portsweep	1061	21	40	49.52
30	abalone-19_vs_10-11-12-13	1622	32	8	49.69

31	winequality-white-3-9_vs_5	1482	25	11	58.28
32	poker-8-9_vs_6	1485	25	25	58.4
33	winequality-red-3_vs_5	691	10	11	68.1
34	kddcup-land_vs_satan	1610	21	30	75.67
35	poker-8-9_vs_5	2075	25	25	82
36	poker-8_vs_6	1477	17	25	85.88

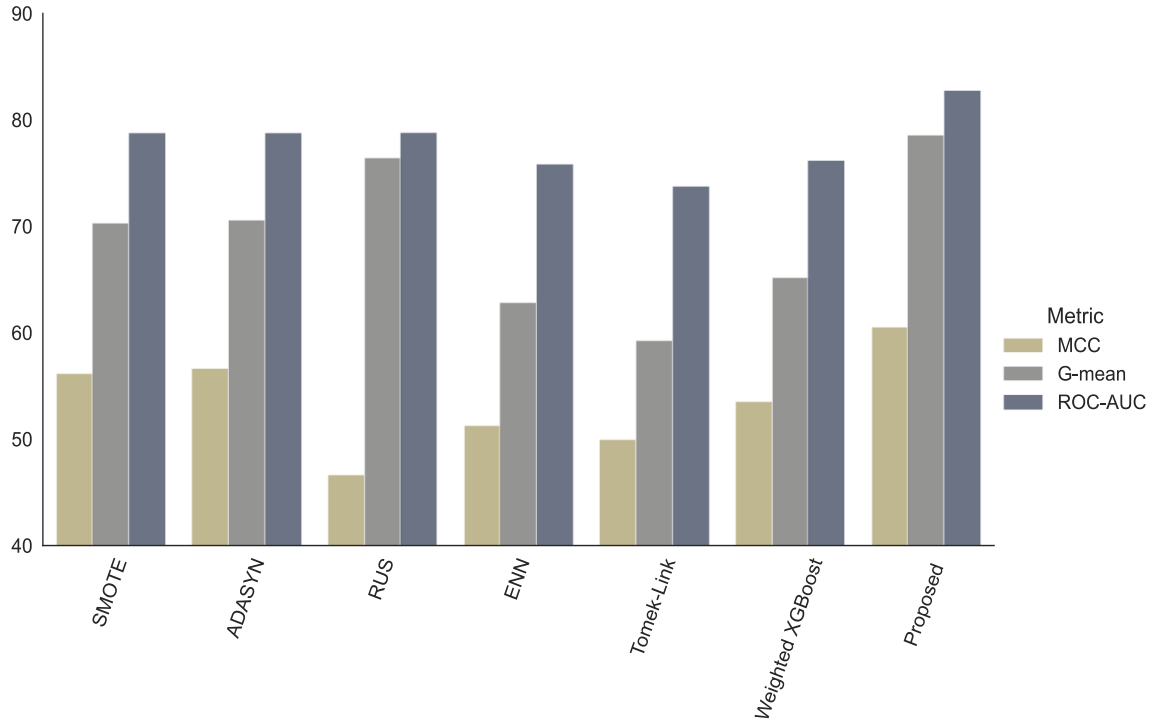

Fig. 11. Performance comparison with other approaches on 36 imbalanced datasets

Table 4: Average of the performance measures obtained from different approaches on 36 imbalanced datasets

Metrics	SMOTE	ADASYN	RUS	Tomek-link	ENN	Weighted XGBoost	**Proposed**
Accuracy	93.57	**94.49**	78.09	94.25	93.64	94	93.3
Sensitivity	62.34	62.5	**79.72**	51.06	56.79	56.44	71.22
Specificity	95.22	95.09	77.87	**96.46**	94.92	95.91	94.35
G-mean	70.3	70.54	76.45	59.22	62.76	65.18	**78.52**
ROC-AUC	78.78	78.79	78.8	73.76	75.86	76.17	**82.78**
Precision	59.11	59.93	39.64	57.64	55.07	59.61	**60.09**
MCC	56.14	56.57	46.62	49.94	51.24	53.49	**60.55**

Chapter 6

Conclusion and Future Works:

In this thesis, we provided an extensive analysis of the sampling technique followed by a critical review. Rigorous experimentation was conducted to analyze the effectiveness of the algorithms in a broad range of imbalanced scenarios. The performance of these techniques is also compared to find the most appropriate approach in different imbalanced settings.

Overall, OS techniques performed better than the other approaches. Especially in highly imbalanced scenarios, they have no alternatives. While the SMOTE variants showed similar performance, algorithms such as LEE, SMOBD, SMOTE-IPF, and LVQ-SMOTE provided comparatively better results. Hybrid sampling techniques also showed comparable performance. Other hybridizations between algorithms should be explored for better results. The under-sampling and ensemble approaches provided quite a good result in lower imbalances. However, as the IR rises, performance decreases drastically. The boosting ensemble performed better than bagging-based ensembles. OS-based ensembles performed better than US-based ensembles. However, in the datasets with high IR, all the ensembles failed to produce good results. Using SMOTE or its variants on the entire dataset produces better results. As compared to using data resampling techniques, we also tested the cost-sensitive learning approach. While they cannot outperform the best-performing sampling approaches, they still provide quite a good performance in lower imbalances. Considering the simplicity of the technique, it can be considered a good alternative when the dataset size is very large, and sampling becomes time-consuming.

We also suggested the hybridization of oversampling and cost-sensitive learning. This technique performed quite well and performed the best among the sampling techniques when tested on 36 imbalanced datasets. This proves the effectiveness of hybridization in handling imbalanced datasets.

In the future, we would like to extend our work to multiclass classification. Besides, we are also expanding our research to handle other factors e.g., class overlap, multiple minority class clusters, and multiple majority class issues that make learning from an imbalanced dataset difficult.

Chapter 7

Demonstration of Outcome Based Education (OBE)

7.1 Introduction

We explore a range of sampling strategies designed to mitigate imbalanced data issues, specifically concentrating on class overlap issues. Rebalancing datasets through adjustments to minority and majority class distributions is made possible in large part by sampling techniques. We carefully assess a variety of sample strategies, taking into account imbalance ratio (IR), other important metrics, and how well they work in situations when classes overlap. We hope to shed light on these methods' advantages and disadvantages by evaluating their performance on a variety of metrics, which will help us better understand how these methods affect model performance in scenarios where class distributions aren't equal in the real world. Our investigation involves investigating new approaches by combining several sample techniques to develop more reliable and efficient ways to deal with imbalanced data. We pay close attention to overlapping classes and offer solutions to resolve this problem. Our goal is to address the challenges of managing several minority classes in imbalanced data sets by utilizing R and Python technologies. Our objective is to provide people with the information and resources they need to cope with imbalanced data using our thorough examination of sampling strategies and application of R and Python technologies. We think we may better handle the issues presented by imbalanced datasets and enhance overall model performance by learning more about sampling techniques and investigating novel approaches.

Addressing COs and POs

Table 5: The following table shows the COs for EEE 4700/4800.

COs	CO Statement	POs
CO1	Identify a contemporary real-life problem related to electrical and electronic engineering by reviewing and analyzing existing research works.	PO2
CO2	Determine functional requirements of the problem considering feasibility and efficiency through analysis and synthesis of information.	PO4
CO3	Select a suitable solution and determine its method considering professional ethics, codes, and standards.	PO8
CO4	Adopt modern engineering resources and tools for the solution of the problem.	PO5
CO5	Prepare a management plan and budgetary implications for the solution of the problem.	PO11
CO6	Analyze the impact of the proposed solution on health, safety, culture, and society.	PO6
CO7	Analyze the impact of the proposed solution on the environment and sustainability.	PO7
CO8	Develop a viable solution considering health, safety, cultural, societal, and environmental aspects.	PO3
CO9	Work effectively as an individual and as a team member for the accomplishment of the solution.	PO9
CO10	Prepare various technical reports, design documentation, and deliver effective presentations for demonstration of the solution.	PO10
CO11	Recognize the need for continuing education and participation in professional societies and meetings.	PO12

Table 6: The following table shows the aspects addressed for certain Program Outcomes (POs) addressed in EEE 4700/4800 for Project and Thesis.

	Statement	Different Aspects	Put Tick (√)
PO3	**Design/development of solutions**: Design solutions for complex electrical and electronic engineering problems and design systems, components, or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.	Public health	√
		Safety	√
		Cultural	
		Societal	
		Environmental	
PO4	**Investigation:** Conduct investigations of complex electrical and electronic engineering problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valid conclusions.	Design of experiments	√
		Analysis and interpretation of data	√
		Synthesis of information	√
PO6	**The engineer and society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to professional engineering practice and solutions to complex electrical and electronic engineering problems.	Societal	
		Health	√
		Safety	√
		Legal	
		Cultural	
PO7	**Environment and sustainability:** Understand and evaluate the sustainability and impact of professional engineering work in the solution of complex electrical and electronic engineering problems in societal and environmental contexts.	Societal	
		Environmental	
PO8		Religious values	

	Ethics: Apply ethical principles embedded with religious values, professional ethics and responsibilities, and norms of electrical and electronic engineering practice.	Professional ethics and responsibilities	√
		Norms	√
PO9	**Individual work and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams and in multi-disciplinary settings.	Individual	√
		Teamwork	√
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.	Comprehend and write effective reports	√
		Design documentation	√
		Make effective presentations	√
		Give and receive clear instructions	√
PO11	**Project management and finance:** Demonstrate knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	Engineering management principles	
		Economic decision-making	
		Manage projects	√
		Multidisciplinary environments	√

Table 7: The following table explains or justifies how the COs and corresponding POs have been addressed in EEE 4700/4800 (Project and Thesis).

COs	POs	Explanation/Justification
CO1	PO2	
CO2	PO4	Most real-life datasets are imbalanced. So, solving the Imbalanced problem is necessary.
CO3	PO8	Ethical considerations were taken into account when designing the experiments. No of the methods used in this project has a conflict with professional norms.
CO4	PO5	We used a combination of cutting-edge software libraries and programming languages designed for data analysis and machine learning to adopt contemporary technical resources and techniques for the problem's solution
CO5	PO11	To prepare a management plan and budgetary implications for the solution of the problem, we meticulously outlined the project phases, and resource requirements. We established reporting schedules, meeting frequencies, and communication channels to ensure effective coordination among team members and stakeholders.
CO6	PO6	
CO7	PO7	
CO8	PO3	
CO9	PO9	Throughout the project, we emphasized the importance of working effectively both as individual contributors and as a cohesive team. Individually, team members demonstrated initiative by continuously enhancing their skills and knowledge in areas relevant to the project, such as imbalanced data analysis techniques and engineering principles. Each member took responsibility for specific tasks, ensuring their timely completion and contributing to the overall project progress.
CO10	PO10	Throughout the project, we produced several technical reports, design documentation, and powerful presentations to illustrate our approach.
CO11	PO12	

Addressing Knowledge Profiles (K3 – K8)

Table 8: The following table shows the Knowledge Profiles (K3 – K8) addressed in EEE 4700/4800.

K	Knowledge Profile (Attribute)	Put Tick (√)
K1	A systematic, theory-based understanding of the natural sciences applicable to the discipline	√
K2	Conceptually based mathematics, numerical analysis, statistics and the formal aspects of computer and information science to support analysis and modeling applicable to the discipline	√
K3	A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline	√
K4	Engineering specialist knowledge provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline	√
K5	Knowledge that supports engineering design in a practice area	√
K6	Knowledge of engineering practice (technology) in the practice areas in the engineering discipline	√
K7	Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability	
K8	Engagement with selected knowledge in the research literature of the discipline	√

Table 9: The following table explains or justifies how the Knowledge Profiles (K3 – K8) have been addressed in EEE 4700/4800.

K	Explanation/Justification
K1	Analysis of different data sets from different ranges of topics is selective based on their Imbalance ratio. The theory-based understanding of different algorithms is crucial here for different engineering applications.
K2	Mathematical concepts such as various machine learning models and their underlying principles, numerical analysis, statistics for assessment, and the formal parts of engineering enable modeling and analysis relevant to the field.
K3	The exploration covered theoretical aspects of imbalanced data analysis, including statistical learning theory concepts, sampling methods, and performance metrics for evaluating imbalanced datasets.
K4	The topics delved into the theoretical frameworks of imbalanced data analysis, exploring statistical foundations and techniques, alongside strategies for mitigating algorithmic bias and fairness metrics.
K5	Different system that was specifically suited to the difficulties encountered in engineering by utilizing our knowledge of sampling strategies, performance evaluation measures, and imbalanced data processing methodologies are explored.
K6	Demonstration of a strong understanding of engineering practice by focusing on the practical application of imbalanced data analysis techniques to real-world problems. Techniques using relevant software tools, such as Python libraries like scikit-learn are examples.
K7	
K8	A thorough literature review and critical analysis was conducted. In engaging with specific research areas, we prioritized research that specifically tackled imbalanced data challenges in engineering. We identified relevant techniques that had been successfully applied in areas similar to our chosen application.

Addressing Attributes of Ranges of Complex Engineering Problem Solving (P1 – P7)

Table 10: The following table shows the attributes of ranges of Complex Engineering Problem Solving (P1 – P7) addressed in EEE 4700/4800.

P	**Range of Complex Engineering Problem Solving**	**Put Tick**
Attribute	Complex Engineering Problems have characteristics P1 and some or all of P2 to P7:	**(√)**
Depth of knowledge required	**P1:** Cannot be resolved without in-depth engineering knowledge at the level of one or more of K3, K4, K5, K6, or K8 which allows a fundamentals-based, first principles analytical approach	√
Range of conflicting requirements	**P2:** Involve wide-ranging or conflicting technical, engineering, and other issues	√
Depth of analysis required	**P3:** Have no obvious solution and require abstract thinking, and originality in analysis to formulate suitable models	√
Familiarity of issues	**P4:** Involve infrequently encountered issues	√
Extent of applicable codes	**P5:** Are outside problems encompassed by standards and codes of practice for professional engineering	√
Extent of stakeholder involvement and conflicting requirements	**P6:** Involve diverse groups of stakeholders with widely varying needs	√
Interdependence	**P7:** Are high-level problems including many parts or sub-problems	√

Table 11: The following table explains or justifies how the attributes of ranges of Complex Engineering Problem Solving (P1 – P7) have been addressed in EEE 4700/4800 (Project and Thesis).

P	Explanation/Justification
P1	We recognized the importance of knowledge areas K3, K4, K5, and K8 in imbalanced data analysis, as they went beyond the application of basic engineering principles. These areas were crucial because they required deeper theoretical understanding. For example, we need to know how a SVM model works to apply it for a specific type of dataset.
P2	We recognized the complexity of handling imbalanced data analysis in different engineering branches like data mining, machine learning, and medical data handling as highlighted by the wide-ranging and potentially conflicting issues involved. We navigated through these challenges by carefully considering each aspect, ensuring a comprehensive approach to addressing imbalanced data analysis in these fields.
P3	We acknowledged that imbalanced data analysis presented a complex challenge due to the absence of a single, straightforward solution. A solution for a dataset of IR of 20 will not work on a dataset with 10 IR value. So, different approaches, and their hybrid are examined to determine the relevance for a specific case.
P4	We encountered infrequently encountered issues when applying imbalanced data analysis within specific engineering domains. These cases are often rare and often a single problem specific. An Intrusion Detection System (IDS) data and its features including IR will not match with the Myocardial Infraction dataset, and these are very different from other datasets encountered in this field.
P5	We observed that standard engineering codes and practices in our work like being aware of plagiarism and following directives for using the works of other researchers.
P6	Throughout the project, we can ensure the involvement of diverse groups of stakeholders with widely varying needs. We recognize that different stakeholders may have different requirements and priorities when it comes to the performance of the solution. For example, some stakeholders may require higher accuracy and precision in the models generated by the solution, while others may prioritize scalability or interpretability.
P7	We encountered a complex challenge in imbalanced data analysis for engineering because it involved numerous interconnected components, we found that effectively addressing

	imbalanced data required us to tackle various sub-problems at each stage of the analysis process. To fine-tune a model different parameters need to be considered to make the model more accurate and it results in a very fruitful way if done right but it is just a sub-step.

Addressing Attributes of Ranges of Complex Engineering Activities (A1 – A5)

Table 12: The following table shows the attributes of ranges of Complex Engineering Activities (A1 – A5) addressed in EEE 4700/4800 (Project and Thesis).

A	**Range of Complex Engineering Activities**	**Put Tick (√)**
Attribute	Complex activities mean (engineering) activities or projects that have some or all of the following characteristics:	
Range of resources	**A1:** Involve the use of diverse resources (and for this purpose resources include people, money, equipment, materials, information, and technologies)	√
Level of interaction	**A2:** Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering, or other issues	√
Innovation	**A3:** Involve creative use of engineering principles and research-based knowledge in novel ways	√
Consequences for society and the environment	**A4:** Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation	√
Familiarity	**A5:** Can extend beyond previous experiences by applying principles-based approaches	√

Table 13: The following table explains or justifies how the attributes of ranges of Complex Engineering Activities (A1 – A5) have been addressed in EEE 4700/4800 (Project and Thesis).

A	Explanation/Justification
A1	Various resources have been utilized in our study, including the utilization of various data repositories and the incorporation of insights from numerous research works.
A2	Resolutions have been implemented in numerous instances within our work. There were occasions when our results did not align closely with our anticipated accuracy or outcomes. During such times, we undertook specific resolutions to enhance performance. Given the challenge of working with imbalanced data, these adjustments were frequently necessary to improve our results.
A3	While we may not have introduced a new algorithm, in certain instances, we achieved higher accuracy compared to some existing research papers.
A4	Our work holds the potential for innovative impact as various branches of data mining are interrelated. Consider the scenario of fraud detection, where the dataset predominantly consists of normal cases with a minimal fraction of fraud cases, resulting in an imbalanced dataset. Since our work focuses on addressing imbalances in datasets, individuals engaged in fraud detection could benefit from our methodologies. This exemplifies how our project can indirectly contribute to societal and environmental benefits.
A5	Indeed, we can extend beyond past experiences by implementing principles-based approaches, which is precisely what we are currently undertaking.

References:

[1] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Learning from Imbalanced Data Sets," *Learning from Imbalanced Data Sets*, 2018, doi: 10.1007/978-3-319-98074-4.

[2] M. Dudjak and G. Martinović, "An empirical study of data intrinsic characteristics that make learning from imbalanced data difficult," *Expert Syst Appl*, vol. 182, p. 115297, Nov. 2021, doi: 10.1016/J.ESWA.2021.115297.

[3] H. Y. J. Kang, E. Batbaatar, D. W. Choi, K. S. Choi, M. Ko, and K. S. Ryu, "Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy," *JMIR Med Inform*, vol. 11, no. 1, Jan. 2023, doi: 10.2196/47859.

[4] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowl Based Syst*, vol. 212, p. 106631, Jan. 2021, doi: 10.1016/J.KNOSYS.2020.106631.

[5] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-NN performance in a challenging scenario of imbalance and overlapping," *Pattern Analysis and Applications*, vol. 11, no. 3–4, pp. 269–280, Sep. 2008, doi: 10.1007/S10044-007-0087-5/FIGURES/7.

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst Appl*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/J.ESWA.2016.12.035.

[7] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, pp. 243–248, Apr. 2020, doi: 10.1109/ICICS49469.2020.239556.

[8] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/JAIR.1.11192.

[9] A. Newaz, M. S. Mohosheu, and M. A. Al Noman, "Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques," *Inform Med Unlocked*, vol. 42, p. 101361, Jan. 2023, doi: 10.1016/J.IMU.2023.101361.

[10] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random Balance: Ensembles of variable priors classifiers for imbalanced data," *Knowl Based Syst*, vol. 85, pp. 96–111, Sep. 2015, doi: 10.1016/J.KNOSYS.2015.04.022.

[11] H. J. Kim, N. O. Jo, and K. S. Shin, "Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction," *Expert Syst Appl*, vol. 59, pp. 226–234, Oct. 2016, doi: 10.1016/J.ESWA.2016.04.027.

[12] G. Kovács, "Smote-variants: A python implementation of 85 minority oversampling techniques," *Neurocomputing*, vol. 366, pp. 352–354, Nov. 2019, doi: 10.1016/J.NEUCOM.2019.06.100.

[13] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop Oversampling for Class Imbalance Learning: A Review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022, doi: 10.1109/ACCESS.2022.3169512.

[14] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” *J Biomed Inform*, vol. 107, p. 103465, Jul. 2020, doi: 10.1016/J.JBI.2020.103465.

[15] A. Newaz, S. Hassan, F. Shahriyar Haq, and C. Author, “An Empirical Analysis of the Efficacy of Different Sampling Techniques for Imbalanced Classification,” Aug. 2022, Accessed: Mar. 15, 2024. [Online]. Available: https://arxiv.org/abs/2208.11852v1

[16] J. J. Rodríguez, J. F. Díez-Pastor, Á. Arnaiz-González, and L. I. Kuncheva, “Random Balance ensembles for multiclass imbalance learning,” *Knowl Based Syst*, vol. 193, p. 105434, Apr. 2020, doi: 10.1016/J.KNOSYS.2019.105434.

[17] V. H. Alves Ribeiro and G. Reynoso-Meza, “Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets,” *Expert Syst Appl*, vol. 147, p. 113232, Jun. 2020, doi: 10.1016/J.ESWA.2020.113232.

[18] K. Yang *et al.*, “Hybrid Classifier Ensemble for Imbalanced Data,” *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 4, pp. 1387–1400, Apr. 2020, doi: 10.1109/TNNLS.2019.2920246.

[19] A. Anaissi, P. J. Kennedy, M. Goyal, and D. R. Catchpoole, “A balanced iterative random forest for gene selection from microarray data,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–10, Aug. 2013, doi: 10.1186/1471-2105-14-261/TABLES/4.

[20] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–16, Mar. 2013, doi: 10.1186/1471-2105-14-106/FIGURES/7.

[21] G. Kovács, “An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets,” *Appl Soft Comput*, vol. 83, p. 105662, Oct. 2019, doi: 10.1016/J.ASOC.2019.105662.

[22] H. Nugroho, K. Wikantika, S. Bijaksana, and A. Saepuloh, “Handling imbalanced data in supervised machine learning for lithological mapping using remote sensing and airborne geophysical data,” *Open Geosciences*, vol. 15, no. 1, Jan. 2023, doi: 10.1515/GEO-2022-0487/DOWNLOADASSET/SUPPL/S2A_MSIL2A_20190109T011721_N0211_R088_T53MPR_20190109T032340.ZIP.

[23] R. C. Prati, G. E. A. P. A. Batista, and D. F. Silva, “Class imbalance revisited,” *Knowl Inf Syst*, vol. 45, no. 1, pp. 247–270, Oct. 2015, doi: 10.1007/S10115-014-0794-3.

[24] A. N. Tarekegn, M. Giacobini, and K. Michalak, “A review of methods for imbalanced multi-label classification,” *Pattern Recognit*, vol. 118, p. 107965, Oct. 2021, doi: 10.1016/J.PATCOG.2021.107965.

[25] “Myocardial infarction complications - UCI Machine Learning Repository.” Accessed: Jul. 14, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications

[26] J. Alcalá-Fdez *et al.*, “KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework,” vol. 17, pp. 255–287, 2011, Accessed: Mar. 15, 2024. [Online]. Available: http://the-data-mine.com/bin/view/Software