

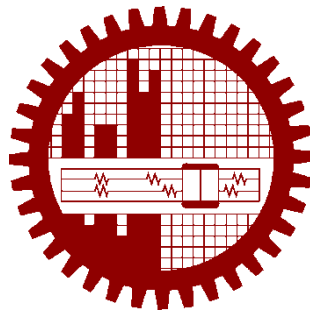
A MULTI-LAYER HYBRID BALANCING TECHNIQUE TO REMOVE DATA IMBALANCE

by

Muhammad Tanveer Islam

0417312025

MASTER OF SCIENCE
IN
INFORMATION AND COMMUNICATION TECHNOLOGY



Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

June, 2021

This thesis titled, “A MULTI-LAYER HYBRID BALANCING TECHNIQUE TO REMOVE DATA IMBALANCE”, submitted by Muhammad Tanveer Islam, Roll No.: 0417312025, Session: April 2017, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Information and Communication Technology on 15th June, 2021.

BOARD OF EXAMINERS



Dr. Hossen Asiful Mustafa
Associate Professor
IICT, BUET, Dhaka

Chairman
(Supervisor)



Dr. Md. Rubaiyat Hossain Mondal
Director and Professor
IICT, BUET, Dhaka

Member
(Ex-Officio)



Dr. Md. Liakot Ali
Professor
IICT, BUET, Dhaka

Member



Dr. Mohammad Shorif Uddin
Professor
Dept. of CSE, Jahangirnagar University, Dhaka

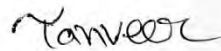
Member
(External)

Candidate's Declaration

This is to certify that the work presented in this thesis entitled, "A MULTI-LAYER HYBRID BALANCING TECHNIQUE TO REMOVE DATA IMBALANCE", is the outcome of the research carried out by Muhammad Tanveer Islam under the supervision of Dr. Hossen Asiful Mustafa, Associate Professor, Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

Signature of the Candidate



Muhammad Tanveer Islam
0417312025

Dedication

I would like to dedicate this thesis to Almighty God, Allah SWT, for giving me the opportunity, determination, and strength to do this research. I would also like to dedicate this work to my beloved parents for their love and endless support.

Contents

Certification	i
Candidate’s Declaration	ii
Dedication	iii
List of Figures	vii
List of Tables	ix
List of Abbreviations	xi
Acknowledgement	xii
Abstract	xiii
1 Introduction	1
Introduction	1
Machine Learning and Data Modeling	2
Imbalance Dataset	3
Oversampling.....	4
Under-Sampling.....	4
Motivations	5
Objectives of the Thesis	6
Objectives with specific aims	6
Possible outcome	6
Thesis Outline	6
2 Literature Review	8
3 Background	14
Adaptive Synthetic Sampling Approach (ADASYN).....	14
SVM-SMOTE	16
SMOTE + Edit Nearest Neighbor (SMOTE + ENN)	17

Random Forest (RF).....	17
Artificial Neural Network (ANN).....	20
Summary	22
4 Proposed Approach	23
Multi Layer Hybrid Data Balancing Scheme (MLH)	23
Initial Layer	23
Process Engineering Layer	25
Final Output Layer.....	27
Summary	27
5 Dataset and Experimental Setup	28
Dataset Description	28
Bank Tele-Marketing.....	28
Online Shoppers Purchasing.....	29
Default of Credit Card	29
Page Blocks	29
Abalone.....	30
Seismic Bumps	30
HR Employee Attrition.....	30
Pima Indians Diabetes	31
Cervical Cancer (Risk Factors).....	31
Polish Companies' Bankruptcy	31
Blood Transfusion	32
Ecoli 1	32
Ecoli 2.....	32
Wilt	33
Autism Screening Adult	33
White Wine Quality	33
Red Wine Quality	34
QSAR Androgen Receptor	34
HCC Survival	34
Mesothelioma's disease	35
Extention of Z-Alizadeh Sani	35
QSAR Oral Toxicity	35
Climate Model Simulation Crashes	35
Electrical Grid Stability Simulated.....	36
QSAR Biodegradation.....	36
Immunotherapy.....	37

Parkinson’s Disease	37
Customer Churn Wireless Telecom	37
Sales: win - loss	37
Experimental Setup and Performance Metrics	39
Summary	41
6 Performance Evaluation	42
Result Analysis for Different Datasets	42
Performance Comparison	51
Performance of Group Data	55
Statistical Analysis	55
Summary	57
7 Conclusions	68
Conclusions	68
Future Prospects of Our Work	69
Bibliography	70

List of Figures

A basic topology of Data-Driven Decision Making.....	2
Visual Representation of imbalance Dataset	4
Difference of Oversampling and Under-Sampling.....	5
Random Forest Architecture	19
Artificial Neural Network (ANN) Architecture	21
4.1 A schematic representation of the proposed approach	24
Data Reduce for five dataset using MLH Balancing Scheme	43
Class attribute count for five dataset	43
Count Percentage Difference (Part 1)	44
Count Percentage Difference (Part 2)	45
Overall Performance Evaluation for Random Forest	51
Overall Performance Evaluation for Neural Network.....	51
Result Compare between Existing Research and Multi-Layer Hybrid (MLH) Balancing Technique (Part 1)	54
Result Compare between Existing Research and Multi-Layer Hybrid (MLH) Balancing Technique (Part 2)	55
Group Performance Analysis	56
Cumulative Gain for Bank Tele-Marketing, Online Shoppers Purchas- ing, Default of Credit Card and Page Blocks	59
Cumulative Gain for Abalone, Seismic Bumps, HR Employee Attrition and Pima Indians Diabetes	60
Cumulative Gain for Cervical cancer, Polish companies bankruptcy, Blood Transfusion and Ecoli 1	61
Cumulative Gain for Ecoli 2, Wilt, Autism Screening Adult and White Wine Quality	62
Cumulative Gain for Red Wine Quality, QSAR Androgen Receptor, HCC Survival and Mesothelioma's disease	63
Cumulative Gain for Z-Alizadeh Sani, QSAR Oral Toxicity, Climate Model Simulation and Electrical Grid Stability Simulated.....	64

Cumulative Gain for QSAR Biodegradation, Immunotherapy, Parkinson's Disease and Polish Companies Bankruptcy (Y1)	65
Cumulative Gain for Polish Companies Bankruptcy (year 2, year 3, year 4, year 5).....	66
Cumulative Gain for Customer Churn Wireless Telecom and Sales: win - loss	67

List of Tables

5.1	Experimental dataset description with attribute number, class counts, and imbalance ratio	39
	Percentage of major, and minor classes for original datasets and the percentage after applying the MLH balancing scheme on the original datasets.	46
	Accuracy, F-Measure and AUC Score for existing and proposed approach using Random Forest. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.	47
	Sensitivity, Specificity and G-Mean for existing and proposed approach using Random Forest. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.	48
	Accuracy, F-Measure and AUC Score for existing and proposed approach using Neural Network. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.	49
	Sensitivity, Specificity and G-Mean for existing and proposed approach using Neural Network. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.	50

List of Abbreviations

ADASYN Adaptive Synthetic Sampling. 5, 14

AI Artificial Intelligence. 1, 2, 5

ANN Artificial Neural Network. 20, 21

CFP Classification by Feature Partitions. 52

DA Data Analysis. 1

DDDM Data-Driven Decision Making. 1, 2

GSVMRU Granular Support Vector Machines - Repetitive Under-sampling Algorithm.
53

HCAB-SMOTE Hybrid Clustered Affinitive Borderline SMOTE. 13, 54

HUSBoost Under Sampling-Based Ensemble Approach. 12

ID Imbalanced Dataset. 3

IR Imbalance Ratio. 6

k-NNFP K Nearest Neighbor Classification on Feature Projections. 52

LAD Left Anterior Descending. 52

LCX left circumflex. 52

ML Machine-Learning. 2, 3, 22

MLH Multi Layer Hybrid Data Balancing Scheme. 23, 54

OUPS Over-Sampling using Propensity Scores. 9, 53

PSM Propensity Score Matching. 9

RBF Radial Basis Function kernel. 16

RCA Right Coronary Artery. 52

RF Random Forest. 12, 17

SL Supervised Learning. 5

SMOTE Synthetic Minority Oversampling Techniques. 5

SMOTEENN SMOTE + Edit Nearest Neighbor. 5, 17

SVM Support Vector Machine. 12, 16

SVM-SMOTE Support Vector Machine with SMOTE. 5, 16

TD Training Dataset. 3

VFI Voting Feature Intervals. 52

Acknowledgement

First and foremost, I express my deepest gratitude to **Almighty Allah** for bestowing his blessings on me and giving me the ability to accomplish the achievement of this work.

I would like to express my deepest sense of thankfulness and gratitude to my thesis supervisor **Dr. Hossen Asiful Mustafa**, Associate Professor, Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka, for leading me into the research field of Data Balancing techniques and Machine-Learning. His scholarly guidance, constant and energetic supervision and valuable advice made this work a successful one. He has been a continuous source of inspiration and a real motivating force throughout my research work. I am also extremely grateful to him for providing me a high-end GPU instance for training the large dataset into Machine-Learning models to accomplish my research work.

Finally, I want to dedicate the essence of my purest respect to my parents, my classmates and my colleagues for supporting me throughout my years of study and through the process of writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Data is one of the essential elements nowadays for discovering business decisions, decision optimization, and scientific research and growing exponentially due to the use of different kinds of applications in various business organizations and production industries. The proper dataset offers organizations and researchers to analyze their showcasing techniques, make effective data-driven choices and make superior advertisements. In real-life scenarios, most data sources create a gap among class attribute elements which reduces to build a proper decision in the prediction. An imbalanced dataset creates a critical problem that affects the business decisions and makes a biased result towards the major class. However, existing data balancing techniques can solve the problems of data balancing. Existing data balancing techniques have a major drawback: these create new artificial samples randomly, which create outliers and hamper the potentiality of the original dataset. Our thesis work proposes a Multi-Layer Hybrid (MLH) Balancing Scheme that combines three over-sampling techniques and processes output in a proper way. This scheme gives a balanced and noise-free output by combining the characteristics of ADASYN, SVM-SMOTE, and SMOTE+ENN. It also creates new data points within the range of the original dataset, which keeps the originality of the new data points. Thus, the generated output from three layers is proper balancing output for machine learning models. We use 34 different imbalanced datasets with different imbalance ratios, and experimental results show balanced and proper output for the proposed scheme. We apply the resultant dataset to Random Forest (RF) and Artificial Neural Network (ANN); comparing existing techniques shows that our scheme gives better results. We used various types of the dataset in our thesis and got a different amount of result for these datasets; so we combined the results and got the average output for different metrics. Using the RF, we achieved, 82%, 83%, 83%, 84% and 91% average Accuracy; 45%, 63%, 72%, 58% and 88% average G-Mean; 39%, 55%, 62%, 51% and 83% average F-Measure for Original Dataset, ADASYN, SMOTEENN, SVMSMOTE and Proposed MLH, respectively. Using the ANN, we achieved, 78%, 77%, 74%, 80% and 79% average Accuracy; 30%, 71%, 73%, 69% and 77% average G-Mean; 26%, 59%, 59%, 60% and 67% average F-Measure for Original Dataset, ADASYN, SMOTEENN, SVMSMOTE and Proposed MLH, respectively. Using our proposed approach, we got a better outcome for the imbalanced dataset than the existing approach and observed a better performance for our proposed approach using the Random Forest.

Chapter 1

Introduction

Introduction

Dataset is one of the key elements in Artificial Intelligence (AI), which is generated every day from different types of automated applications used in our daily life. It creates a significant role in different sectors of our daily life and plays a significant role in business, finance, health care, and scientific research. Data-Driven Decision Making (DDDM) is a cycle that includes gathering information dependent on quantifiable objectives or KPIs, investigating examples and realities from the data knowledge and using them to create procedures and exercises that advantage the business in various regions. Understanding problem architecture, measuring data, implementing software and gathering meaningful information are some basic steps for DDDM and it is a very important process for Data Analysis (DA). DA allows the manager to comprehend the elements of their business, envision market moves, and oversee the chances. When looking after stock, estimating arrangements, employing ability, managers are grasping examination and orderly measurable thinking to settle on choices that improve proficiency and benefits of organizations. Data and analytics are upsetting existing plans of action and environments. From utilizing granular information to customize items and administrations to scaling computerized stages to coordinate purchasers and dealers, organizations are utilizing business investigation to empower all the quicker and truth-based dynamics. Indeed, contemplates show that data-driven organizations can settle on better decisions appreciate, high operational productivity, improved consumer loyalty, and powerful benefit and income levels. Business and financial data have become more complex in terms of calculation which requires a sophisticated mechanism for knowledge discovery. However, the success rate of data mining solutions in the business largely relies upon the interoperability of data as well as creating an effective

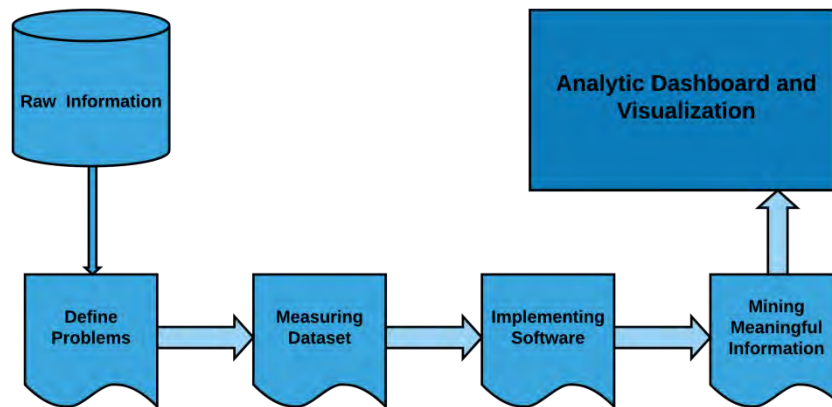


Figure 1.1: A basic topology of Data-Driven Decision Making (DDDM)

machine learning model. Figure 1.1 gives the basic idea of DDDM.

Analytical specialists today have a vast area of scientific capacities and strategies available. They can reach from the most central strategies to prescient analytics that gives progressed models to estimate and predict the future. Data is the new oil in modern life and helps to get the ideal path for organizations and comprehend it is to digitize their cycles. Digitizing client associations can give importance to the data, by which organizations can take care of procedure, deals, promoting, and item improvement. Inside digitization produces data and managers can use to improve their tasks, including directing, transportation, asset distribution, booking, capacity planning, and assembling. Data analytics with its extensive use cases and assorted applications can currently arise as the cornerstone of vital business dynamics. From empowering organizations to settle on customer situated promoting choices to encouraging them to address key operational shortcomings, the investigation is changing the insight towards the significance of the information. Advanced statistical models are assisting this reason by giving important parts of knowledge out of unpredictable informational collections and by empowering organizations to investigate new business domains. Make an inclination for expert management of data mining services rather than other tools and techniques to guarantee solid and exact conveyance of insightful answers for better business development.

Machine Learning and Data Modeling

Machine-Learning (ML) models are increasingly being used to predict future outcomes in different scenarios. ML is an important area of artificial intelligence. It creates a self-learning computers mode without express programming. The iterative part of machine learning and AI is significant on the grounds that as models are presented to new

data, they can easily adjust. Those gain outcome from previous calculations to deliver dependable, repeatable results. Things like developing volumes and assortments of accessible information, computational preparing that is less expensive and all the more remarkable, and reasonable information storage. These things mean it's conceivable to rapidly and naturally produce models that can examine greater, more perplexing data and convey quicker, more exact outcomes for an exceptionally enormous scope. And by using a precise model, an organization has a better possibility of recognizing beneficial chances. Computer can learn new information by Machine Learning algorithm. Presently many business organizations are growing more strong models for examining more perplexing data while conveying quicker, more exact outcomes for huge scopes. ML instruments empower organizations to all the more rapidly distinguish productive chances and expected dangers. New strategies in the field are advancing quickly and extended the utilization of ML to almost boundless potential outcomes. Enterprises that rely upon immense amounts of data, need a framework to examine it proficiently and precisely.

Imbalance Dataset

For creating a ML model, the training dataset (TD) is the main element to give intelligence to a machine. Unfortunately, many datasets are imbalanced in real life and are not suitable for creating a ML model with appropriate functionality. In an Imbalanced Dataset (ID), the ratio of different classes is high. Any dataset with inconsistent conveyance between majority and minority classes can be considered as an imbalance dataset, and in real-life applications, the seriousness of class imbalance can shift from minor to major. A dataset can be viewed as imbalanced if the classes row count has a huge difference. The dominant part class makes up the vast majority of the dataset, though the minority class, with restricted dataset portrayal, is frequently viewed as the class of interest. Figure 1.2 gives the visual representation of imbalance dataset where two classes represent the major class and the minor class. Major class data points rule the minor class of data points in terms of data count. For solving the imbalance issue, generally, two main processes are used in the machine learning sector. Those are oversampling and under-sampling techniques. Both of them have some advantages and disadvantages.

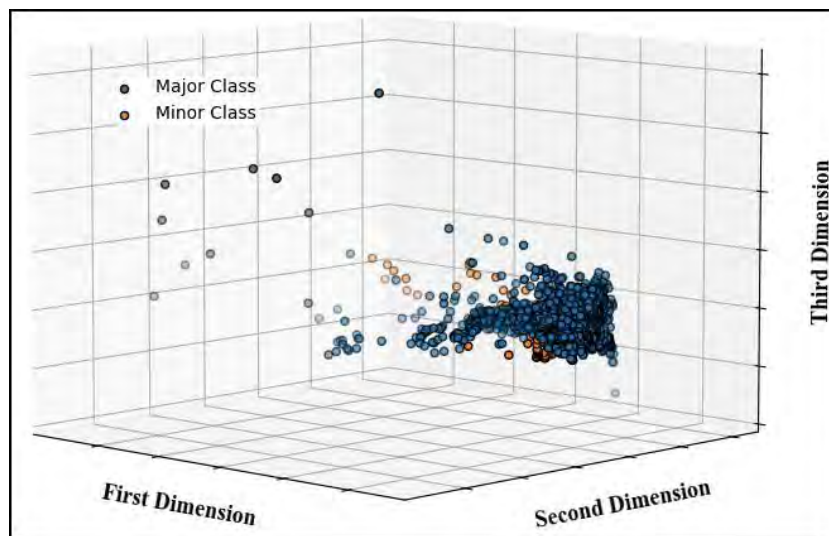


Figure 1.2: Visual Representation of imbalance Dataset

Oversampling

OverSampling [1] is a data balancing technique where the minor class is boosted to have an equal count of the major class. It creates the new data points by observing the characteristics of the minor class from original dataset using a random number generation process. Creating new data points of the minor class increases the data points comparing to the original dataset. It also creates the balancing property between the class attribute and avoids the one side result from the machine learning model. Generating new data points using a random number generation process, oversampling creates outlier and redundant data in the output. This process hampers the classification results of the machine learning model and effects future prediction.

Under-Sampling

Under-Sampling [1] is a data balancing technique that is similar to the OverSampling process. It reduces the minor class count and creates the same number of counts in the major and minor classes. For this reason, it decreases the data points from the original dataset that reduces the potentiality of the original dataset. The output sampling dataset loses the original information from the parent dataset. Figure1.3 gives the graphical representation of Oversampling and Under-Sampling process.

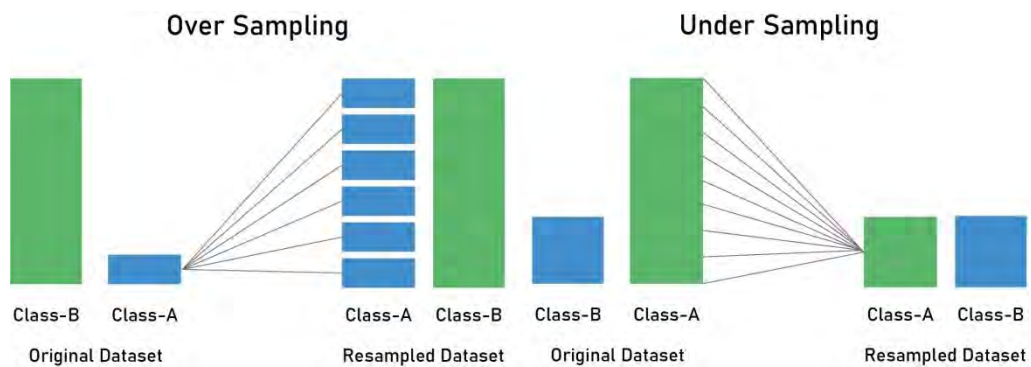


Figure 1.3: Difference of Oversampling and Under-Sampling [1]

Motivations

An imbalanced classification issue is an example of a classification problem where the appropriation of models across the known classes is one-sided or slanted. Imbalanced classifications represent a test for predictive modeling as the vast majority of the AI calculations utilized for characterization were planned around the suspicion of an equivalent number of models for each class. This outcomes in models that have poor prescient execution, explicitly for the minority class. This issue causes the minority class is more significant and accordingly the issue is more sensitive to classification blunders for the minority class than the majority class. The Imbalanced Dataset issue shows up in many real-life applications like text classification, fault detection, face recognition, social demonstrating, stock markets, medical diagnosis, etc. [2]. An Imbalanced Dataset affects the outcome of a supervised learning (SL) algorithm as it needs a target variable that contains the result for the particular input data; a model with an Imbalanced Dataset will give accurate result only for the major class but give a poor prediction for minor class [3] [4]. However, the minor class sometimes gives valuable information and it is critical to identify it properly from an Imbalanced Dataset.

Synthetic Minority Oversampling Techniques (SMOTE) [5] is a promising oversampling strategy that was proposed to enhance irregular oversampling. SMOTE is still popular due to its simplicity and effectiveness. As SMOTE creates samples by generating a random number (between 0-1), it creates random points from original datasets. Although Adaptive Synthetic Sampling (ADASYN) produces appropriate number of data points for each observation of the minority class, Support Vector Machine with SMOTE (SVM-SMOTE) creates new data points by computing the support vectors of the original dataset and SMOTE + Edit Nearest Neighbor (SMOTEENN) removes the outliers data points from the SMOTE output using K-nearest neighbor algorithm, ADASYN, SVM-SMOTE and SMOTEENN are some oversampling technologies that also create new data points by generating a random number. For this reason, these

technologies create noisy and outlier datasets that remove the originality and effect in classification models. Noisy data also creates overfitting and underfitting problem which effects in future outcomes. These problems cannot be solved by the existing oversampling techniques. However, researchers show that oversampling is quite useful than under-sampling when the Imbalance Ratio (IR) [6] [7] is higher in the dataset to keep the potentiality of the original dataset.

Objectives of the Thesis

Objectives with specific aims

The objective of this thesis is to design a new scheme to balance imbalance dataset using multi-layer hybrid model. To achieve this objective, we have identified the following specific aims:

- Find the optimum K value of K-Nearest Neighbor for the parameters of ADASYN [8], SVM-SMOTE [9].
- Design a hybrid model using combined characteristics of [8], SVM-SMOTE [9] and SMOTE + Edit Nearest Neighbor techniques [7]. The model will keep the potentiality originality of the dataset after balancing a dataset.
- Validate the proposed model with 30+ publicly available datasets with data imbalance ratio, i.e., major dataset count/minor dataset count, range is 1.6 to 59.
- Compare the proposed model with existing schemes.

Possible outcome

Successful completion of this research will result in a multi-layer hybrid data balancing scheme which can be effectively used to remove imbalance in various datasets to improve classification performance.

Thesis Outline

In the rest of this thesis, we present the details of our approach for robust gait recognition.

-
- **Chapter2** provides a survey of the existing data balancing techniques and evaluates them for their strengths and limitations.
 - **Chapter3** provides elaboration about some of the existing theories of the oversampling techniques and applied classification algorithms.
 - **Chapter4** describes the mechanism of our proposed model with flow diagram.
 - **Chapter5** describes the dataset, experimental setup and implementation of our research.
 - **Chapter6** gives the experimental evaluation of the proposed framework on 34 publicly available imbalance datasets. It also compares our results with other existing oversampling algorithms in the different experimental setup on these datasets.
 - **Chapter7** concludes the work with a summary of contributions and presents some perspectives about possible future research directions.

Chapter 2

Literature Review

We motivate working with the imbalance dataset due to some other research paper in some recent years and from the analysis of those research, we emphasize working with our proposed approach. Below, we discuss some related work of our research approach.

- **Combined Data Sampling (SMOTE + Tomek Links)**

Hartayuni et al. [10] used a hybrid approach combined with SMOTE (Over Sampling) and Tomek Links (Under Sampling) techniques to avoid imbalance problem from the dataset. They compare SMOTE, Tomek Links, Original Dataset with their combined approach and got the results using the SVM classification algorithm. They used 5 datasets in the framework and gave the results. In any case, in a few extraordinary cases combine sampling strategy is no way better than the utilize of strategies Tomek links. Based on accuracy rates in this research, utilizing a combine sampling strategy given superior outcomes than SMOTE and Tomek links in 5-fold cross approval.

For blood transfusion dataset they got 73.18% accuracy rate for SMOTE, 78.06% accuracy rate for Tomek Links and 86.22% accuracy rate for their combine sampling method. Using Ecoli 1 dataset, they got 96.90% accuracy rate for SMOTE, 90% accuracy rate for Tomek Links and 96.90% accuracy rate for their combine sampling method. Also they got 90.06% accuracy rate for abalone dataset using combine approach. They have the balance F-Measure for their combine approach. However, highly imbalanced datasets do not work well using this approach.

- **Data Sampling with Cross-Validation Approach**

Santos et al. [11] perform cross-validation during oversampling and avoid the overfitting problems. A visit test imperfection of oversampling algorithms to the complete dataset, coming about in one-sided models and overly-optimistic esti-

mates. They recognize from overfitting, appearing that the previous is related with the cross-validation strategy, whereas the last mentioned is affected by the chosen oversampling calculation. Moreover they perform by comparison of well-established oversampling calculations, backed by a information complexity analysis. In this experiment, they chose the sampling method by calculating the data complexity based on cleaning procedures, cluster-based example synthetization and adaptive weighting of minority examples.

Regarding over-optimism issue, they proposed two cross-validation (CV) method in their study; 1. CV after oversampling and 2. CV during oversampling. For first method, similar type of samples may occur in both training and testing part which creates an overoptimistic error. For the second method, testing pattern will be different from training pattern which will avoid the overoptimistic error. In this situation, the designs included within the test set are never oversampled or seen by the show within the preparing arrange; in this way, permitting a appropriate assessment of the model's capability to generalize from the preparing information. With respect to the issue of overfitting, a few analysts straightforwardly relate it to all oversampling methods, whereas others allude to the overoptimistic comes about of a CV approach as "overfitting", which confounds both concepts and prevents their distinguishing proof. Similar copies may occur in both method which creates overfitting problems for oversampling techniques. In their experiment, the Synthetic Minority Oversampling Technique was coupled with Tomek Links.

- **Hybrid OverSampling : SMOTE and Propensity Scores (OUPS)**

Rivera et al. [12] proposed a model Over-Sampling using Propensity Scores (OUPS) combined with SMOTE and propensity score matching (PSM). OUPS method calculates the propensity score and assigns to each observation of the minor class. The propensity score represents the overall likelihood for a specific group of observation and it is a metric. It calculates the variates in the feature space of the dataset. All data points in minor class will regenerate the new data points for oversampling but in descending order of propensity score. One drawback of PSM is reducing the number of samples which increases the variance. As it reproduces for all the data points in the minor class which generates outliers and noisy data points. They used Logistic Regression (LR), Support Vector Machine (SVM), Neural Network (NN) and Linear Discriminant Analysis (LDA) in their study and compare result for original and simulated data sets.

OUPS outperformed SMOTE and PSM in terms of precision and affectability. SMOTE beat all other examining methods in specificity and in F-measure. For

expanded accuracy for classification, OUPS created less untrue negatives as a result of tall affectability. False negatives represent to misclassification of a minority bunch illustration as having a place to the larger part bunch. In spite of the fact that OUPS did not beat SMOTE in terms of specificity and F-measure, it did perform moderately well with 90.78% on normal for specificity and a near F-measure making OUPS a dependable approach.

- **Combination with OverSampling and UnderSampling Approach**

Cateni et al. [13] proposed a resampling method combined with an oversampling and an undersampling technique for binary classification. They proposed Similarity based UnderSampling and Normal Distribution-based Oversampling (SUNDO) methods which combined with SMOTE-based oversampling and informed clustering-based undersampling. This strategy assists to create a balanced dataset without critical loss of information and without the addition of a great number of synthetic designs. They also used Support Vector Machine (SVM), Decision Tree (DT), labelled Self-Organizing Map (LSOM) and Bayesian Classifiers (BC) in their study for binary classification problems and compare their proposed approach. In this technique, the training dataset is separated into two classes and then, oversampling is done on minor class, and undersampling on major class. Imbalanced dataset is divided into 75% of training dataset and 25% of validation dataset to maintain the same proportion between the two classes.

- **Cluster Based OverSampling: Hepatocellular Carcinoma Patients**

Santos et al. [14] proposed an oversampling approach combined with k-mean clustering. Existing considers have several confinements that have not to be tended to: a couple of do not center completely on Hepatocellular Carcinoma patients, others have a few application barriers, and none considers the heterogeneity between patients nor the closeness of misplaced data, a common impediment in healthcare settings. In this approach, a complex Hepatocellular Carcinoma database composed of heterogeneous clinical highlights is considered. They propose a modern cluster-based oversampling approach incredible to small and imbalanced data, which distinguish for the heterogeneity of patients with Hepatocellular Carcinoma. The preprocessing methods of this work are based on information ascription considering suitable remove measurements for both heterogeneous and lost information (HEOM) and clustering thinks about to survey the basic quiet bunches within the considered dataset (K-means).

They applied the Representative set approach and Augmented sets approach by preprocessing and cleaning the HCC dataset [14]. In the Representative set approach, the entire dataset is divided into K number of portions and SMOTE is

applied to each portion. Resampling datasets of all portions are merged together to create a sub-sample. According to this research work, creating the sub-sample process runs for a particular time and generate a particular number of sub-samples and merges them properly to get a unique representative dataset. In the approach of the Augmented set, they randomly select 20% data from all sub-samples which is generated in Representative set approach and gets the best combination of sub-samples by a majority voting scheme. Both approaches work well for heterogeneous datasets but cannot solve class imbalance problem. As it chooses 20% data from all sub-samples randomly, it can take outliers also.

The proposed technique depended on a cluster-based oversampling approach where two classifiers (NN and LR) were independently utilized in two novel approaches, alluded to as Agent Set Approach and Increased Sets Approach. The most contrast of these two sets of approaches comprises employing a modern cluster-based strategy that addresses the challenges already identified at the starting of the ponder.

- **Combined Approach: SMOTE and Biased-SVM**

Wang [15] also used a hybrid technique for imbalance dataset issue combining with SMOTE and Biased-SVM approach. Biased-SVM is used to get support vectors from the dataset and sampling the support vectors using SMOTE technique. Biased-SVM is used to get an appropriate classifier. To progress the computation productivity of the calculation, it is proposed by combining Destroyed and Biased-SVM approaches, which restrain the bolster vector not as it were inside the bolster vectors, but moreover inside the complete minority lesson. To confirm the viability of the proposed calculation, four diverse UCI datasets are received to approve this approach through reenactments. The comes about show the proposed approach can receive superior execution than the initial approaches. Their proposed approach gives better sensitivity for the minority class.

- **Real-time online purchasing prediction: Imbalance Dataset**

Sakar et al. [16] utilized a real-time online customer behavior examination framework comprising of two modules which at the same time predict the visitor's shopping expectation and Web location deserting probability. In the first part, they anticipate the obtaining deliberate of the guest utilizing amassed pageview information kept track during the visit in conjunction with a few session and client data. The extricated features are nourished to Random Forest (RF), Support Vector Machines (SVM), and Multilayer Perceptron (MLP) classification model as input. They utilize oversampling and feature choice methods to progress the execution and adaptability of the classifiers. The outcome appears that MLP that's

calculated utilizing versatile backpropagation calculation with weight backtracking produces altogether higher precision and F1 Score than RF and SVM. Another finding is that clickstream information achieved from the routeway taken after amid the online visit pass on critical data almost the acquiring deliberate of the guest, combining them with session information-based features that have special data almost the obtaining intrigued moves forward the victory rate of the framework. In the second approach, utilizing as it were consecutive click stream information, they prepare a long short-term memory-based repetitive neural arrange that creates a sigmoid yield appearing as the likelihood gauge of visitor's purposeful to take off the location without finalizing the exchange in an expectation skyline.

Their approach bolsters the possibility of exact and adaptable obtaining purposeful forecast for virtual shopping environment utilizing clickstream and session data information.

- **Hybrid Under-Sampling Method (HUSBoost): Imbalanced dataset**

Mahmudul Hasan Popel et al. [17] propose a modern crossover under sampling-based ensemble approach (HUSBoost) to handle imbalanced information which incorporates three essential steps- information cleaning, information adjusting, and classification steps. They evacuated the noisy information utilizing Tomek-Links. After that, they made a few adjusted subsets by applying the random undersampling (RUS) strategy to the larger part class instances. These under-sampled majority and minority instances constitute the subsets of the imbalanced information. Having the same number of majority and minority class, they got to be adjusted subsets of information. At that point in each adjusted subset, random forest (RF), AdaBoost with the decision tree (CART), and AdaBoost with Support Vector Machine (SVM) are actualized in parallel where they utilized a delicate voting approach to induce the combined result. From these ensemble classifiers, they got the normal result from all the adjusted subsets. They utilized 27 datasets with diverse imbalanced proportion in arrange to confirm the viability of our proposed demonstrate and compare the test comes about of our demonstrate with RUSBoost and EasyEnsemble strategy. For 27 datasets (10 cases for each information set) and inside 270 cases, RUSBoost performed way better in 61 cases and EasyEnsemble in 50 cases where the proposed strategy (HUSBoost) in this inquire about work given superior comes about in the rest of the 159 cases. They utilized 5 fold and 10 fold cross-validation for comparing the test outcomes with a shifting sum of information conjointly to realize promising comes about on distinctive execution assessment measurements. They moreover compared

outcomes about with two exceptionally prevalent under-sampling based ensemble methods (RUSBoost and EasyEnsemble) and in most of the cases, HUSBoost given way better outcomes.

- **Hybrid Clustered Affinitive Borderline SMOTE (HCAB-SMOTE): Binary Classification**

Al Majzoub et al. [18] proposed a novel approach for creating synthesized information known as Hybrid Clustered Affinitive Borderline SMOTE (HCAB-SMOTE). It used to play down the number of created instances whereas expanding the classification recall. It combines under-sampling for expelling larger part commotion instances and oversampling approaches to improve the density of the borderline. It employments k-means clustering on the borderline zone and identify which clusters to oversample to get better outcomes. Test outcomes about appear that HCAB-SMOTE outflanked SMOTE, Borderline SMOTE, AB-SMOTE and CAB-SMOTE approaches which were created some time recently coming to HCABSMOTE, because it given the most noteworthy accuracy with the slightest number of produced instances.

- **OverSampling Approach: Forecasting Wheat Yields**

Chemchem [19] proposed a method of anticipating wheat production dependent on machine learning and it can calculate the accurate value of wheat production losses in France. They used a highly imbalanced dataset and applied a hybridization of SMOTE and machine learning approach for forecasting wheat production on particular dataset. By their approach, they get better operating characteristic of the ROC receiver and get 90.07% accuracy on the testing phase for comparative analysis by hybridizing the SMOTE algorithm with the Random Forest algorithm.

In our proposed approach, we used different highly imbalance datasets (Wilt [20], Polish companies bankruptcy [21], Abalone (Imbalanced: 19) [22], White Wine Quality [23] and Red Wine Quality [18]) and got suitable results compared to existing approaches. Moreover, our approach can eliminate the outliers from the output.

Chapter 3

Background

There are some established technologies exist which can solve the imbalance problem. In the background of our study, we discuss the existing methods which we used in our approach. Those are ADASYN, SVM-SMOTE, and SMOTE-ENN. We also discuss two classification models: Random Forest, and Artificial Neural Network, which are used to analyze performance of our approach.

Adaptive Synthetic Sampling Approach (ADASYN)

Adaptive Synthetic Sampling (ADASYN) is an oversampling approach and it is similar to the SMOTE technique. It improves the learning of SMOTE in two way. First one is, removing the bias data points and second is adaptively moving the decision boundary towards the critical space of the data points. The key idea behind this approach is the density distribution ratio (r_i) of the original dataset. By calculating the neighborhood ratio for each minor data point, it generates new data points for boosting minor class. First it calculates the degree of class imbalance by calculating the ratio of minor and major class. The degree should be less than the threshold. If data points under any k nearest neighbor (kNN) group are homogenous (every point is minor), no point will be generated related to this group. If data points under any kNN group are heterogeneous (mix with major and minor data points), minor data points will be generated by randomly choosing minor data points inside the group. The new dataset will be generated by calculating the major data points ratio (neighborhood ratio) inside each kNN group. The relation between creating the number of data points for each minor data point is inversely proportional to the number of major data points inside a kNN group. If the minor data points amount is less than the major data points, then the neighborhood ra-

ratio is higher and a large number of new data points are created that are related to the particular group. So, it maintains a dynamic density distribution for each group created by kNN value. The disadvantage of this method is that it cannot work for sparsely distributed dataset as it works with the idea of density distribution ratio. If the dataset is sparse then it simply terminates the process. Also, KNN value is not dynamic in this process and need to define it manually. The process of ADASYN using below equations.

The degree of class imbalance refers by equation 3.1,

$$d = C_{minor}/C_{major} \quad (3.1)$$

$$d < d_t \quad (3.2)$$

Where C_{minor} = minor class count, C_{major} = major class count, d_t = threshold and $d \in (0, 1]$

The number of synthetic data examples that need to be generated for the minority class is calculated by equation 3.3.

$$T = (C_{major} - C_{minor}) * \beta \quad (3.3)$$

Where T is the generate number and β is a constant ($\beta = 1$ means fully balance output). Finally, output will be created by repeating calculation of equation 3.5. Equation 3.4 helps to execute loop from 1 to t .

$$t = T * r_i \quad (3.4)$$

$$X_{new} = X + (X - X_z) * \lambda \quad (3.5)$$

Where X_{new} = new data point, X = minor data points, X_z = randomly selected data point inside group, t is the total execute number of loop and $\lambda \in [0, 1]$

SVM-SMOTE

Support Vector Machine with SMOTE (SVM-SMOTE) is an oversampling method creating data points at the boundary line between the major and minor classes of the original dataset. According to this process, minority class will expand toward the majority class in such a manner where major data points are lower in density. With this methodology, minor data points can perceive even in the areas close to the limit where there doesn't exist any agent of them in the training set. For this reason, it creates new data points with proper density and less overlapping. First, it will create a hyperplane using the Support Vector Machine (SVM) algorithm and find the support vectors of the minor class. Using Radial Basis Function kernel (RBF), it finds the maximum margin of two classes and gets the best hyperplane of the dataset. A hyperplane can be defined using equation 3.6.

$$f(x) = \beta_0 + \beta^T x \quad (3.6)$$

Where, β is known as the weight vector and β_0 as the bias.

To maximize the hyperplane, equation 3.7 can be used.

$$|\beta_0 + \beta^T x| = 1 \quad (3.7)$$

Where x symbolizes the training examples closest to the hyperplane.

In particular, for the accepted hyperplane, the numerator is equal to one and the distance to the support vectors is given by equation 3.8:

$$distance_{sv} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|} \quad (3.8)$$

Radial Basis Kernel (RBF) can be defined as equation 3.9.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2); \gamma > 0 \quad (3.9)$$

By support vectors of minor class, new instances will be created randomly by joining each minor data with every point of its kNN group using the interpolation (inside the nearest neighbor group) or extrapolation (outside the nearest neighbor group) technique depending on the density of major data points around it. If major data points count is less than half of its nearest neighbors, new instances will be created outside of the

kNN group to expand minority class area toward the majority class. This extension is necessary to remove less density of major instances. In the opposite situation, a new instance will be created inside the kNN group for less density of minor class. The boundary of the minor class will be consolidated similar to SMOTE. But, the point will be selected by the order of the kth nearest neighbor instead of randomizing the selection of the nearest neighbor. Equation 3.10 and 3.11 represent the new data point for extrapolate and interpolate, respectively.

$$X_{new}^+ = SV_i^+ + (SV_i^+ - A_k) * \lambda \quad (3.10)$$

$$X_{new}^+ = SV_i^+ + (A_k - SV_i^+) * \lambda \quad (3.11)$$

Where SV_i^+ is positive support vectors, A_k is Array containing k positive nearest neighbors of each positive and $\lambda \in [0, 1]$.

SMOTE + Edit Nearest Neighbor (SMOTE + ENN)

SMOTE + Edit Nearest Neighbor (SMOTEENN) technique used SMOTE for oversampling and is a very powerful technique for cleaning the oversampling output. It removes the noisy output from the oversampling dataset and gives a proper output from the original dataset. Edited Nearest Neighbor (ENN) used $k=3$ for the nearest neighbor to remove noisy data points and cleaning the output. ENN removes any example whose class label differs from the class of at least two of its three nearest neighbors. For binary classification, if any data points belong to the major class and three nearest neighbors mis-classify the data point, then the major data points are removed. If data points belong to the minor class and nearest neighbor mis-classifies, then major data points of this particular nearest neighbor are removed. In this way, it cleans the data points from both classes and removes mis-classifying data points.

Random Forest (RF)

Random Forest (RF) [24] is one of the most prominent and easily manageable learning techniques in machine learning. The RF is utilized in a variety of fields, such as banking, the stock exchange, medication and online business. In finance, for instance, it is utilized to recognize clients bound to reimburse their obligation on schedule, or utilize

a bank's services frequently. In this space it is used to identify fraudsters out to trick the bank. In exchanging, RF can be utilized to decide a stock's future conduct. In the medical services space, it is utilized to recognize the right blend of segments in medication and to break down a patient's clinical history to distinguish illnesses. Arbitrary backwoods are utilized in web-based business to decide if a client will really like the item or not.

RF can use for both classification and regression problems. It makes various decision trees dependent on subsets of the dataset utilizing substitution technique at the preparing stage and predicts the class by classification approach or by calculating mean estimation in regression approach [25,26]. The more trees in the RF makes the model increasingly imposing and profound for analysis. Important feature attributes also can be identified by this algorithm. By this way, dropping of unnecessary features is easy for machine learning technique. It also helps to avoid the overfitting and vice versa. The hyperparameters of RF is also important to increase the forecasting power of the model or making faster model. We can set the hyperparameters by selecting the number of trees and cross-validation method. It takes care of the overfitting issue by creating several sub-trees [27]. Combining a decision tree with a bagging classification is not necessary for this algorithm because of easily utilizing the classifier-class. It adds randomness to the model for growing the trees in the execution phase. It does not find the vital attributes from the splitting node. It finds the best feature among a random subset of attributes. This outcomes in a wide variety that outcomes in a better model. Therefore, only an arbitrary subset of the attributes is taken into consideration by the algorithm for splitting a node. Figure3.1 gives the visual representation of the mechanism for Random Forest Model.

It can use Bagging or Bootstrap aggregation. Bagging is used to reduce the variance of a decision tree. Here, the subset of entire preparing datasets picked and models are built for each subset. We can get a robust and effective result by taking the average of all the predictions from different trees using Bagging. Bootstrap is a measurable methodology to arrive at the midpoint of qualities from the sub-test and it gives the better output utilizing the mean value. RF requires attribute selection measure and pruning method for designing several sub-trees.

The most concerning issue in machine learning is overfitting; yet more often than not this won't occur on RF. If there are sufficient trees in the RF, the classifier won't overfit the model. The primary impediment of RF is that countless trees can make the calculation excessively moderate and inadequate for continuous expectations. These algorithms are quick to prepare, yet very delayed to make forecasts whenever they are trained. A more precise expectation requires more trees, which brings about a

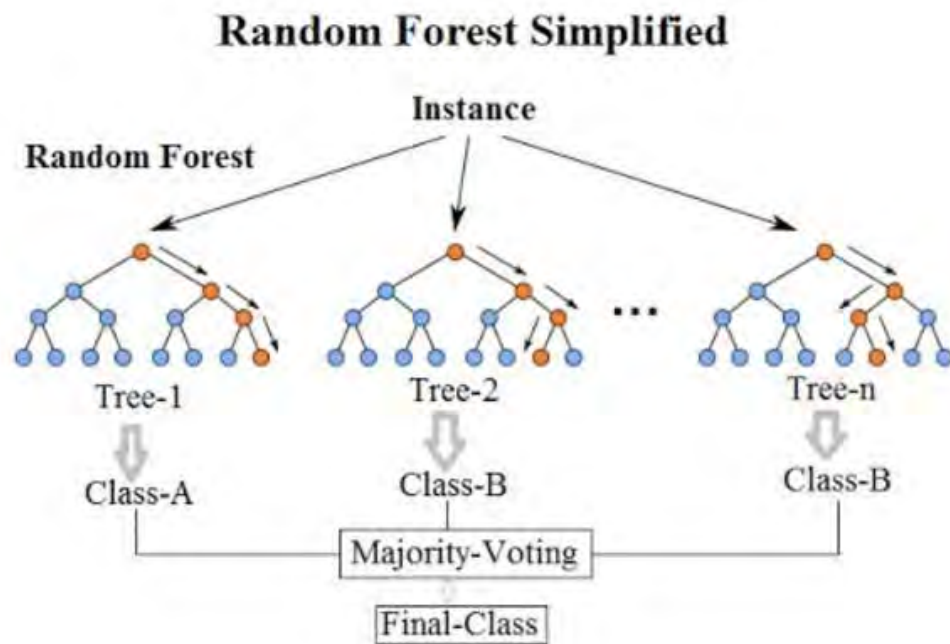


Figure 3.1: Random Forest Architecture

slower model. In most real-world applications, the RF is sufficiently quick however there can absolutely be circumstances where run-time execution is significant and different methodologies would be liked. RF is a prescient displaying apparatus and not an illustrative instrument, which means in case you're searching for a depiction of the connections in your information, different methodologies would be better.

Information Gain Ratio (Entropy) [28] and the Gini Index [29] are the degrees of probability for attribute selection and impurity measurement for the class attribute in the decision tree.

We can represent the GINI index and Entropy using equation 3.12 and equation 3.13 respectively.

$$Gini(t) = 1 - \sum_{j=1}^n p_j^2 \quad (3.12)$$

Where, p_j = Relative Frequency of class j in dataset (D).

$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_- \quad (3.13)$$

Where, S is the training Example, P_+ is the proportion of positive example and P_- is

the proportion of negative example.

Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) [30] is a computational model, which is similar to the structure of human neural systems. ANNs have three layers that are interconnected. An ANN has lots of artificial neurons called nodes, which are interconnected by hubs. These preparing units are comprised of input layer, hidden layer and output layer. The input nodes get different structures constantly of data dependent on an inner weighting framework, and the neural organization endeavors to find out about the data introduced to create one output report. ANNs utilize a bunch of learning rules called backpropagation, a truncation for in reverse proliferation of blunder, to consummate their output. ANN are making ready forever changing applications to be created for use in all areas of the economy. ANN have been applied in different activities to make an interpretation of website pages into different languages, to having a menial helper request food supplies online to chatting with chatbots to tackle issues. Email service providers use ANNs to identify and erase spam from a client's inbox; resource chiefs use it to figure the course of an organization's stock; web based business stages use it to customize proposals to their crowd; chatbots are created with ANNs for normal language handling; profound learning calculations use ANN to foresee the probability of an occasion; and the rundown of ANN joining goes on across various areas, ventures, and nations [31].

It is useful for a large number of datasets and can handle a large dataset properly in training phase. ANN is utilized for prescient modelling. The explanation being that ANN as a rule attempts to over-fit the relationship. ANN is utilized in situations where what has occurred in past is rehashed precisely in same way. It is a kind of AI method which has tremendous memory. However, it doesn't function admirably in the event that where scoring populace is altogether extraordinary contrasted with preparing test. For this reason, it functions admirably in instances of picture acknowledgment and voice acknowledgment. In training stage, it learns to perceive designs in data. During this supervised stage, the network compares its exact yield delivered and what it was intended to create—the ideal output. The distinction between the two results is changed utilizing backpropagation. This implies that the network works in reverse, going from the output node to the input node to change the weight of its associations between the units until the distinction between the real and wanted result creates the most minimal error.

The primary layer comprises of information neurons. Primary neurons send informa-

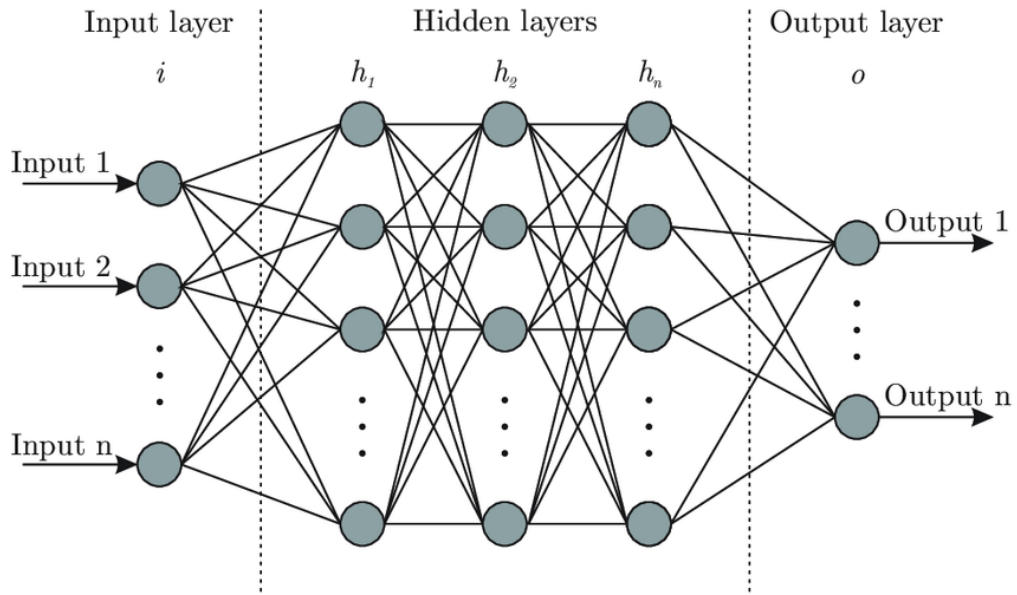


Figure 3.2: Artificial Neural Network (ANN) Architecture

tion to the subsequent layer, which then sends to output neurons to the third layer. ANNs is viewed as nonlinear information demonstrating system where existing powerful connections among sources of information and displayed result [32]. In the training phase, dataset routes from node to node between different layers and updates weight value for every node automatically according to the training dataset. Figure3.2gives the architecture of the Artificial Neural Network (ANN).

In ANN, each concealed hub contains a weight, which is refreshed to fit our model for the training dataset in the training stage. Therefore, in the training stage, the weight update works utilizing beneath procedure in our model.

$$Z = \sum_{i=1}^n w_i x_i + b \quad (3.14)$$

Where, n = the number of inputs from the incoming layer, i = counter from 1 to n , w_i = initial weight for every note, x_i is the input variable for our dataset, b represents biased value and Z refers for activation function.

$$y = \begin{cases} 1, & \text{if } Z \geq \theta. \\ 0, & \text{otherwise.} \end{cases} \quad (3.15)$$

Where, y is the actual output and θ represents threshold.

To calculate error, we will use equation3.16.

$$E = \sum \frac{1}{2}(t - y)^2 \quad (3.16)$$

Where, t is the targeted output.

$$\Delta W = LR \times E \times x_i \quad (3.17)$$

Where, LR is the learning rate and ΔW represents the change of weight for every node in each iteration.

$$W_{new} = W_{old} + \Delta W \quad (3.18)$$

Summary

In this thesis, we used these techniques which we discussed in this background section. We discussed ADASYN, SVM-SMOTE and SMOTE + ENN which are the popular existing oversampling balancing techniques for imbalanced datasets. We also discussed Random forest (RF) and Artificial Neural Network (ANN) which are the existing classification method in ML. All of those are related to our thesis work. In our thesis work, we proposed a multi-layer hybrid balancing framework that is combined with the discussed balancing techniques in the background section. We used the characteristics of those balancing techniques to achieve our framework. When we used our balancing approach to the imbalanced datasets and get the balanced dataset as output from the original input then we used the discussed classification algorithms to the balanced datasets. The outcome from our approach is better than the other existing balancing techniques.

Chapter 4

Proposed Approach

Multi Layer Hybrid Data Balancing Scheme (MLH)

In this thesis, we propose a multi-layer hybrid data balancing scheme (MLH) using a combination of ADASYN, SVM-SMOTE, and SMOTE + ENN approach. As most data balancing techniques works based on random number generate process to balance the original dataset, those creates outliers and outrange data. This is a critical issue for the machine learning models which hampers the potentiality of the original dataset. To overcome the problem, our approach removes the outrange data points from the oversampling techniques and merge them together to get a combined characteristics output. It helps to get all usefulness of ADASYN, SVM-SMOTE, and SMOTE + ENN in output dataset with perfect balancing. Our proposed scheme contains multiple layers. In the initial layer, we use ADASYN, SVM-SMOTE, and SMOTE + ENN approach separately. All separate chunk of dataset from three oversampling techniques pass to the Process Engineering Layer. In the Process Engineering Layer, we remove the outlier data points by parallel processing. After that, we merge the parallel processing input and remove duplicate from the combined dataset. In the final layer, we apply the SMOTE + ENN scheme to the output of the Process Engineering Layer and get the final balanced dataset. Figure4.1 shows the schematic representation of the proposed scheme and the overall steps of our proposed approach is described in the following.

Initial Layer

In this layer, we combined the output of three oversampling techniques and remove the outrange output from the oversampling techniques. Data potentiality is also preserved

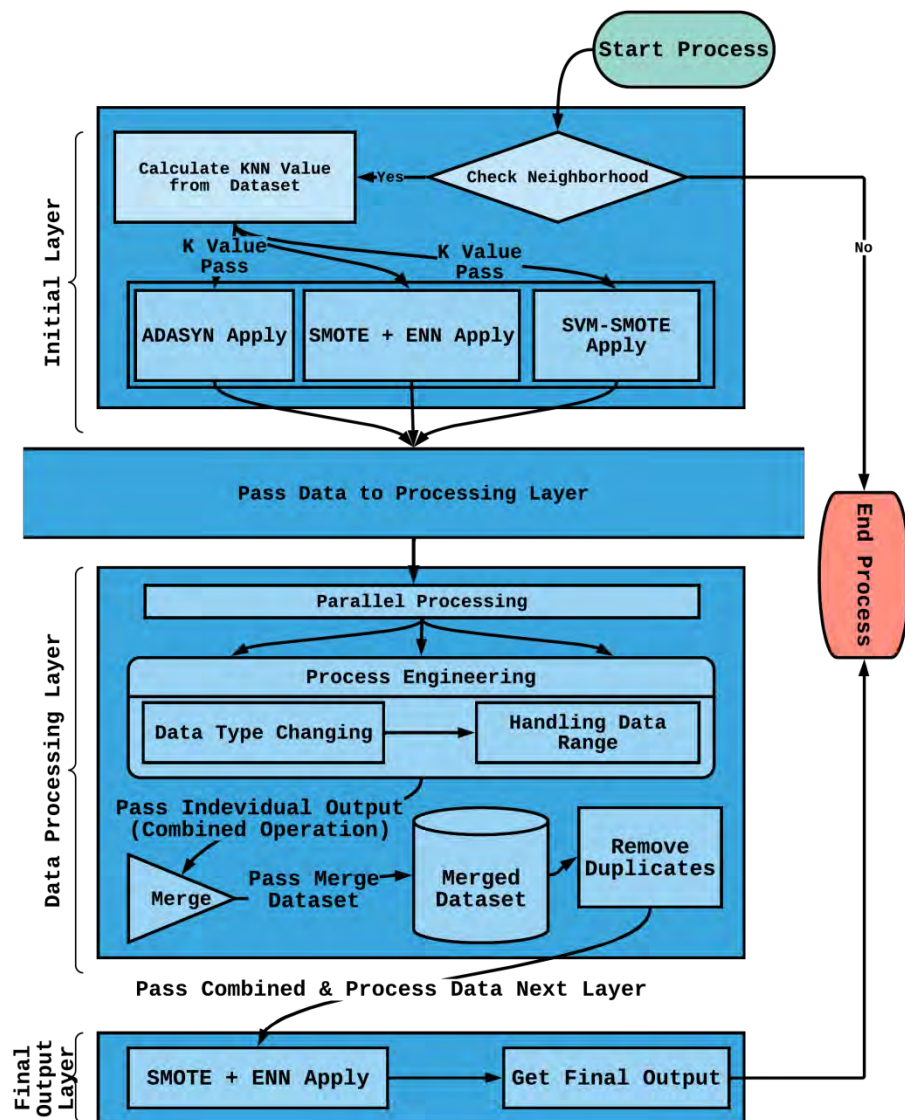


Figure 4.1: A schematic representation of the proposed approach

in this layer and pass the process output to the second layer. The procedures of this layer are following:

Check Neighborhood

Our hybrid scheme contains ADASYN over-sampling, which does not work for the sparsely distributed dataset. So, we need to calculate the neighborhood of the classes inside a particular kNN group to check the sparsity; we do this by using k nearest neighbor algorithm. If an input dataset does not give any neighborhood value, then the dataset cannot get any result using the ADASYN, and consequently our proposed scheme will terminate.

Calculate KNN

Our scheme can dynamically detect the appropriate K from the original dataset by calculating the accuracy of the minor class. If the number of minor class count of the original dataset contains small number (data count less than or equal to 20), then the k-value is set to 3. For other cases, we use the K Nearest Neighbor algorithm [33] for different numbers of K-value with 5-fold cross-validation [34] and check the appropriate K-value for the original dataset by measuring the maximum accuracy for minor class. As cross-validation can use data points in the process dataset, it can achieve accurate result and remove the overfitting from the machine learning model. It also helps to build machine learning model using all dataset and give a clear understanding of the dataset. For this reason, we use this technique to find the k-value from the original dataset. The range of K-value is between 3 to 20, and we choose only the odd number between the specific range as a parameter for the K Nearest Neighbor algorithm. We choose odd number because the K Nearest Neighbor algorithm finds the k numbers of nearest neighbor for all training data points and finds the class label for the testing data points, the number of K value can create a tie result if the k value is not odd. This appropriate K value is used as a parameter to calculate the ADASYN and SVM-SMOTE approach.

Apply Oversampling Techniques

After calculating K, we execute ADASYN and SVM-SMOTE using the calculated K parameter. For SMOTE + ENN, we use the default value K=3. We use ADASYN as it works with the distribution ratio of the nearest neighbor group and gives a proper distribution of oversampling data points. We use SVM-SMOTE to create a boundary line between major and minor dataset and generate a less degree of overlapping. SMOTE + ENN gives us a noise-free clean dataset from the original dataset. Individually, these three approaches work differently; however, our proposed scheme can generate a properly distributed, and noise-free dataset with boundary line by combining the results of these approaches.

Process Engineering Layer

Sometimes ADASYN, SVM-SMOTE, and SMOTE + ENN approaches hamper the data type of original dataset which affects the potentiality of the original dataset. So, in our process engineering part, we clean the three oversampling in two phases. This process

engineering part is applied separately for the three oversampling processes. After applying process engineering part, we merge the output data and remove the duplicates from the output.

Changing Data Types

In the phase, we pass the oversampling output as the input and change the data type like the original dataset. By checking the data type from original dataset, our approach can automatically change the type of oversampling output. As existing oversampling techniques create output sample randomly, it cannot control the data type in their process. So, by identifying the mismatch of data type, our approach can change the out which is the first part of the data processing stage.

For example, if an attribute in the original dataset has Integer type property and applied oversampling approach created this attribute as Real type property, then our scheme changes this attribute to an Integer. For this reason, our scheme keeps the originality and potentiality of the original dataset.

Handling Data Range

We also observed that data points exceed the range of the original dataset after applying the three existing approaches for oversampling. For this reason, the output of the three existing approaches creates redundant data points. In the second phase of process engineering, our scheme solves these problems by removing the data points outside of the boundary to keep the originality of the actual dataset. By this way, we can also remove the outlier from the existing oversampling approach.

Combine and Duplicate Remove

After processing the data using Changing Data Types and Handling Data Range for three different oversampling approaches separately, we merge the outputs and then, remove the duplicates from the combined output. By removing duplicates, we can keep data potentiality as well as remove data overlapping. Combine approach also helps to achieve all characteristics of three existing oversampling approach what we used in our proposed approach.

Final Output Layer

We used ADASYN, SVM-SMOTE, and SMOTE + ENN oversampling approach in the first layer to boost the minor class of the original dataset. For this reason, the minor class becomes major in the first layer and an imbalanced dataset is generated. So, after combining the dataset, we pass the combined output as input to the next layer and apply the SMOTE + ENN approach in this layer. As the second layer uses SMOTE + ENN approach, it removes the noise from the major class which is minor for the original dataset. Thus, it creates noise and outlier free data for the minor class of the original dataset and gives a better recall value for minor class (Positive class) of original data. Due to creating a dataset in a distributed manner from the original dataset and removing outliers, our proposed scheme avoids overfitting and keeps the potentiality of the original dataset.

4.2 Summary

Although there are other techniques besides ADASYN, SVM-SMOTE, and SMOTE+ ENN for balancing datasets, we achieved better balancing results by combining these three approaches in our scheme for various dimensions of different imbalanced datasets. Use of our data engineering and tuning process along with the combination of these approaches make our final output noise and outlier free. As we applied our data engineering separately to ADASYN, SVM-SMOTE, and SMOTE+ ENN output, our scheme maintains the authenticity of data and removes the out-range data from the output. Our approach removes the redundant data points by using the SMOTE + ENN in the second layer. For this kind of architecture, we are able to get better results compared to other approaches even though the experimental datasets are highly imbalanced.

Chapter 5

Dataset and Experimental Setup

Dataset Description

In our experiment, we used different kind of imbalance dataset, e.g., financial, medical, Bio Life, and employee performance related dataset which contains different level of imbalances. All datasets contain different dimensions and different labels of attributes with robust characteristics. Using equation 5.1, we calculate the Imbalance Ratio (IR) for every dataset. If IR is higher, then the dataset is more imbalanced. For validating our proposed scheme, we used 29 imbalance binary classification datasets with IR value range between 1.6 to 59 where different dataset gives different result after making them balanced. The Polish companies bankruptcy [21], one of the datasets used in our experiment, is divided into 5 years. We apply our scheme separately for each year dataset as well as for all years. Thus, our total experimental datasets are 34. Below gives the overall description of the 29 dataset which applied in our study.

Bank Tele-Marketing

Bank Tele-Marketing [35] is a business-oriented dataset which collected from the banking organizations of Portugal. It is the collection of marketing campaigns over the phone call. It contains 41188 rows and 20 columns. 19 of them are attribute columns and one is class column. It contains client information, last transaction information of campaigning clients, social information of the clients. It is a multivariant dataset which mixed with numerical and categorical attributes. Job, Marital Status, Education, Credit in Default, House Loan and Personal Loan are the categorical attributes which contains the unknown (Missing) value for this dataset.

Online Shoppers Purchasing

The area of this dataset is business related and it contains the online shopping purchasing information. Online Shoppers Purchasing [16] is highly imbalanced binary classification dataset. It has 12330 data points with 18 attributes. It contains 10 numerical and 8 categorical attributes. The 'Revenue' attribute is the class column for this dataset. "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" define the quantity of various kinds of pages visited by the visitors in that session and all out time spent in every one of these page classifications. The "Bounce Rate", "Exit Rate" and "Page Value" features define to the measurements estimated by "Google Analytics" for each page in the online business webpage. The estimation of "Bounce Rate" highlight for a website page alludes to the level of guests who enter the webpage from that page and afterward leave without setting off some other solicitations to the examination worker during that session. The estimation of "Exit Rate" highlight for a particular website page is determined with respect to all online visits to the page, the rate that were the toward the end in the session. The "Page Value" include speaks to the normal incentive for a website page that a client visited prior to finishing a web-based business exchange. The dataset also incorporates working framework, program, locale, traffic type, guest type as returning or new guest, a Boolean worth showing whether the date of the visit is end of the week, and month of the year.

Default of Credit Card

This dataset [36] describes the default payments of customers in Taiwan which compares the classification accuracy and statistical analysis using data mining. This dataset analysis and represents the risk factors of customers. It contains 30000 data points with 24 attributes. 23 attributes are the feature attributes and 1 is class attribute. It is a binary classification dataset with highly imbalanced. All attributes are numeric columns in this dataset. X1, X2, X3, X4, X5 represents the given credit amount, Gender, Education, Marital Status and Age, respectively. X6-X11 repayment History of Payment for six months; X12-X17 Bill statement; X18-X23 Previous Payment.

Page Blocks

The dataset [37] comprises of characterizing all the page design of a record that has been identified by a segmentation cycle. It contains 5473 data points with 10 attributes.

The dataset characteristics of Page Block data is multivariate which contains real and integer type attributes. The 5473 data points come from 54 unique documents where Each observation considers as a block. Data are in a format readable by C4.5.

Abalone

This dataset [22] uses to predict the age of abalone from physical measurements. By this dataset using machine learning techniques, researchers can automatically detect the age which avoid boring task. The dataset characteristics of Abalone data is multivariate which contains real and integer type attributes. There are 4177 data points with 10 attributes.

Seismic Bumps

Mining is happening threats which are usually called mining risks. A danger from mining is a seismic risk that regularly happens in numerous underground mines. Seismic peril is the hardest noticeable and unsurprising of normal risks. Besides, with the data about the chance of risky circumstance event, a suitable management administration can decrease the danger of rock burst or pull out laborers from the undermined territory. A decent expectation of expanded seismic movement is in this manner a matter of incredible down to earth significance. The dataset [38] describes the issue of high energy seismic bumps predicting in a coal mine. It contains two longwalls located from a Polish coal mine. The introduced dataset is portrayed by the lopsided dispersion of positive and negative models. In the informational index, there is just 170 positive example which is class 1.

HR Employee Attrition

The key to success in any organization is attracting and retaining top talent. But due to some circumstances, employees are prompt to leave. This dataset [18] helps HR analyst to solve this problem. It contains 1470 data points with 35 attributes information. It is multivariant dataset mix with categorical and real data points.

Pima Indians Diabetes

This dataset [39] is produced by the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of the dataset is to symptomatically foresee whether a patient has diabetes, in light of certain demonstrative estimations remembered for the dataset. Specifically, all patients here are females with greater than or equal to 21 years of age of Pima Indian legacy. It contains 768 data points and 9 attributes (8 of them are feature attributes). It is combined with real and integer value and having a binary classification property.

Cervical Cancer (Risk Factors)

This is the dataset [40] about cervical malignant growth events. Cervical cancer growth is one the most regular disease sicknesses that happen to women. This dataset is demonstrating a few factors that may impact cervical disease. The dataset was gathered at 'Hospital Universitario de Caracas' in Caracas, Venezuela which contains segment data, propensities, and memorable clinical records of 858 patients. A few patients chose not to respond to a portion of the inquiries on account of protection concerns (missing qualities).

Polish Companies' Bankruptcy

The dataset [21] is about the bankruptcy expectation of Polish organizations. The information was gathered from Emerging Markets Information Service (EMIS), which is an information base containing data on developing business sectors around the world. The bankrupt organizations were examined in the period 2000-2012, while the as yet working organizations were assessed from 2007 to 2013. Basing on the gathered information five classification cases were recognized, which relies upon the prediction time frame. In 1st year, the information contains monetary rates from the first year of the anticipating time frame and comparing class name that demonstrates chapter 11 status following 5 years. The information contains 7027 occasions (budget reports), 271 speaks to bankrupted organizations, 6756 firms that didn't bankrupt in the anticipating time frame. In 2nd year, the information contains monetary rates from the second year of the prediction time frame and relating class mark that demonstrates liquidation status following 4 years. The information contains 10173 occurrences (budget summaries), 400 speaks to bankrupted organizations, 9773 firms that didn't bankrupt in the estimating time frame. In 3rd year, the information contains monetary rates from the third

year of the anticipating time frame and relating class name that shows insolvency status following 3 years. The information contains 10503 examples (budget summaries), 495 speaks to bankrupted organizations, 10008 firms that didn't bankrupt in the anticipating time frame. In 4th year, the information contains monetary rates from the fourth year of the anticipating time frame and comparing class name that shows liquidation status following 2 years. The information contains 9792 occurrences (fiscal reports), 515 speaks to bankrupted organizations, 9277 firms that didn't bankrupt in the anticipating time frame. In the 5th year, the information contains monetary rates from the fifth year of the anticipating time frame and comparing class name that shows insolvency status following 1 year. The information contains 5910 occurrences (fiscal summaries), 410 speaks to bankrupted organizations, 5500 firms that didn't bankrupt in the anticipating time frame.

Blood Transfusion

This dataset [41] received the benefactor information base of Blood Transfusion Service Center in Hsin-Chu City in Taiwan to show the RFMTC showcasing model (a changed rendition of RFM). They did their blood transfusion service transport to one college in Hsin-Chu City to accumulate blood gave about like clockwork. To assemble a FRMTC model, they chose 748 contributors at arbitrary from the benefactor information base. These 748 benefactor information, every one included R (Recency - months since last donation), F (Frequency - all out number of donation), M (Monetary - absolute blood gave in c.c.), T (Time - months since first donation), and a double factor speaking to whether he/she gave blood in March 2007 (1 represent giving blood; 0 represents not giving blood). It contains no missing value and having 4 feature attributes in this dataset.

Ecoli 1

This is the ecoli-0-3-4 dataset [10] which has the information of the ecoli. It has 200 data points and 7 feature attributes. Although it is comparatively small dataset, it is highly imbalance dataset. All attributes are real in this dataset.

Ecoli 2

This is the ecoli-0-1-3-7 dataset [10] which has the information of the ecoli. It has 281 data points and 7 feature attributes. Although it is comparatively small dataset,

it is highly imbalance dataset (97.51% Negative Class and 2.49% Positive Class). All attributes are real in this dataset.

Wilt

Wilt dataset [20] detects the diseased of Japanese Oak Wilt and Japanese Pine Wilt using satellite imagery. The information comprises of picture portions, produced by fragmenting the pansharpened image. The fragments contain phantom data from the Quickbird multispectral image bands and surface data from the panchromatic (Pan) picture band. The testing information is for the row with Segmentation scale 15 sections and unique multi-spectral picture Spectral data. The goal is to model the probability of trees presents wilt, conditioned on the image features. It contains 4339 training samples and 500 testing samples where training sample is highly imbalanced dataset. GLCM mean texture, mean green value), Mean red value, Mean NIR value and Standard deviation are the feature attributes in this dataset.

Autism Screening Adult

Autistic Spectrum Disorder (ASD) [42] is a neuro advancement condition related with critical medical care costs, and early determination can essentially decrease these. This dataset is identified with mental imbalance screening of grown-ups that contained 20 highlights to be used for additional examination particularly in deciding persuasive medically introverted qualities and improving the grouping of ASD cases. In this dataset, creators recorded ten social highlights (AQ-10-Adult) in addition to ten people attributes that have end up being compelling in recognizing the ASD cases from controls in conduct science. The source of the dataset is Fadi Fayez Thabtah, Department of Digital Technology, Manukau Institute of Technology, Auckland, New Zealand. It is a Multivariate/Sequential/Time-Series/Domain-Theory dataset which contains 20 feature attributes and 704 instances. It is a classification type dataset with binary type and having the missing attributes. The area of this dataset is Medical, health and social science.

White Wine Quality

This dataset [23] is a binary classification related data which contains 1482 samples with 11 feature attributes. It is a highly imbalance dataset which represents the quality of white wine.

Red Wine Quality

This dataset [18] is a binary classification related data which contains 1599 samples with 11 feature attributes. It is a highly imbalance dataset which represents the quality of white wine. It does not have any missing values.

QSAR Androgen Receptor

This dataset [43] is utilized to create arrangement QSAR models for the segregation of binder/positive (199) and non-binder/negative (1488) particles by methods for various AI techniques. Characteristics (molecular fingerprints) were determined at the Milano Chemometrics and QSAR Research Group on a bunch of synthetic substances gave by the National Center of Computational Toxicology, at the U.S. Natural Protection Agency in the system of the CoMPARA collective demonstrating project, which focused the improvement of QSAR models to distinguish folios to the Androgen Receptor. It contains 1024 feature attributes and one class attribute.

HCC Survival

HCC dataset [14] was acquired at a University Hospital in Portugal and contains a few segment, risk factors, research facility and full survival features of 165 genuine patients determined to have HCC. The dataset contains 49 highlights chose by the EASL-EORTC (European Association for the Study of the Liver - European Organization for Research and Treatment of Cancer) Clinical Practice Guidelines, which are the present status of-the-workmanship on the management of HCC. This heterogeneous dataset contains 23 quantitative factors, and 26 subjective factors. By and large, missing information speaks to 10.22% of the entire dataset and just eight patients have total data in all fields (4.85%). The objective factors are the endurance at 1 year, and was encoded as a twofold factor: 0 (bites the dust) and 1 (lives). A specific level of class-lopsidedness is additionally present (63 cases marked as sure and 102 as negative). A clear portrayal of the HCC dataset is given in [14]. Another cluster-based oversampling technique is applied in this dataset to improve endurance expectation of hepatocellular carcinoma patients.

Mesothelioma's disease

This is the medical related dataset [44] which contains the syndrome and outcome of Mesothelioma's disease. Dangerous mesotheliomas (MM) are forceful tumors of the pleura. These tumors are associated with asbestos introduction, it might also be identified with past simian infection 40 (SV40) contamination and very feasible for hereditary inclination. Molecular systems can also be ensnared in the advancement of mesothelioma. Provincial living is related with the advancement of mesothelioma. It contains 324 Instances and 33 feature attributes in this dataset which has no missing value.

Extention of Z-Alizadeh Sani

The Z-Alizadeh Sani dataset [45] contains the records of 303 patients where it contains 58 numbers of feature. As per clinical literature, outcome can be considered as markers of CAD for a patient. In any case, some of them have never been utilized in data mining-based methodologies for CAD determination. The highlights are organized in four gatherings: demographic, side effect and assessment, ECG, and research facility and reverberation highlights. Every patient could be in two potential classifications CAD or Normal. A patient is ordered as CAD, if his/her measurement narrowing is more prominent than or equivalent to half, and in any case as Normal.

QSAR Oral Toxicity

This dataset [46] is utilized to create arrangement QSAR models for the separation of extremely poisonous/positive (741) and not harmful/negative (8251) atoms by methods for various AI techniques. Attributes (atomic fingerprints) were determined at the Milano Chemometrics and QSAR Research Group on a bunch of synthetics gave by the ICCVAM Acute Toxicity Workgroup, as a team with the U.S. Natural Protection Agency (U.S. EPA, National Center for Computational Toxicology), which composed the Predictive Models for Acute Oral Systemic Toxicity community undertaking to create in silico models to foresee intense oral foundational poisonousness for filling administrative requirements. It contains 1024 binary atomic fingerprints and 1 trial class.

Climate Model Simulation Crashes

This data [47] utilizes the records of reenactment crashes experienced during atmosphere model uncertainty quantification (UQ) outfits. Various arrangements of irritated

boundary were executed as a feature of a wide exertion to evaluate and oblige vulnerabilities in the climatic, ocean ice, and sea model segments of CCSM4. The disappointments detailed here happened during recreations that bothered boundary esteems in the Parallel Ocean Program 20 (POP2), the sea segment of CCSM4. For these trials, POP2 was combined with the ocean ice model, while information-based parts were utilized for the land and air. The simulations were coordinated for 10 yr, and the framework was constrained with climatological air-ocean motion information utilizing typical year compelling from Large and Yeager (2009). Three separate Latin hypercube outfits were led, each containing 180 group individuals. 46 out of the 540 reenactments fizzled for numerical reasons at blends of parameter esteems. The objective is to utilize characterization to foresee recreation results (fall flat or succeed) from input parameter esteems, and to utilize affectability investigation and highlight determination to decide the reasons for reenactment crashes.

Electrical Grid Stability Simulated

Decentral Smart Grid Control (DSGC) [48] is another framework executing request reaction without huge changes of the foundation. The examination is performed for various arrangements of information esteems utilizing the technique like that portrayed in [49]. This dataset contains 10000 occurrences with 13 element attributes. $\tau[x]$ is the response season of member, $p[x]$ is the ostensible force devoured/delivered, $g[x]$ is the coefficient (gamma) corresponding to value flexibility, $stab$ is the maximal genuine piece of the trademark condition root and $stabf$ is the result section property.

QSAR Biodegradation

This dataset [50] is inherent from the Milano Chemometrics and QSAR Research team. The exploration prompting these outcomes has gotten subsidizing from the European Community Seventh Framework Program. The information has been utilized to create QSAR (Quantitative Structure Activity Relationships) models for the investigation of the connections between synthetic structure and biodegradation of particles. Biodegradation exploratory estimations of 1055 synthetic substances were gathered from the page of the National Institute of Technology and Evaluation of Japan (NITE). Grouping models were created to segregate prepared (356) and not prepared (699) biodegradable particles by methods for three separate modelling techniques.

Immunotherapy

The dataset exhibited with immunotherapy treatment technique [51] which comprise of 90 patients and having seven component ascribes in it with some numeric values in it. 71 of them are positive and 19 of them are negative which is an imbalance dataset but in very small amount. No of properties comprise on sexual orientation of the patients, age of the patients, in how long a patient be dealt with, no of moles of patient, which sort of moles are there, region of skin covered by moles, induration measurement (mm2) of moles and treatment result after usage of the immunotherapy treatment.

Parkinson's Disease

The information utilized in this examination [52] were accumulated from 188 patients with Parkinson's Disease (PD) (107 men and 81 ladies) with ages going from 33 to 87 at the Department of Neurology in Istanbul University. The control group comprises of 64 solid people (23 men and 41 ladies) with ages changing somewhere in the range of 41 and 82. During the information assortment measure, the mouthpiece is set to 44.1 KHz and following the doctor assessment. Different discourse signal handling calculations including Time Frequency Features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform based Features, Vocal Fold Features and TWQT highlights have been applied to the discourse accounts of Parkinson's Disease (PD) patients to separate clinically helpful data for PD evaluation.

Customer Churn Wireless Telecom

This competition [18] is tied in with anticipating whether a client will change media communications supplier, something known as "agitating". It is an unevenness dataset which contains 5000 cases with 19 feature attributes. It contains client data and nature. It also gives the possibility of the call duration of a client.

Sales: win - loss

IBM Watson Sales-Win-Loss dataset [18] contains the sales information of the supplier's groups. It gives to predict the outcome of sales win or loss for a particular information. It not only gives the analytics results but also gives the opportunities of the knowing the basic idea for a wining customer. It contains 78026 instances with 17 features attributes.

Table 5.1 represents the description of original datasets with Imbalance Ratio (IR). Table 5.1 also refers to the category of the datasets where mainly our datasets are categorized by Bio-Life, Business, Medical Research and Social criteria. Bio-Life datasets contain bio-informatics, structural, scientific research and computational analysis data. It is a life-related dataset with lots of variation. Business data contains economic and financial transactions from profitable organizations. Medical research is related to the survey of diseases and their prevention. Social criteria are similar to the social activities and behavior of human beings.

$$\text{ImbalanceRatio}(IR) = \frac{\text{MajorDataCount}}{\text{MinorDataCount}} \quad (5.1)$$

Table 5.1: Experimental dataset description with attribute number, class counts, and imbalance ratio

Dataset Name	Dataset Category	Attribute #	Class Attribute Count		IR
			Major	Minor	
Bank Tele-Marketing [35]	Business	19	36548	4640	7.877
Online Shoppers Purchasing [16]	Business	17	10422	1908	5.462
Default of Credit Card [36]	Business	23	23364	6636	3.521
Page Blocks [37]	Business	10	4913	559	8.789
Abalone [22]	Bio Life	8	4142	187	22.15
Seismic Bumps [38]	Business	18	2408	170	14.165
HR Employee Attrition [18]	Social	34	1233	237	5.203
Pima Indians Diabetes [39]	Medical Research	8	500	268	1.866
Cervical cancer (Risk Factors) [40]	Medical Research	32	803	55	14.6
Polish companies bankruptcy [21]	Business	64	12394	626	19.799
Blood Transfusion [41]	Medical Research	4	570	178	3.202
Ecoli 1 [10]	Bio Life	7	180	20	9
Ecoli 2 [10]	Bio Life	7	274	7	39.143
Wilt [20]	Bio Life	5	4265	74	57.635
Autism Screening Adult [42]	Social	19	515	189	2.725
White Wine Quality [23]	Business	11	1457	25	58.28
Red Wine Quality [18]	Business	11	1546	53	29.17
QSAR Androgen Receptor [43]	Bio Life	1024	1488	198	7.515
HCC Survival [14]	Medical Research	49	102	63	1.619
Mesothelioma's disease [44]	Medical Research	34	228	96	2.375
Extention of Z-Alizadeh Sani [45]	Medical Research	58	216	87	2.483
QSAR Oral Toxicity [46]	Bio Life	1024	8250	741	11.134
Climate Model Simulation Crashes [47]	Bio Life	18	494	46	10.739
Electrical Grid Stability Simulated [48]	Bio Life	13	6380	3620	1.762
QSAR Biodegradation [50]	Bio Life	41	699	356	1.963
Immunotherapy [51]	Medical Research	7	71	19	3.737
Parkinson's Disease [52]	Medical Research	21	564	192	2.938
Polish Companies Bankruptcy (Y1) [21]	Business	64	6756	271	24.93
Polish Companies Bankruptcy (Y2) [21]	Business	64	9773	400	24.432
Polish Companies Bankruptcy (Y3) [21]	Business	64	10008	495	20.218
Polish Companies Bankruptcy (Y4) [21]	Business	64	9277	515	18.014
Polish Companies Bankruptcy (Y5) [21]	Business	64	5500	410	13.415
Customer Churn Wireless Telecom [18]	Social	19	4293	707	6.072
Sales: win - loss [18]	Business	17	60398	17627	3.426

Experimental Setup and Performance Metrics

In this Thesis, we implemented our proposed scheme using three frameworks of Python Programming language. Scikit Learn [53] and Imbalanced-Learn [54] are two python frameworks which are used for data analysis, and machine-learning implementation and handling of imbalance problems, respectively. Using the Pandas framework [55], we manipulate and process our datasets to fit into a machine learning model. For evaluating

our model, we divided our datasets into the train-test splitting method where the training data are used for data balancing, and model creation and the testing data are used for model evaluation. As we use the train-test splitting method, the verification dataset becomes unused and we can validate our machine-learning model using the subset of the original dataset. For validation, we compare the performance of machine learning model with original data and the balanced data from our proposed approach. We use the following performance metrics for evaluation.

1. Accuracy (Acc):

Accuracy is an universal performance metric which is the ratio of correct result and total result count. Equation 5.2 represents the accuracy score where tp refers successfully detected positive class, tn refers successfully detected negative class, fp refers wrongly detected positive class and fn refers wrongly detected negative class.

$$Accuracy(Acc.) = \frac{(tp + tn)}{(tp + fp + fn + tn)} \quad (5.2)$$

2. Precision or Positive Predictive Value (PPV):

Precision is the ratio of correctly detected positive and total predicted positive observation. Equation 5.3 represents the Precision.

$$Precision = \frac{tp}{(tp + fp)} \quad (5.3)$$

3. Sensitivity (Se) or Recall or True Positive Rate (TPR):

Sensitivity [10] is the efficiency of positive class (Minor Class) of the dataset. Equation 5.4 represents the Sensitivity score.

$$Sensitivity(Se) = \frac{tp}{(tp + fn)} \quad (5.4)$$

4. Specificity (Sp) or True Negative Rate (TNR):

Specificity [10] is the efficiency of negative class (Major Class) of the dataset. Equation 5.5 represents the Sensitivity score.

$$Specificity(Sp) = \frac{tn}{(tn + fp)} \quad (5.5)$$

5. F-Measure or F-score:

F-Measure [10] is useful for imbalance dataset. It measures the balances of pre-

cision and recall (Sensitivity). It gives the balance property of the accurate and predicted result of positive class. Equation 5.6 represents the F-Measure.

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision + Recall)} \quad (5.6)$$

Where, Equation 5.3 and 5.4 refers the Precision and Recall, respectively.

6. Geometric Mean (G-Mean):

G-Mean [10] is the combine result of Sensitivity and Specificity. It is useful when dataset is imbalanced and needs relation of positive and negative class. Equation 5.7 represents the G-Mean.

$$G - Mean = \sqrt{Sensitivity * Specificity} \quad (5.7)$$

7. Receiver Operating Characteristics (ROC) Curve:

ROC is a curve which represent the Area Under Curve (AUC) and creates a relation between sensitivity and specificity. In ROC curve, the x-axis and y-axis represents fp and tp, respectively. Bigger AUC is better for classification algorithms.

The accuracy rate is a universal evaluation metric, but it gives bias results and cannot create a balance between major and minor classes. For this reason, other performance matrices are necessary to evaluate the machine learning algorithms.

Summary

In this section, we briefly discuss the datasets which we used in our thesis work. The description, source, category, attribute type, imbalance ratio (IR) of each dataset is explained in this section. We can also observe that we used four types of datasets in our thesis which are similar in their domain. Those are Business and financial types, Medical research, Social and Bio life in nature. We can also see the IR variation from our overall datasets. Our approach can handle the different degrees of imbalance which we can see from Table 5.1. Apart from this, we also discuss the environment of implementation for our thesis. We discuss the programming language applied for implementation and the library which are related to this implementation. For performance evaluation of the imbalance dataset, we need some special types of performance metrics. We discuss those metrics like G-Mean, Recall and F-Measure, etc. in this section. By those metrics, we can measure the efficiency of our proposed approach.

Chapter 6

Performance Evaluation

Result Analysis for Different Datasets

We examine our proposed scheme with various dimensions of the imbalanced datasets and get classification results using Random Forest and Neural Network models. Table 6.1 shows the percentage of major and minor class for Original Dataset and the percentage after applying our proposed scheme. As we can see in the table, our proposed scheme can balance the dataset efficiently. To evaluate classification performance, we calculate classification results for 5 scenarios: (i) without oversampling, (ii) ADASYN, (iii) SVM-SMOTE, (iv) SMOTE + ENN, and (v) proposed scheme. Table 6.2 gives the accuracy rate, F-Measure and AUC Score; Table 6.3 gives the Sensitivity, Specificity and G-Mean, respectively for Random Forest model. On the other hand, Table 6.4 gives the accuracy rate, F-Measure and AUC Score; Table 6.5 gives the Sensitivity, Specificity and G-Mean, respectively for ANN model. By observing the AUC scores in these 4 tables, we can see that our scheme gives better performance and balancing results for 32 of the 34 datasets for RF model and for 24 datasets in NN model.

Comparing our MLH Balancing Scheme results with original dataset, we can observe an irregular change for 5 of all datasets which we applied in our study. Default of Credit Card [36], QSAR Oral Toxicity [46], Extension of Z-Alizadeh Sani [45], Mesothelioma's disease [44] and HCC Survival [14] are the five datasets where information is decrease for our proposed approach. As our approach removes outlier data points and works well for high dimensional characteristics of the dataset, it gives bad results for using our proposed approach. Our oversampling approach boosting the dataset for minor class but give imperfect result for the major class of original dataset. Comparing with other oversampling approaches, our approach does not work well for these five

datasets. But overall it works well for the others dataset. Figure 6.1 represent the visual representation for information reduction of our approach comparing with original dataset and existing oversampling techniques. Figure 6.2 gives the brief idea of major and minor class condition for the 5 dataset.

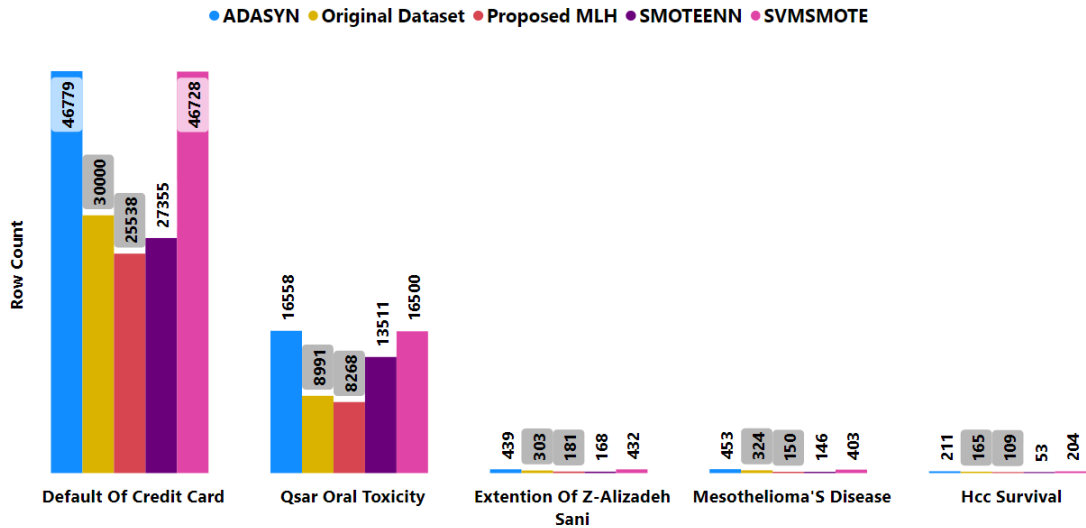


Figure 6.1: Data Reduce for five dataset using MLH Balancing Scheme

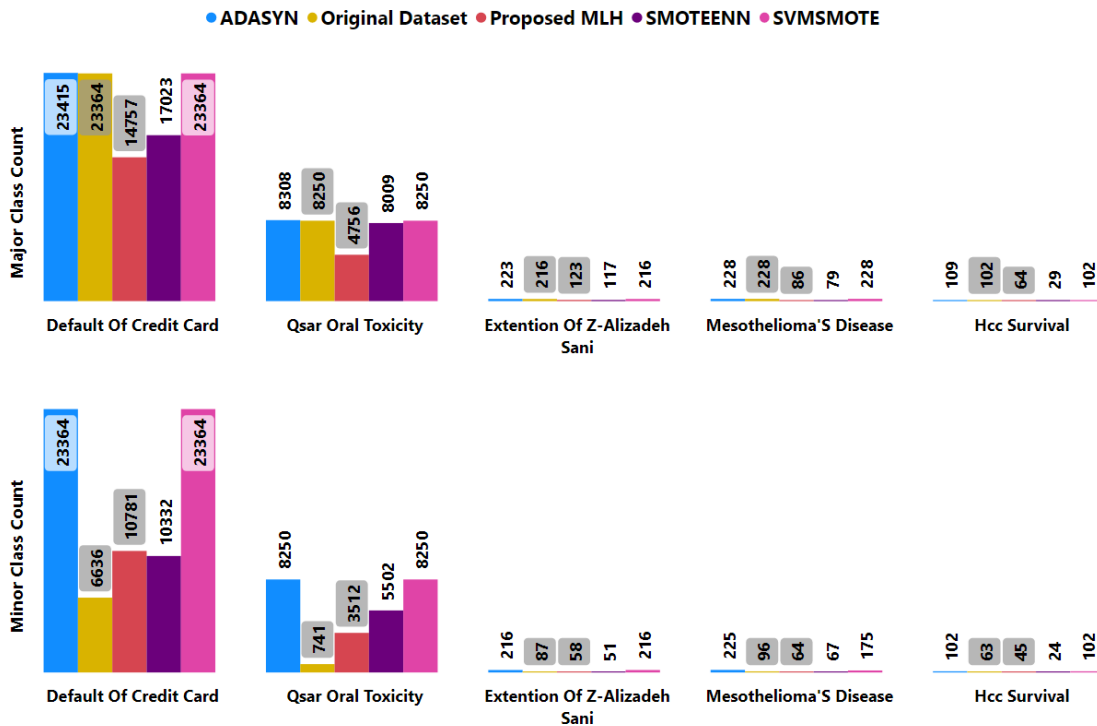


Figure 6.2: Class attribute count for five dataset

We can also observe the difference between major and minor class is quiet most lower

then others oversampling techniques in our study. As our approach reduce outliers from the generated output, it creates some larger difference between major and minor class count percentage. Although our approach creating some distance in count percentages than existing oversampling techniques, it gives a outlier free balance dataset after oversampling. Figure 6.3 and 6.4 gives the visual representation of count percentage difference for all dataset.

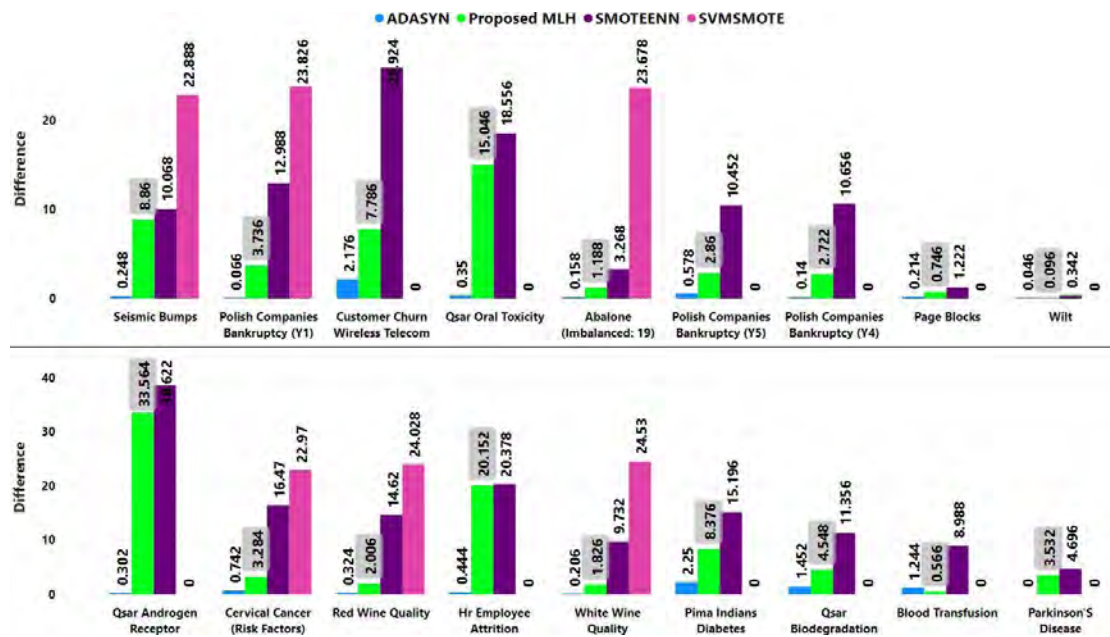


Figure 6.3: Count Percentage Difference (Part 1)

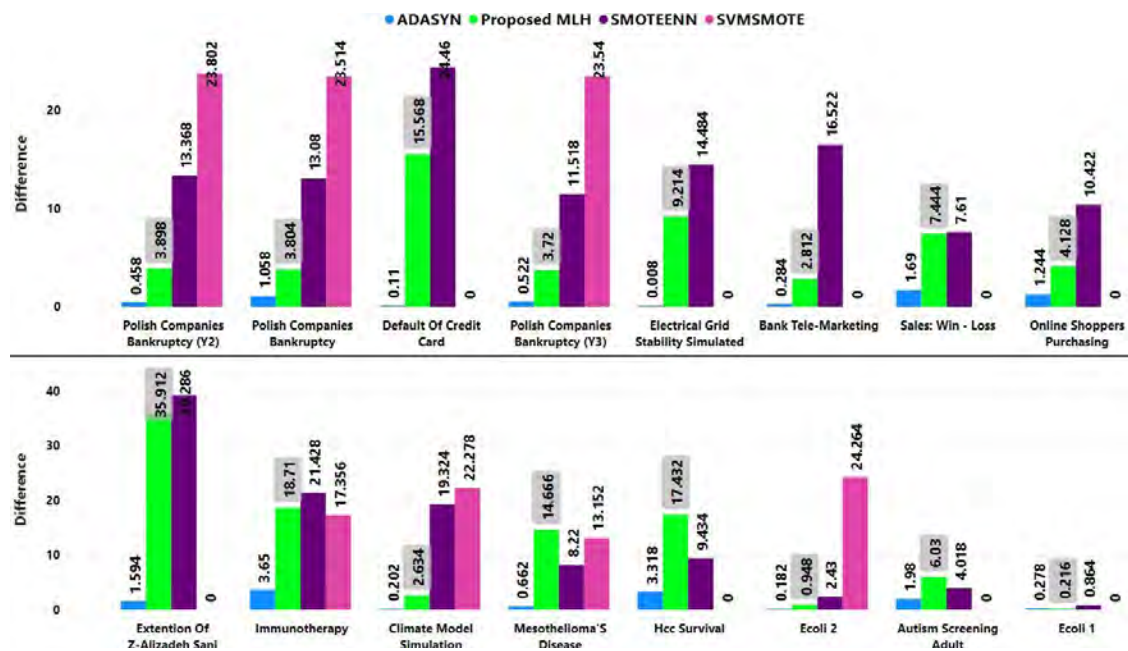


Figure 6.4: Count Percentage Difference (Part 2)

If we compare Random Forest with Neural Network, we can observe that Random Forest gives a better result than Neural Network. Figure 6.5 and 6.6 represent the average performance matrices for Random Forest and Neural Network. Here, X-axis represents Balancing Techniques and Y-axis represents Average Performance Matrices. Our scheme gives better performance for the Random Forest algorithm. From figure 6.5, we can observe that the average-Accuracy and the average-Sensitivity are positively correlated with each other. At 91.223%, Multi-Layer Hybrid (MLH) Balancing Technique has the highest average-Accuracy and is 11.27% higher than Original Dataset, which has the lowest average accuracy of 81.982%. By observing G-Mean value (Balancing Property) we can say that our scheme gives a balancing result for both classes.

Table 6.1: Percentage of major, and minor classes for original datasets and the percentage after applying the MLH balancing scheme on the original datasets.

Dataset Name	Original Dataset		MLH Balancing Scheme	
	Major Class	Minor Class	Major Class	Minor Class
Bank Tele-Marketing [35]	88.74%	11.27%	48.59%	51.41%
Online Shoppers Purchasing [16]	84.53%	15.47%	47.94%	52.06%
Default of Credit Card [36]	77.88%	22.12%	42.22%	57.78%
Page Blocks [37]	89.78%	10.22%	49.63%	50.37%
Abalone [22]	95.68%	4.32%	49.41%	50.59%
Seismic Bumps [38]	93.41%	6.59%	45.57%	54.43%
HR Employee Attrition [18]	83.88%	16.12%	39.92%	60.08%
Pima Indians Diabetes [39]	65.10%	34.90%	45.81%	54.19%
Cervical cancer (Risk Factors) [40]	93.59%	6.41%	48.36%	51.64%
Polish companies bankruptcy [21]	95.19%	4.81%	48.10%	51.90%
Blood Transfusion [41]	76.20%	23.80%	49.72%	50.28%
Ecoli 1 [10]	90.00%	10.00%	49.89%	50.11%
Ecoli 2 [10]	97.51%	2.49%	49.53%	50.47%
Wilt [20]	98.30%	1.71%	49.95%	50.05%
Autism Screening Adult [42]	73.15%	26.85%	46.99%	53.02%
White Wine Quality [23]	98.31%	1.69%	50.91%	49.09%
Red Wine Quality [18]	96.69%	3.32%	49.00%	51.00%
QSAR Androgen Receptor [43]	88.26%	11.74%	33.22%	66.78%
HCC Survival [14]	61.82%	38.18%	58.72%	41.28%
Mesothelioma's disease [44]	70.37%	29.63%	42.67%	57.33%
Extention of Z-Alizadeh Sani [45]	71.29%	28.71%	32.04%	67.96%
QSAR Oral Toxicity [46]	91.76%	8.24%	57.52%	42.48%
Climate Model Simulation Crashes [47]	91.48%	8.52%	48.68%	51.32%
Electrical Grid Stability Simulated [48]	63.80%	36.20%	45.39%	54.61%
QSAR Biodegradation [50]	66.26%	33.74%	47.73%	52.27%
Immunotherapy [51]	78.89%	21.11%	40.65%	59.36%
Parkinson's Disease [52]	74.60%	25.40%	48.23%	51.77%
Polish Companies Bankruptcy (Y1) [21]	96.14%	3.86%	48.13%	51.87%
Polish Companies Bankruptcy (Y2) [21]	96.07%	3.93%	48.05%	51.95%
Polish Companies Bankruptcy (Y3) [21]	95.29%	4.71%	48.14%	51.86%
Polish Companies Bankruptcy (Y4) [21]	94.74%	5.26%	48.64%	51.36%
Polish Companies Bankruptcy (Y5) [21]	93.06%	6.94%	48.57%	51.43%
Customer Churn Wireless Telecom [18]	85.86%	14.14%	46.11%	53.89%
Sales: win - loss [18]	77.41%	22.59%	46.28%	53.72%

Table 6.2: Accuracy, F-Measure and AUC Score for existing and proposed approach using Random Forest. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.

Dataset Name	Accuracy					F-Measure					AUC Score				
	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°
Bank Tele-Marketing	71.05%	73.05%	78.66%	74.31%	94.31%	44.01%	50.61%	68.79%	53.97%	92.83%	80.16%	80.17%	80.63%	80.54%	98.12%
Online Shoppers Purchasing	86.40%	87.74%	86.40%	87.38%	95.34%	66.26%	74.87%	74.13%	73.92%	90.64%	90.99%	90.02%	89.87%	90.49%	98.56%
Default of Credit Card	70.04%	72.42%	70.92%	73.98%	81.74%	48.22%	58.82%	65.32%	60.30%	77.74%	78.56%	76.96%	78%	78.85%	88.72%
Page Blocks	97.69%	97.06%	97.48%	97.48%	97.69%	92.99%	91.76%	93.10%	92.86%	93.33%	99.05%	99.43%	99.03%	99.58%	99.70%
Abalone	79.92%	77.56%	82.28%	79.13%	93.31%	0%	9.52%	60.18%	10.17%	82.10%	76.43%	75.30%	85.28%	76.27%	94.23%
Seismic Bumps	76.82%	80%	73.64%	77.73%	85%	0%	35.29%	44.23%	22.22%	72.27%	75.82%	71.16%	71.23%	71.85%	92.20%
HR Employee Attrition	79.66%	76.31%	70.86%	78.62%	83.65%	18.49%	27.10%	35.35%	27.14%	65.18%	66.51%	65.76%	68.36%	64.95%	86.26%
Pima Indians Diabetes	66.35%	77.88%	81.73%	72.12%	86.54%	58.82%	77.67%	83.19%	72.38%	87.27%	81.32%	81.54%	86.17%	80.59%	92.72%
Cervical cancer (Risk Factors)	83.33%	82.29%	81.25%	82.29%	90.62%	0%	0%	10%	0%	64%	49.53%	54.65%	54.30%	58.83%	82.07%
Polish Companies Bankruptcy	77.23%	79.83%	80.94%	78.71%	92.70%	6.12%	30.04%	49.01%	18.87%	81.73%	84.02%	84.39%	83.92%	83.54%	96.76%
Blood Transfusion	69.23%	62.31%	70.77%	66.92%	76.15%	35.48%	44.94%	57.78%	48.19%	68.69%	65.70%	62.43%	65.29%	67.26%	78.90%
Ecoli 1	97.83%	97.83%	93.48%	97.83%	97.83%	94.74%	95.24%	86.96%	94.74%	95.24%	100%	99.72%	99.72%	100%	100%
Ecoli 2	96.43%	100%	96.43%	96.43%	100%	0%	100%	66.67%	0%	100%	100%	100%	100%	100%	100%
Wilt	62.80%	84%	84%	84.80%	86.20%	1.06%	75%	76.88%	77.65%	79.40%	80.05%	94.01%	92.10%	92.42%	94.46%
Autism Screening Adult	94.97%	95.53%	95.53%	96.09%	98.32%	93.88%	94.67%	94.94%	95.42%	98.06%	99.64%	99.58%	99.20%	99.50%	99.95%
White Wine Quality	87.95%	86.75%	85.54%	89.16%	95.18%	0%	15.38%	14.29%	18.18%	80%	61.51%	55.68%	41.64%	72.60%	86.44%
Red Wine Quality	85.56%	85.56%	82.22%	85.56%	96.67%	0%	38.10%	27.27%	13.33%	88%	73.48%	77.52%	68.83%	70.88%	94.36%
QSAR Androgen Receptor	83.73%	83.08%	77.22%	83.95%	75.05%	31.19%	30.36%	45.03%	33.93%	54.54%	83.90%	75.54%	75.30%	77.43%	87.13%
HCC Survival	71.43%	76.19%	69.05%	80.95%	83.33%	50%	61.54%	64.86%	73.33%	80%	78.85%	89.90%	81.85%	82.93%	90.62%
Mesothelioma's disease	100%	100%	98.90%	98.90%	97.80%	100%	100%	98.55%	98.55%	96.97%	100%	100%	99.97%	100%	100%
Extention of Z-Alizadeh Sani	95.38%	93.85%	95.38%	92.31%	96.92%	92.68%	91.67%	93.62%	89.80%	95.65%	99.79%	99.84%	100%	98.68%	100%
QSAR Oral Toxicity	86.96%	88.12%	86.14%	88.78%	89.60%	47.68%	55.56%	62.83%	60%	74.90%	89.06%	83.86%	84.86%	86.82%	94.56%
Climate Model Simulation Crashes	80.44%	82.61%	85.87%	83.70%	95.65%	0%	33.33%	62.86%	40%	87.50%	82.02%	86.75%	90.77%	89.41%	98.57%
Electrical Grid Stability Simulated	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
QSAR Biodegradation	87.11%	86.08%	84.54%	88.14%	94.84%	84.85%	84.39%	83.33%	86.55%	94.44%	93.97%	93.17%	94.05%	92.65%	98.23%
Immunotherapy	74.07%	70.37%	66.67%	77.78%	77.78%	36.36%	50%	60.87%	57.14%	62.50%	94.75%	81.79%	66.67%	81.17%	76.85%
Parkinson's Disease	69.74%	73.03%	75%	75%	79.60%	52.08%	65.55%	71.21%	67.80%	75.97%	83.73%	84.96%	81.13%	86.36%	86.25%
Polish Companies Bankruptcy (Y1)	79.77%	81.71%	79.38%	80.54%	95.33%	7.14%	27.69%	34.57%	19.36%	88%	67.77%	70.15%	67.70%	65.58%	97.53%
Polish Companies Bankruptcy (Y2)	78.55%	79.36%	79.89%	78.55%	94.64%	2.44%	15.38%	38.02%	2.44%	85.92%	76.03%	73.74%	68.94%	71.44%	94.88%
Polish Companies Bankruptcy (Y3)	75.94%	76.69%	79.70%	76.44%	94.49%	7.69%	21.85%	43.36%	11.32%	87.78%	76.96%	80.80%	81.79%	79.70%	97.32%
Polish Companies Bankruptcy (Y4)	74.28%	75.33%	77.16%	73.23%	93.18%	9.26%	29.85%	52.46%	19.05%	86.60%	78.45%	78.26%	77.50%	79.47%	95.43%
Polish Companies Bankruptcy (Y5)	71.26%	79.76%	79.35%	78.95%	93.93%	26.80%	59.68%	63.31%	57.38%	90.45%	86.54%	86.46%	84.44%	87.24%	97.34%
Customer Churn Wireless Telecom	93.92%	93.76%	89.52%	93.36%	96.72%	73.61%	76.36%	67.65%	73.82%	88.83%	90.04%	91.40%	90.54%	90.60%	97.99%
Sales: win - loss	81.54%	83.18%	83.29%	83.49%	91.47%	70.27%	75.62%	77.66%	75.87%	88.64%	90.87%	90.81%	90.57%	91.04%	96.82%

Table 6.3: Sensitivity, Specificity and G-Mean for existing and proposed approach using Random Forest. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.

Dataset Name	Sensitivity					Specificity					G-Mean				
	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°
Bank Tele-Marketing	29.31%	35.56%	60.56%	38.79%	94.83%	97.54%	96.85%	90.15%	96.85%	93.98%	53.47%	58.69%	73.89%	61.30%	94.40%
Online Shoppers Purchasing	57.07%	78.01%	83.25%	76.44%	96.34%	95.36%	90.72%	87.36%	90.72%	95.04%	73.77%	84.13%	85.28%	83.27%	95.68%
Default of Credit Card	33.58%	47.44%	65.96%	47.59%	76.81%	95.94%	90.16%	74.44%	92.73%	85.24%	56.76%	65.40%	70.07%	66.43%	80.91%
Page Blocks	86.90%	92.86%	96.43%	92.86%	91.67%	100%	97.96%	97.71%	98.47%	98.98%	93.22%	95.38%	97.07%	95.62%	95.25%
Abalone	0%	6.38%	72.34%	6.38%	82.98%	0%	93.72%	84.54%	95.65%	95.65%	0%	24.46%	78.20%	24.71%	89.09%
Seismic Bumps	0%	23.53%	45.10%	13.72%	84.31%	0%	97.04%	82.25%	97.04%	85.21%	0%	47.78%	60.90%	36.50%	84.76%
HR Employee Attrition	10.28%	19.63%	35.51%	17.76%	68.22%	99.73%	92.70%	81.08%	96.22%	88.11%	32.02%	42.65%	53.66%	41.33%	77.53%
Pima Indians Diabetes	46.30%	74.07%	87.04%	70.37%	88.89%	88%	82%	76%	74%	84%	63.83%	77.94%	81.33%	72.16%	86.41%
Cervical cancer (Risk Factors)	0%	0%	6.25%	0%	50%	0%	0%	96.25%	0%	98.75%	0%	0%	24.53%	0%	70.27%
Polish Companies Bankruptcy	3.19%	18.62%	39.36%	10.64%	70.21%	99.68%	98.39%	93.55%	99.36%	99.52%	17.84%	42.80%	60.68%	32.51%	83.59%
Blood Transfusion	25%	45.46%	59.09%	45.46%	77.27%	91.86%	70.93%	76.74%	77.91%	75.58%	47.92%	56.78%	67.34%	59.51%	76.42%
Ecoli 1	90%	100%	100%	90%	100%	100%	97.22%	91.67%	100%	97.22%	94.87%	98.60%	95.74%	94.87%	98.60%
Ecoli 2	0%	100%	100%	0%	100%	0%	100%	96.30%	0%	100%	0%	100%	98.13%	0%	100%
Wilt	0.54%	64.17%	71.12%	70.59%	71.12%	100%	95.85%	91.69%	93.29%	95.21%	7.31%	78.43%	80.76%	81.15%	82.29%
Autism Screening Adult	90.79%	93.42%	98.68%	96.05%	100%	98.06%	97.09%	93.20%	96.12%	97.09%	94.35%	95.24%	95.90%	96.08%	98.53%
White Wine Quality	0%	10%	10%	10%	80%	0%	97.26%	95.89%	100%	97.26%	0%	31.19%	30.97%	31.62%	88.21%
Red Wine Quality	0%	30.77%	23.08%	7.69%	84.62%	0%	94.80%	92.21%	98.70%	98.70%	0%	54.01%	46.13%	27.55%	91.39%
QSAR Androgen Receptor	19.10%	19.10%	48.32%	21.35%	77.53%	99.19%	98.39%	84.14%	98.92%	74.46%	43.53%	43.35%	63.76%	45.96%	75.98%
HCC Survival	37.50%	50%	75%	68.75%	87.50%	92.31%	92.31%	65.38%	88.46%	80.77%	58.84%	67.94%	70.03%	77.98%	84.07%
Mesothelioma's disease	100%	100%	100%	100%	94.12%	100%	100%	98.25%	98.25%	100%	100%	100%	99.12%	99.12%	97.01%
Extention of Z-Alizadeh Sani	86.36%	100%	100%	100%	100%	100%	90.70%	93.02%	88.37%	95.35%	92.93%	95.24%	96.45%	94.01%	97.65%
QSAR Oral Toxicity	32.43%	40.54%	63.96%	45.95%	84.68%	99.19%	98.79%	91.11%	98.38%	90.71%	56.72%	63.28%	76.34%	67.23%	87.64%
Climate Model Simulation Crashes	0%	22.22%	61.11%	27.78%	77.78%	0%	97.30%	91.89%	97.30%	100%	0%	46.50%	74.94%	51.99%	88.19%
Electrical Grid Stability Simulated	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
QSAR Biodegradation	78.65%	82.02%	84.27%	83.15%	95.51%	94.29%	89.52%	84.76%	92.38%	94.29%	86.12%	85.69%	84.52%	87.64%	94.89%
Immunotherapy	22.22%	44.44%	77.78%	44.44%	55.56%	100%	83.33%	61.11%	94.44%	88.89%	47.14%	60.86%	68.94%	64.79%	70.27%
Parkinson's Disease	37.31%	58.21%	70.15%	59.70%	73.13%	95.29%	84.71%	78.82%	87.06%	84.71%	59.63%	70.22%	74.36%	72.09%	78.71%
Polish Companies Bankruptcy (Y1)	3.70%	16.67%	25.93%	11.11%	81.48%	100%	99.02%	93.60%	99.02%	99.02%	19.24%	40.62%	49.26%	33.17%	89.82%
Polish Companies Bankruptcy (Y2)	1.25%	8.75%	28.75%	1.25%	76.25%	99.66%	98.64%	93.86%	99.66%	99.66%	11.16%	29.38%	51.95%	11.16%	87.17%
Polish Companies Bankruptcy (Y3)	4.04%	13.13%	31.31%	6.06%	79.80%	99.67%	97.67%	95.67%	99.67%	99.33%	20.07%	35.81%	54.73%	24.58%	89.03%
Polish Companies Bankruptcy (Y4)	4.85%	19.42%	46.60%	11.65%	81.55%	100%	96.04%	88.49%	96.04%	97.48%	22.03%	43.18%	64.22%	33.45%	89.16%
Polish Companies Bankruptcy (Y5)	15.85%	45.12%	53.66%	42.68%	86.58%	98.79%	96.97%	92.12%	96.97%	97.58%	39.58%	66.15%	70.31%	64.34%	91.92%
Customer Churn Wireless Telecom	59.89%	71.19%	77.40%	66.10%	92.09%	99.53%	97.48%	91.52%	97.86%	97.48%	77.21%	83.30%	84.16%	80.43%	94.75%
Sales: win - loss	59.21%	70.81%	78.81%	70.41%	90.30%	94.57%	90.40%	85.91%	91.13%	92.15%	74.83%	80.01%	82.28%	80.10%	91.22%

Table 6.4: Accuracy, F-Measure and AUC Score for existing and proposed approach using Neural Network. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.

Dataset Name	Accuracy					F-Measure					AUC Score				
	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°
Bank Tele-Marketing	68.62%	73.39%	73.89%	74.39%	76.32%	33.39%	68.52%	68.74%	67.38%	67.95%	78.27%	78.33%	78.24%	78.31%	80.32%
Online Shoppers Purchasing	84.07%	86.15%	74.63%	87.13%	86.52%	56.08%	73.29%	60.42%	74.33%	74.06%	88.82%	89.84%	87.93%	89.67%	91.28%
Default of Credit Card	62.60%	63.92%	61.98%	66.10%	65.60%	23.53%	59.22%	63.02%	58.63%	61.96%	71.58%	70.04%	70.52%	71.60%	71.50%
Page Blocks	96.02%	93.92%	93.50%	95.81%	92.45%	87.42%	84.97%	83.60%	88.89%	80.85%	98.87%	98.60%	98.87%	99.26%	98.43%
Abalone	81.50%	81.89%	80.71%	83.46%	83.07%	0%	64.06%	62.60%	65%	65.60%	88.43%	90.71%	90.48%	90.17%	92.10%
Seismic Bumps	76.82%	68.18%	61.82%	72.73%	57.73%	0%	46.97%	45.46%	46.43%	46.86%	66.24%	73.29%	71.88%	72.94%	72.82%
HR Employee Attrition	77.57%	54.93%	47.59%	73.17%	50.73%	0%	43.57%	41.59%	33.33%	41.40%	60.93%	65.41%	65.22%	65.56%	65.44%
Pima Indians Diabetes	72.12%	71.15%	80.77%	74.04%	84.62%	65.06%	72.22%	82.14%	73.79%	85.18%	82.89%	80.37%	88.93%	82.11%	92.63%
Cervical cancer (Risk Factors)	83.33%	72.92%	71.88%	81.25%	77.08%	0%	0%	30.77%	0%	31.25%	51.68%	45.98%	57.23%	49.57%	66.29%
Polish Companies Bankruptcy	76.73%	72.77%	69.31%	77.23%	73.76%	0%	54.73%	53.03%	55.56%	57.77%	72.16%	75.67%	76.17%	76.47%	81.82%
Blood Transfusion	66.15%	59.23%	56.92%	60.77%	60%	0%	58.27%	56.92%	59.20%	57.38%	72.25%	72.01%	72.09%	72.33%	71.93%
Ecoli 1	78.26%	93.48%	89.13%	89.13%	93.48%	0%	86.96%	76.19%	70.59%	86.96%	59.17%	96.94%	95.83%	95.83%	97.22%
Ecoli 2	96.43%	85.71%	89.29%	100%	100%	0%	33.33%	40%	100%	100%	100%	100%	100%	100%	100%
Wilt	64.40%	81.80%	87.80%	86.40%	85.60%	9.18%	73.62%	84.32%	82.10%	81.15%	95.11%	91.70%	94.49%	94.39%	94.32%
Autism Screening Adult	59.22%	84.36%	86.03%	89.94%	87.71%	7.60%	81.58%	84.47%	88.31%	87.36%	89.74%	93.04%	90.43%	95.78%	96.54%
White Wine Quality	87.95%	65.06%	57.83%	89.16%	74.70%	0%	21.62%	18.60%	30.77%	32.26%	54.52%	49.86%	49.45%	49.04%	70.82%
Red Wine Quality	85.56%	81.11%	73.33%	82.22%	81.11%	0%	45.16%	47.83%	33.33%	51.43%	46.95%	82.92%	83.52%	67.23%	81.62%
QSAR Androgen Receptor	85.25%	84.38%	74.40%	85.25%	74.62%	48.48%	47.06%	45.37%	49.25%	55.51%	79.59%	70.31%	73.28%	73.48%	85.99%
HCC Survival	61.90%	64.29%	64.29%	66.67%	42.86%	11.11%	51.61%	44.44%	46.15%	57.14%	60.58%	58.17%	78.36%	61.30%	79.57%
Mesothelioma's disease	93.41%	96.70%	81.32%	95.60%	82.42%	90.62%	95.78%	78.48%	94.29%	80%	99.43%	99.38%	89.63%	99.43%	92.88%
Extention of Z-Alizadeh Sani	93.85%	87.69%	70.77%	92.31%	76.92%	90%	84%	69.84%	89.36%	74.58%	99.68%	99.05%	95.03%	99.05%	98.52%
QSAR Oral Toxicity	87.46%	86.80%	83.50%	86.47%	89.77%	58.24%	56.04%	56.90%	56.84%	75.59%	83.50%	79.21%	82.16%	78.53%	92.22%
Climate Model Simulation Crashes	80.44%	83.70%	86.96%	80.44%	93.48%	0%	54.54%	71.43%	35.71%	85%	80.63%	86.41%	90.65%	88.14%	94.37%
Electrical Grid Stability Simulated	98.97%	99.08%	92.66%	98.74%	98.28%	98.75%	98.90%	91.53%	98.50%	97.97%	99.97%	99.97%	98.33%	99.96%	99.98%
QSAR Biodegradation	86.60%	82.47%	84.54%	84.02%	86.60%	84.52%	81.52%	83.70%	83.06%	85.87%	93.44%	93.17%	92.58%	93.62%	95.08%
Immunotherapy	66.67%	70.37%	51.85%	62.96%	59.26%	0%	60%	48%	37.50%	47.62%	79.01%	64.20%	51.85%	65.43%	62.96%
Parkinson's Disease	66.45%	71.05%	73.03%	77.63%	86.84%	40%	65.62%	70.07%	74.24%	85.29%	83.79%	81.44%	82.97%	82.20%	93.36%
Polish Companies Bankruptcy (Y1)	78.99%	66.93%	67.70%	75.10%	79.77%	0%	42.95%	42.76%	40.74%	60%	58.31%	70.34%	68.27%	65.23%	82.06%
Polish Companies Bankruptcy (Y2)	78.55%	64.34%	61.66%	70.51%	73.73%	0%	40.89%	41.15%	37.50%	56.64%	61.78%	66.81%	66.14%	63.08%	82.92%
Polish Companies Bankruptcy (Y3)	75.19%	77.19%	73.18%	78.20%	82.46%	0%	57.67%	59.62%	53.97%	68.18%	75.57%	76.42%	80.62%	78.30%	88.03%
Polish Companies Bankruptcy (Y4)	72.97%	61.94%	66.14%	72.97%	77.95%	0%	46.89%	51.32%	43.09%	64.10%	70.21%	67.79%	71.78%	68.10%	85.07%
Polish Companies Bankruptcy (Y5)	67.61%	73.68%	73.28%	76.92%	79.76%	6.98%	61.54%	62.92%	60.14%	72.22%	80.38%	75.65%	80.11%	77.07%	88.46%
Customer Churn Wireless Telecom	88.08%	75.28%	76.32%	86%	81.44%	30.70%	44.72%	48.79%	56.14%	56.23%	73.56%	80.78%	85.05%	84.99%	88.27%
Sales: win - loss	75.87%	78.64%	77.70%	72.45%	78.22%	57.78%	72.89%	71.60%	70.34%	71.86%	84.07%	86.26%	85.31%	85.22%	86.61%

Table 6.5: Sensitivity, Specificity and G-Mean for existing and proposed approach using Neural Network. Here, 1° = Original Dataset, 2° = ADASYN, 3° = SMOTE + ENN, 4° = SVM-SMOTE and 5° = Proposed Approach.

Dataset Name	Sensitivity					Specificity					G-Mean				
	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°	1°	2°	3°	4°	5°
Bank Tele-Marketing	20.26%	74.57%	73.92%	68.10%	64.66%	99.32%	72.64%	73.87%	78.39%	83.72%	44.86%	73.60%	73.90%	73.06%	73.57%
Online Shoppers Purchasing	43.46%	81.15%	82.72%	79.58%	82.20%	96.48%	87.68%	72.16%	89.44%	87.84%	64.75%	84.35%	77.26%	84.37%	84.97%
Default of Credit Card	13.86%	63.10%	78.01%	57.83%	67.47%	97.22%	64.49%	50.59%	71.98%	64.28%	36.70%	63.79%	62.82%	64.52%	65.86%
Page Blocks	78.57%	97.62%	94.05%	95.24%	90.48%	99.75%	93.13%	93.38%	95.93%	92.88%	88.53%	95.35%	93.72%	95.58%	91.67%
Abalone	0%	87.23%	87.23%	82.98%	87.23%	0%	80.68%	79.23%	83.58%	82.13%	0%	83.89%	83.13%	83.28%	84.64%
Seismic Bumps	0%	60.78%	68.63%	50.98%	80.39%	0%	70.41%	59.76%	79.29%	50.89%	0%	65.42%	64.04%	63.58%	63.96%
HR Employee Attrition	0%	77.57%	83.18%	29.91%	77.57%	0%	48.38%	37.30%	85.68%	42.97%	0%	61.26%	55.70%	50.62%	57.74%
Pima Indians Diabetes	50%	72.22%	85.18%	70.37%	85.18%	96%	70%	76%	78%	84%	69.28%	71.10%	80.46%	74.09%	84.59%
Cervical cancer (Risk Factors)	0%	0%	37.50%	0%	31.25%	0%	0%	78.75%	0%	86.25%	0%	0%	54.34%	0%	51.92%
Polish Companies Bankruptcy	0%	70.74%	74.47%	61.17%	77.13%	0%	73.39%	67.74%	82.10%	72.74%	0%	72.05%	71.02%	70.86%	74.90%
Blood Transfusion	0%	84.09%	84.09%	84.09%	79.54%	0%	46.51%	43.02%	48.84%	50%	0%	62.54%	60.15%	64.08%	63.07%
Ecoli 1	0%	100%	80%	60%	100%	0%	91.67%	91.67%	97.22%	91.67%	0%	95.74%	85.64%	76.38%	95.74%
Ecoli 2	0%	100%	100%	100%	100%	0%	85.18%	88.89%	100%	100%	0%	92.30%	94.28%	100%	100%
Wilt	4.81%	67.91%	87.70%	83.42%	82.89%	100%	90.10%	87.86%	88.18%	87.22%	21.94%	78.22%	87.78%	85.77%	85.03%
Autism Screening Adult	3.95%	81.58%	89.47%	89.47%	100%	100%	86.41%	83.50%	90.29%	78.64%	19.87%	83.96%	86.43%	89.88%	88.68%
White Wine Quality	0%	40%	40%	20%	50%	0%	68.49%	60.27%	98.63%	78.08%	0%	52.34%	49.10%	44.41%	62.48%
Red Wine Quality	0%	53.85%	84.62%	30.77%	69.23%	0%	85.71%	71.43%	90.91%	83.12%	0%	67.94%	77.74%	52.89%	75.86%
QSAR Androgen Receptor	35.96%	35.96%	55.06%	37.08%	82.02%	97.04%	95.97%	79.03%	96.77%	72.85%	59.07%	58.74%	65.96%	59.90%	77.30%
HCC Survival	6.25%	50%	37.50%	37.50%	100%	96.15%	73.08%	80.77%	84.62%	7.69%	24.52%	60.45%	55.04%	56.33%	27.74%
Mesothelioma's disease	85.29%	100%	91.18%	97.06%	94.12%	98.25%	94.74%	75.44%	94.74%	75.44%	91.54%	97.33%	82.94%	95.89%	84.26%
Extention of Z-Alizadeh Sani	81.82%	95.46%	100%	95.46%	100%	100%	83.72%	55.81%	90.70%	65.12%	90.45%	89.40%	74.71%	93.05%	80.70%
QSAR Oral Toxicity	47.75%	45.95%	59.46%	48.65%	86.49%	96.36%	95.96%	88.89%	94.95%	90.50%	67.83%	66.40%	72.70%	67.96%	88.47%
Climate Model Simulation Crashes	0%	50%	83.33%	27.78%	94.44%	0%	91.89%	87.84%	93.24%	93.24%	0%	67.78%	85.56%	50.89%	93.84%
Electrical Grid Stability Simulated	98.34%	99.45%	95.58%	100%	100%	99.41%	98.82%	90.59%	97.84%	97.06%	98.88%	99.14%	93.05%	98.92%	98.52%
QSAR Biodegradation	79.78%	84.27%	86.52%	85.39%	88.76%	92.38%	80.95%	82.86%	82.86%	84.76%	85.85%	82.59%	84.67%	84.12%	86.74%
Immunotherapy	0%	66.67%	66.67%	33.33%	55.56%	0%	72.22%	44.44%	77.78%	61.11%	0%	69.39%	54.43%	50.92%	58.27%
Parkinson's Disease	25.37%	62.69%	71.64%	73.13%	86.57%	98.82%	77.65%	74.12%	81.18%	87.06%	50.08%	69.77%	72.87%	77.05%	86.81%
Polish companies bankruptcy (Y1)	0%	59.26%	57.41%	40.74%	72.22%	0%	68.97%	70.44%	84.24%	81.77%	0%	63.93%	63.59%	58.58%	76.85%
Polish companies bankruptcy (Y2)	0%	57.50%	62.50%	41.25%	80%	0%	66.21%	61.43%	78.50%	72.01%	0%	61.70%	61.96%	56.90%	75.90%
Polish companies bankruptcy (Y3)	0%	62.63%	79.80%	51.52%	75.76%	0%	82%	71%	87%	84.67%	0%	71.66%	75.27%	66.95%	80.09%
Polish companies bankruptcy (Y4)	0%	62.14%	66.02%	37.86%	72.82%	0%	61.87%	66.19%	85.97%	79.86%	0%	62%	66.10%	57.06%	76.26%
Polish companies bankruptcy (Y5)	3.66%	63.42%	68.29%	52.44%	79.27%	99.39%	78.79%	75.76%	89.09%	80%	19.07%	70.68%	71.93%	68.35%	79.63%
Customer Churn Wireless Telecom	18.64%	70.62%	79.66%	63.28%	84.18%	99.53%	76.05%	75.77%	89.75%	80.99%	43.08%	73.28%	77.69%	75.36%	82.57%
Sales: win - loss	44.79%	77.93%	76.28%	88.62%	75.43%	94.01%	79.06%	78.53%	63.01%	79.85%	64.89%	78.49%	77.40%	74.73%	77.61%

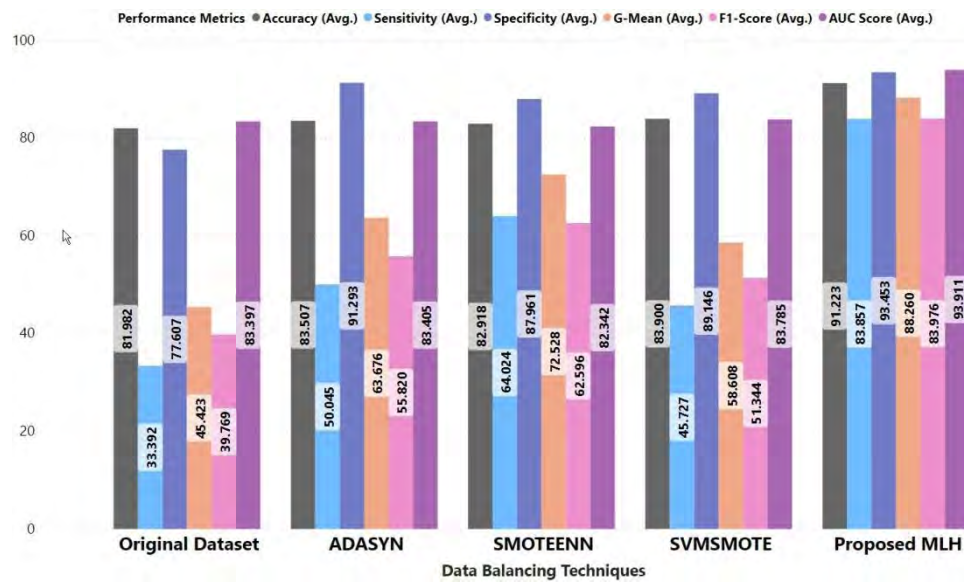


Figure 6.5: Overall Performance Evaluation for Random Forest

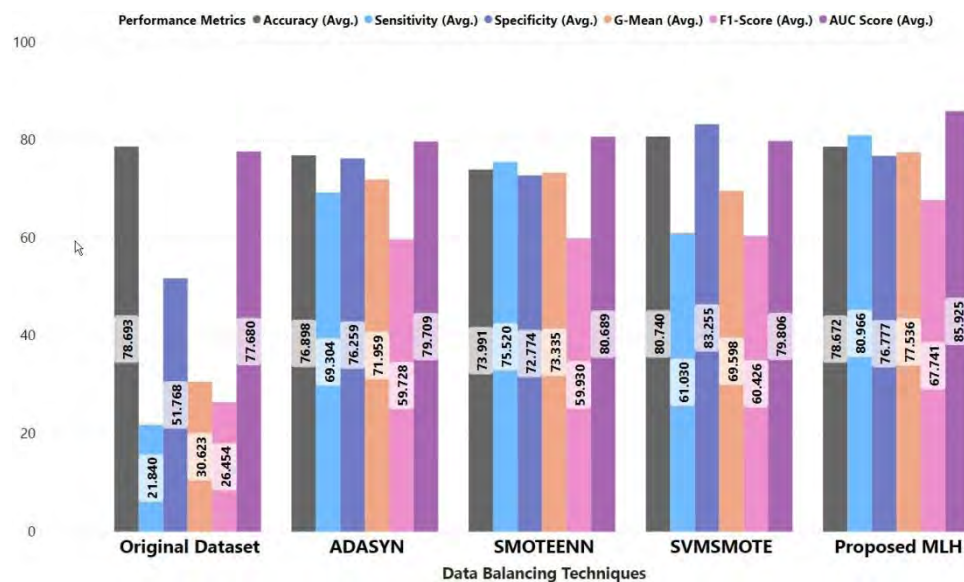


Figure 6.6: Overall Performance Evaluation for Neural Network

Performance Comparison

Sakar et al. [16] used oversampling in online shopping purchasing dataset and got 87.24% accuracy with 86.0% F-Measure, 84.0% recall (TPR) and 92.0% specificity (TNR) for multilayer perceptron classifier. But, we got 95.43% accuracy, 96.34% recall (TPR), 95.04% specificity (TNR), 90.64% F-Measure and 95.69% G-Mean on test dataset for Random forest using our proposed oversampling method.

Johnson et al. [20] used SMOTE techniques in their imbalance dataset and mapped the training dataset with model and got of 84.0% Precision and 70.1% Recall. In our proposed approach, we get 89.87% precision and 73.262% recall using the Random Forest algorithm for the testing dataset. Although this dataset is highly imbalanced, our proposed scheme handles this problem effectively. As our scheme works well for real type attributes, our scheme gives a better result than their approach.

Alizadehsani et al. [56] got 86.14%, 83.17%, and 83.50% accuracy rates for the diagnosis of the stenosis of the left anterior descending (LAD) artery, left circumflex (LCX) artery and right coronary artery (RCA), respectively. Alizadehsani et al. [45] also achieved 94.08% accuracy using data mining methods and the feature creation algorithm, which is higher than the known approaches in the literature. However, using our scheme, we removed the outliers and get 96.92% accuracy, and 100% recall (minor class accuracy) for the dataset using Random Forest.

Lucas et al. [47] achieved 0.96 of AUC score. In our scheme, we get AUC of 98.57% using Random Forest algorithm. Mansouri et al. [50] got 0.91 sensitivity, 0.95 specificity and 0.07 error rate for 5-fold cross-validation method using Consensus analysis. But, using our scheme, we get 95.51% sensitivity, 94.29% specificity and 94.85% accuracy for RF algorithm.

Santos et al. [14] used representative set approach and augmented sets approach (cluster-based oversampling) for HCC dataset. They got 0.737 and 0.752 mean accuracy, 0.689 and 0.700 mean AUC score, 0.640 and 0.665 F-Measure for representative set approach and augmented sets approach, respectively using Neural Network. They also achieved 0.725 and 0.730 mean accuracy, 0.668 and 0.673 mean AUC score, 0.648 and 0.652 F-Measure for representative set approach and augmented sets approach, respectively using Logistic Regression. With our proposed scheme, we got 83.33% accuracy, 87.5% sensitivity, 80.0% F-Measure and 90.63% AUC score for random forest algorithm. Our scheme removes the outliers and avoid overfitting problems effectively. So, for this dataset, our scheme gets a better balancing result for small characteristics dataset.

Demiröz et al. [37] used Voting Feature Intervals (VFI), Classification by Feature Partitions (CFP), Naive Bayesian Classifier (NBC) and K Nearest Neighbor Classification on Feature Projections (k-NNFP) and achieved 87.39%, 89.77%, 89.8% and 90.52% accuracy, respectively for Page Block dataset. But using our oversampling scheme, we got 97.69% accuracy with 91.67% sensitivity using random forest algorithm although it is a highly imbalance dataset.

Kahramanli et al. [57] proposed a Hybrid neural network and got 84.24% accuracy for their approach. Sain et al [10] used combine sampling support vector machine for Pima

diabetes dataset and got 84.96% accuracy, 86.8% AUC score using their process. But using our scheme, we got 86.54% accuracy and 92.72% AUC score for the diabetes dataset using random forest algorithm. They also used Ecoli 1 and Ecoli 2 dataset and using their approach, they got 96.90%, 93.66% accuracy; 83.8%, 89.1% AUC score; 82.2%, 88.4% G-Mean value, respectively. But using our proposed scheme, we achieve 97.83%, 100% accuracy; 100%, 100% AUC score; 98.6%, 100% G-Mean value for Ecoli 1 and Ecoli 2 dataset, respectively.

Rivera et al. [12] used Bank Tele-Marketing dataset for oversampling using propensity scores (OUPS) approach and got 83.77% accuracy, 40.44% sensitivity and 96.71% specificity. But using our proposed scheme, we achieved 94.31% accuracy, 94.83% sensitivity and 93.98% specificity for random forest algorithm. Moro et al. [58] also used this dataset and Neural Network algorithm provide 80% AUC score for this dataset where our scheme gives 98% AUC score using Random Forest model.

Tang et al. used SVM-WEIGHT (Cost sensitive learning for SVM modeling) and effectiveness, and efficiency model Granular Support Vector Machines - Repetitive Under-sampling algorithm (GSVMRU) and got 84% and 81.9% G-Mean using SVM-WEIGHT and GSVMRU, respectively for Abalone (Imbalanced: 19) dataset. They also achieved 86.6% AUC score using SVM-WEIGHT approach [22]. With our scheme, we got 93.3% accuracy, 89.09% G-Mean and 94.2% AUC score for test dataset using random forest algorithm.

Zięba et al. [21] used Ensemble Boosted Trees using the synthetic features (EXGB) in Polish companies bankruptcy dataset and separately get 0.959, 0.944, 0.940, 0.941, 0.955 AUC score for year 1, year 2, year 3, year 4, year 5, respectively. But using our proposed scheme, we get 97.53%, 94.88%, 97.32%, 95.43%, 97.34% AUC score and 88.0%, 85.92%, 87.78%, 86.59%, 90.45% F-Measure for year 1, year 2, year 3, year 4, year 5, respectively. We also combined the whole 5 years dataset and get 96.764% AUC score and 81.734% F-Measure for our proposed scheme.

Ballabio et al. [46] used Consensus Analysis in QSAR oral toxicity dataset and got 80% sensitivity, 96% specificity, 88% Non-Error Rate (NER) for test dataset whereas we got 89.5% accuracy, 84.6% sensitivity and 90.5% specificity for Random Forest model. In our proposed scheme, we can improve positive class (Minor Class) from 80% to 84.6%, but less improvement of negative class (Major Class) as compared to Ballabio et al. [46] work.

For Red Wine dataset, Al Majzoub et al. [18] proposed Hybrid Clustered Affinitive Borderline SMOTE approach for binary classification and got 0.529 recall and 0.643 F-Measure. But using scheme, we got 88% F-Measure, 84.62% sensitivity (recall) and

96% accuracy using our scheme. For Bank Tele-Marketing dataset, Al Majzoub et al. [18] got 0.796 recall and 0.821 F-Measure using their approach. But using the proposed scheme, we achieved 94.31% accuracy, 94.83% Sensitivity and 93.98% specificity for random forest algorithm.

For HR Employee Attrition, Customer Churn Wireless Telecom and Sales: win - loss Al Majzoub et al. [18] used proposed Hybrid Clustered Affinitive Borderline SMOTE (HCAB-SMOTE) approach and got 60.5%, 82.2% and 84.4% Recall; 69.6%, 87.9% and 86.4% F-Measure, respectively. But using our approach, we got 83%, 96.7% and 91% accuracy; 68%, 92% and 90% Recall; 65% 88.8% and 88.6% F-Measure for HR Employee Attrition, Customer Churn Wireless Telecom and Sales: win - loss, respectively. Figure 6.7 and 6.8 shows the comparative analysis between Existing Research and the proposed Multi-Layer Hybrid (MLH) Balancing Technique where X-axis represents dataset name with result name and Y-axis represents percentage of result.

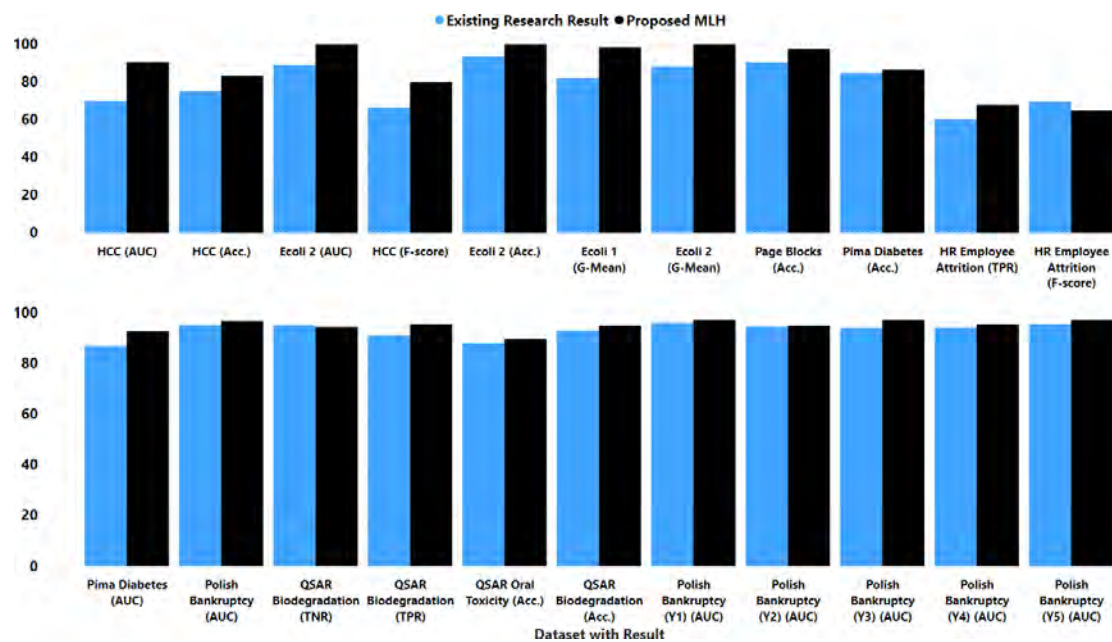


Figure 6.7: Result Compare between Existing Research and Multi-Layer Hybrid (MLH) Balancing Technique (Part 1)

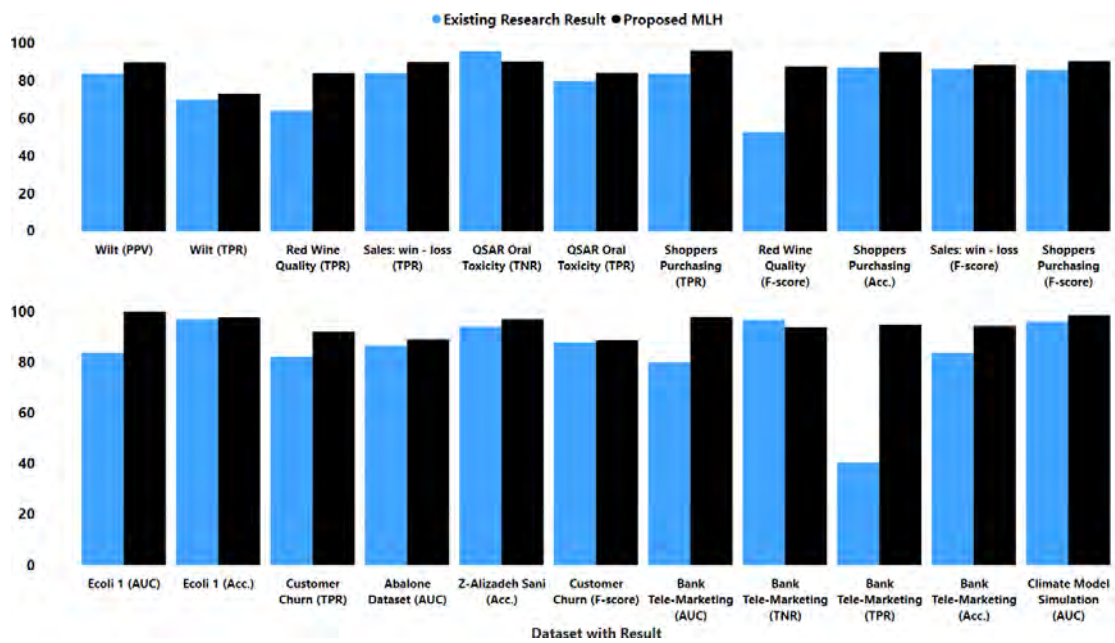


Figure 6.8: Result Compare between Existing Research and Multi-Layer Hybrid (MLH) Balancing Technique (Part 2)

Performance of Group Data

As we categorized our dataset into four main part (Bio-Life, Business, Medical Research and Social criteria). From Figure6.9, we can see Bio-Life data gives the better performance than all groups. Using proposed architecture, we can also observe the average Accuracy, Recall [10] and G-Mean [10] is much more better than existing approach. Proposed approach gives the 89% Accuracy, Recall and G-Mean with 80.5% F-Measure. So from figure6.9we can clearly understand the effectiveness of our proposed model for any kind of imbalanced dataset.

Statistical Analysis

In our experiment, we also used statistical analysis to verify the potentiality of our dataset. Accuracy rate is not useful every time for imbalance dataset in binary classification. ROC AUC score is also not enough for the machine learning model and gives a very abstract idea.Cumulative Gain Chart [59] is very effective solution to verify the model using the probability score. It shows the gain of a model by picking random information from the dataset and gives the potentiality for a particular amount of dataset. To discuss this, let's take an example with the financial aspect. In a bank telemarketing,



Figure 6.9: Group Performance Analysis

the owner wants to advertise a loan scheme to customers using SMS with cost of 0.05-dollar per SMS. They have 10 million customers and 20% of them are frequently using these schemes. So, instead of sending SMS all customer, it is better to use only 20% of total customers which will save money and will not disturb those customers who have no interest in the new product. In this situation, we need to select a better model that can achieve most of those interested customers from 20% dataset. For this scenario, the cumulative gain chart is necessary to find the best model.

In our cumulative gain charts, X-axis represents the percentage of the dataset and the Y-axis represents the response (gain) for a particular class. The relationship of the effectiveness of a model is proportional to the area under a cumulative gain chart. As the area of a chart is bigger, the model is better. Equation 6.1 refers the area of a cumulative gain.

$$CG_{Area} = \sum_{i=0}^N 0.5 * [(CR_i - BL_i) + (CR_{i+1} - CR_{i+1})] * (D_{i+1} - D_i) \quad (6.1)$$

where CR = cumulative gain percent i, BL = baseline at i and D = Decile at i.

In our experiment, we also used some imbalanced financial datasets where minor class effects by major class in the classification phase. For this reason, we compare our proposed approach with the existing approach by using positive class (minor class); in most cases, our approach gives better results (greater area of the gain chart) than the others. Figure 6.10, 6.11, 6.12, 6.13, 6.14, 6.15, 6.16, 6.17 and 6.18 shows the cumulative gain achieved by our proposed approach and existing approach for minor class where Bank Tele-Marketing [35], Online Shoppers Purchasing [16], Default of Credit Card [36], Seismic Bumps [38], QSAR Oral Toxicity [46], QSAR Biodegradation [50], Polish Companies Bankruptcy [21], HR Employee Attrition [18], Wireless-Telecom dataset [18], Blood Transfusion [41] and Sales: win - loss [18] gives better performance than existing over-sampling techniques. But in some cases, our approach does not give good performance for Ecoli 1 [10], Ecoli 2 [10], Immunotherapy [51] and Cervical cancer [40] due to lack of data dimensionality in original dataset.

Summary

In this section, the result of our proposed approach is compared with the existing balancing techniques. Results are shown for the individual dataset and also for grouping

the same category datasets. In most cases, our approach gives better output than the existing approach. Our approach also can handle the high imbalance dataset by analyzing the original dataset. We also apply the cumulative gain (Statistical Analysis) in our proposed approach and it gives better output than the existing approach by getting a different amount of dataset. So from our simulation of the different datasets, we can say that our approach is better than other existing comparing approaches. But in some cases, our approach cannot get better performance than the existing approaches.

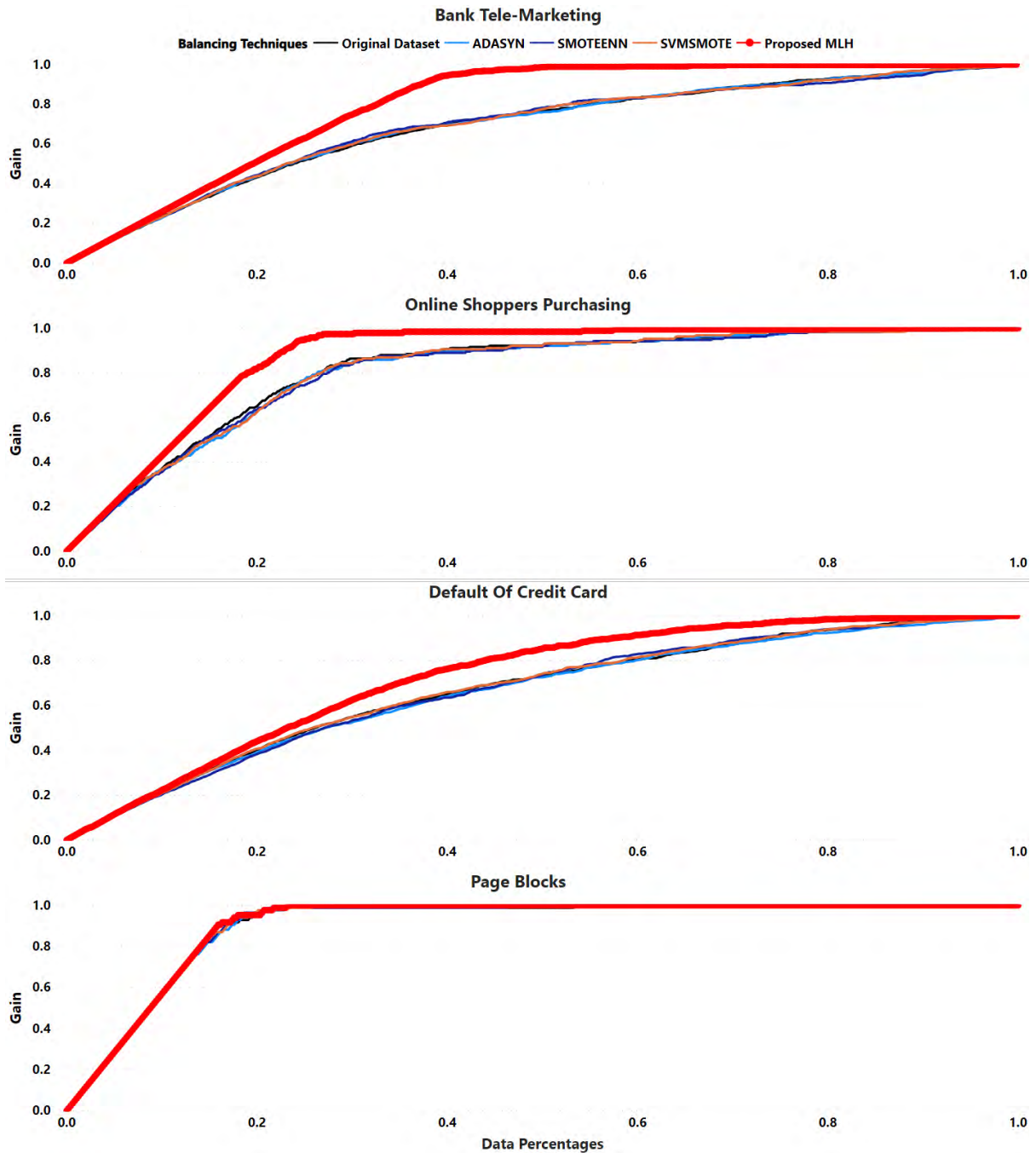


Figure 6.10: Cumulative Gain for Bank Tele-Marketing, Online Shoppers Purchasing, Default of Credit Card and Page Blocks

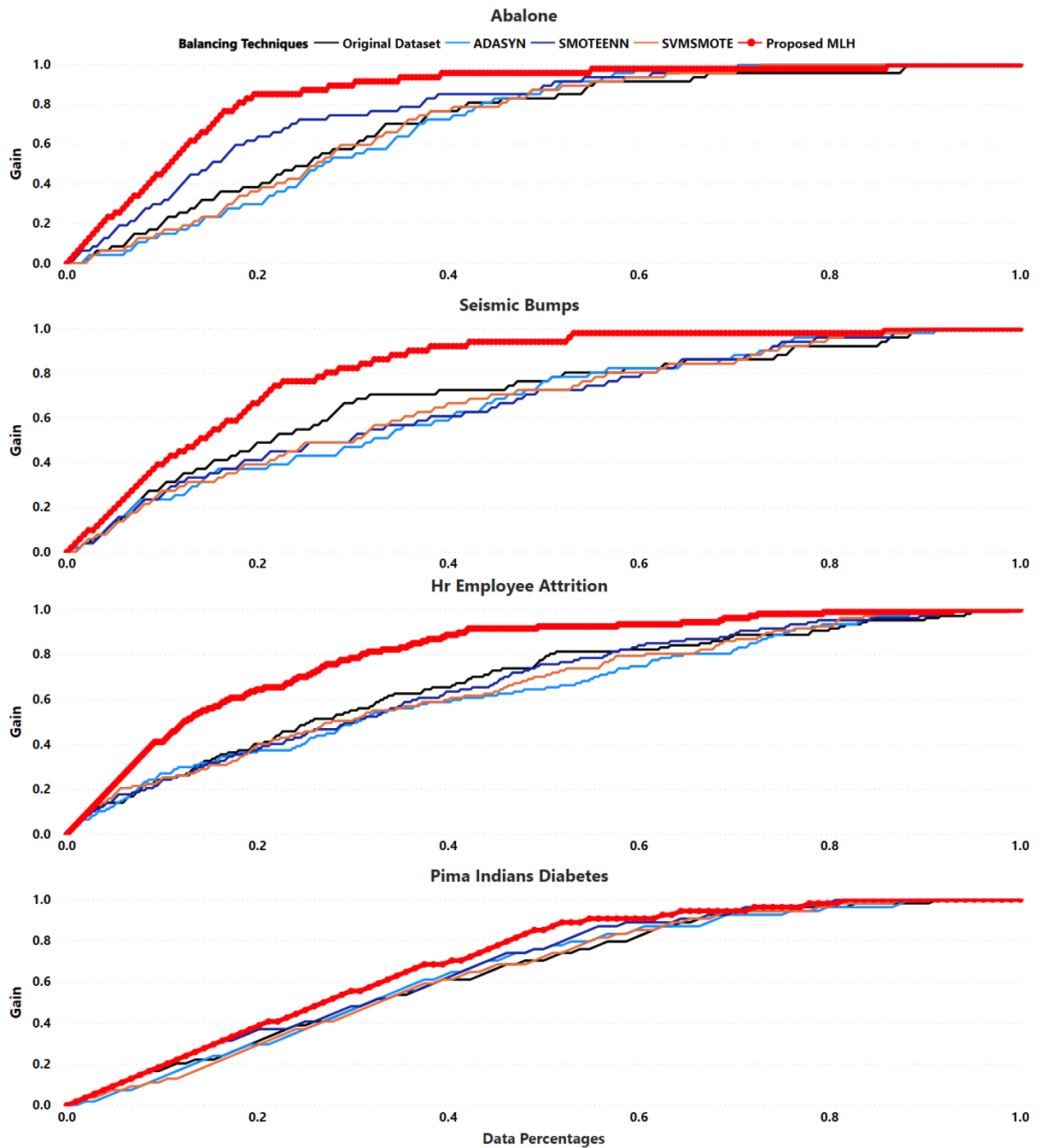


Figure 6.11: Cumulative Gain for Abalone, Seismic Bumps, HR Employee Attrition and Pima Indians Diabetes

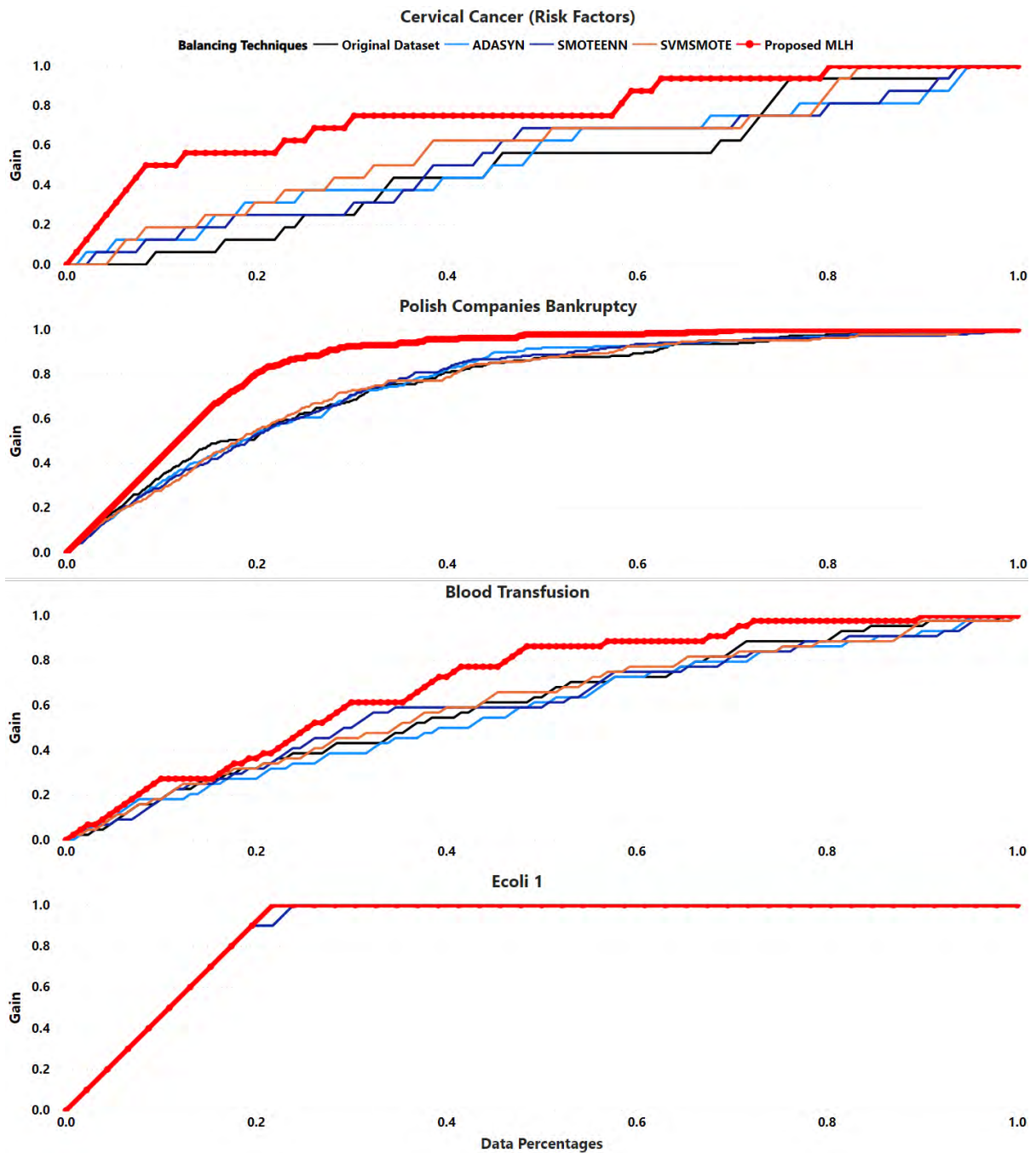


Figure 6.12: Cumulative Gain for Cervical cancer, Polish companies bankruptcy, Blood Transfusion and Ecoli 1

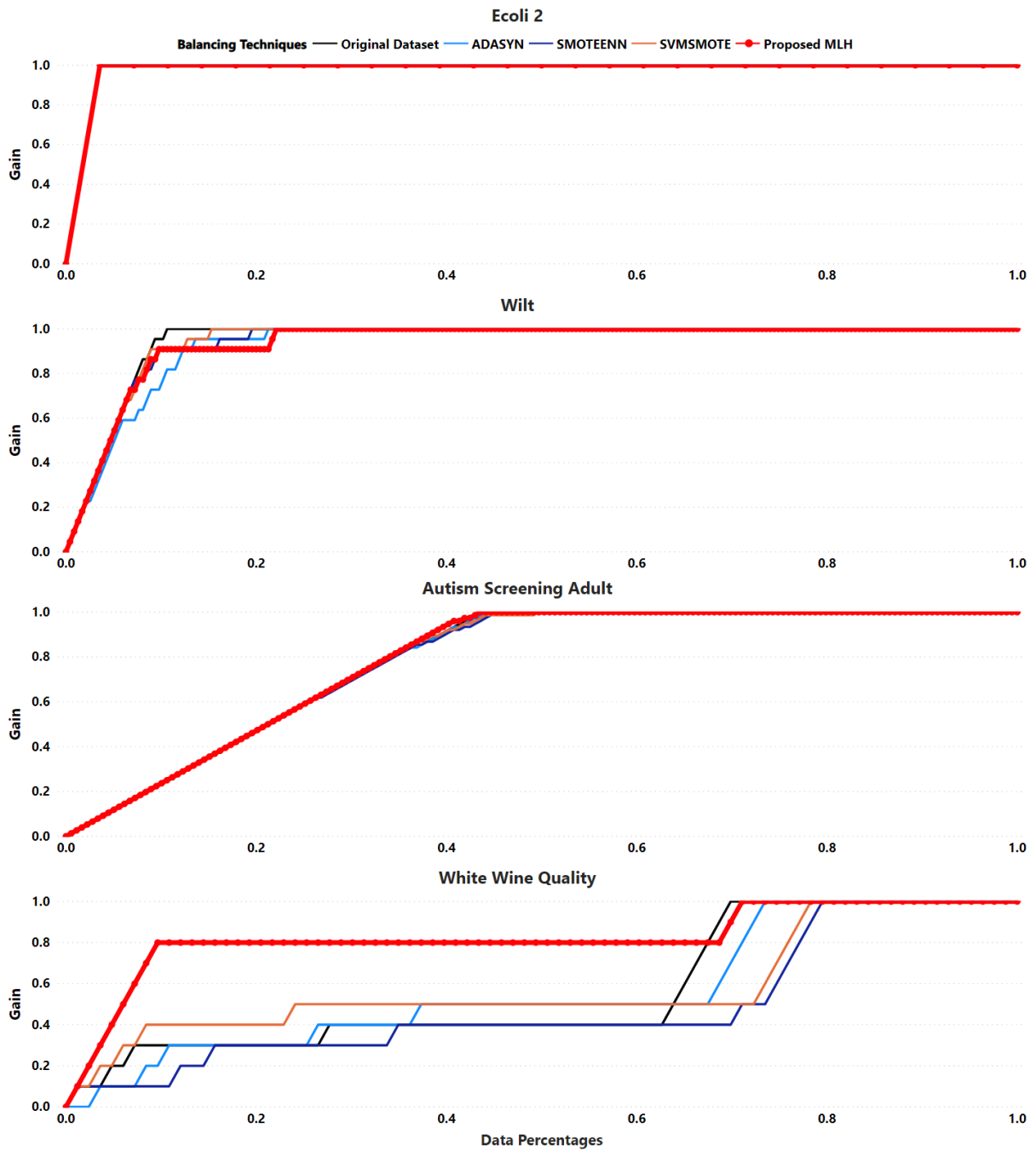


Figure 6.13: Cumulative Gain for Ecoli 2, Wilt, Autism Screening Adult and White Wine Quality

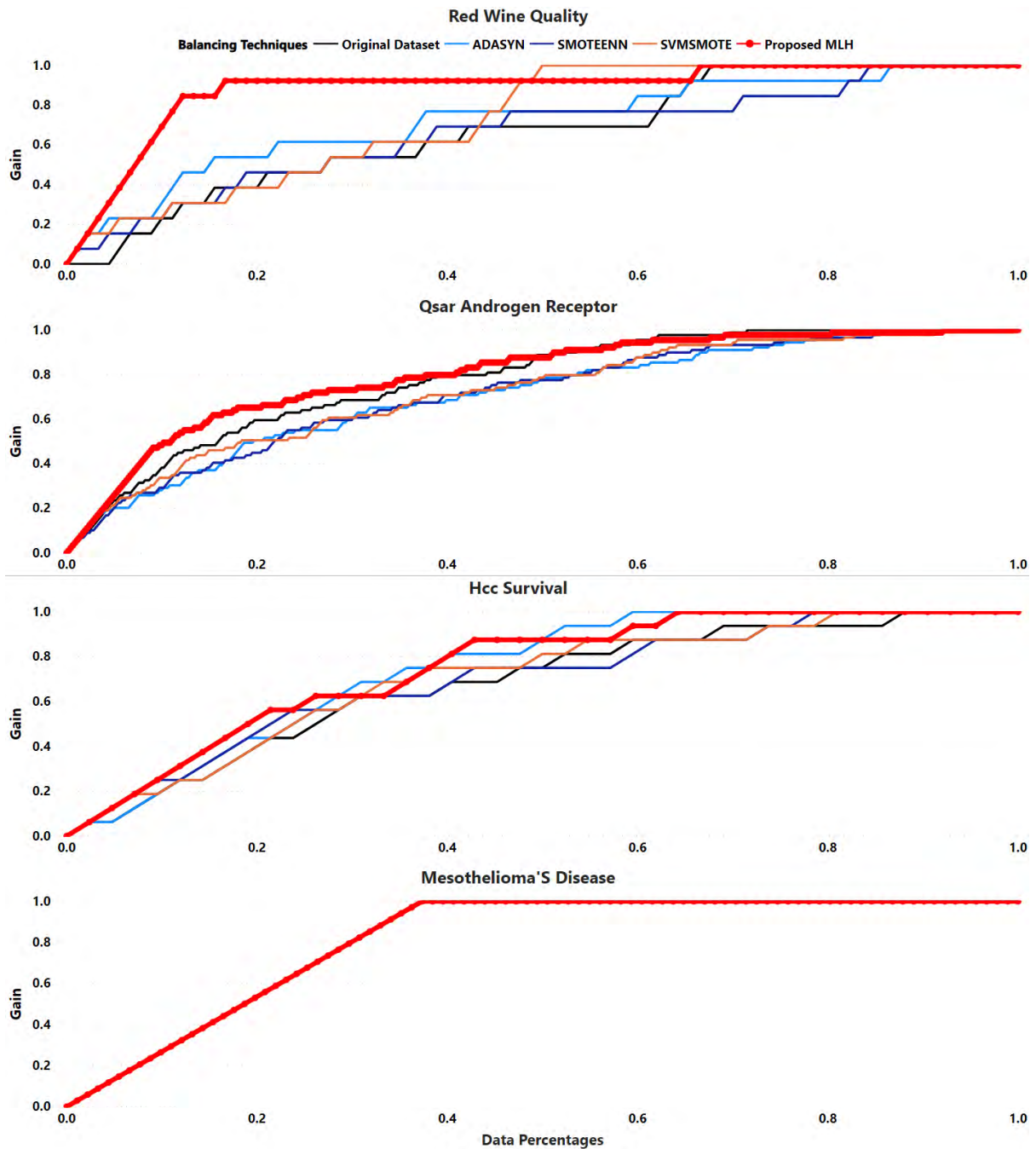


Figure 6.14: Cumulative Gain for Red Wine Quality, QSAR Androgen Receptor, HCC Survival and Mesothelioma's disease

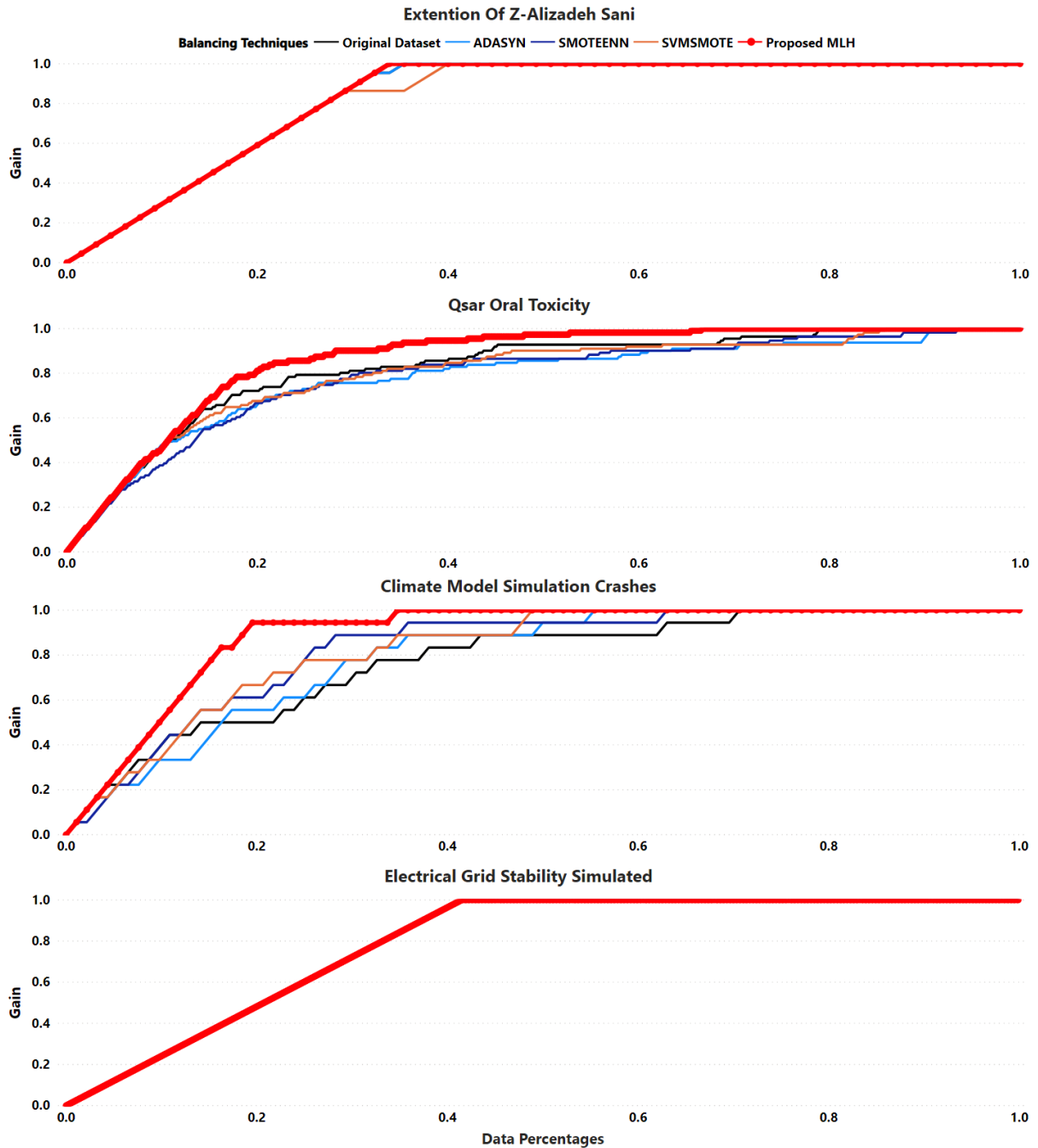


Figure 6.15: Cumulative Gain for Z-Alizadeh Sani, QSAR Oral Toxicity, Climate Model Simulation and Electrical Grid Stability Simulated

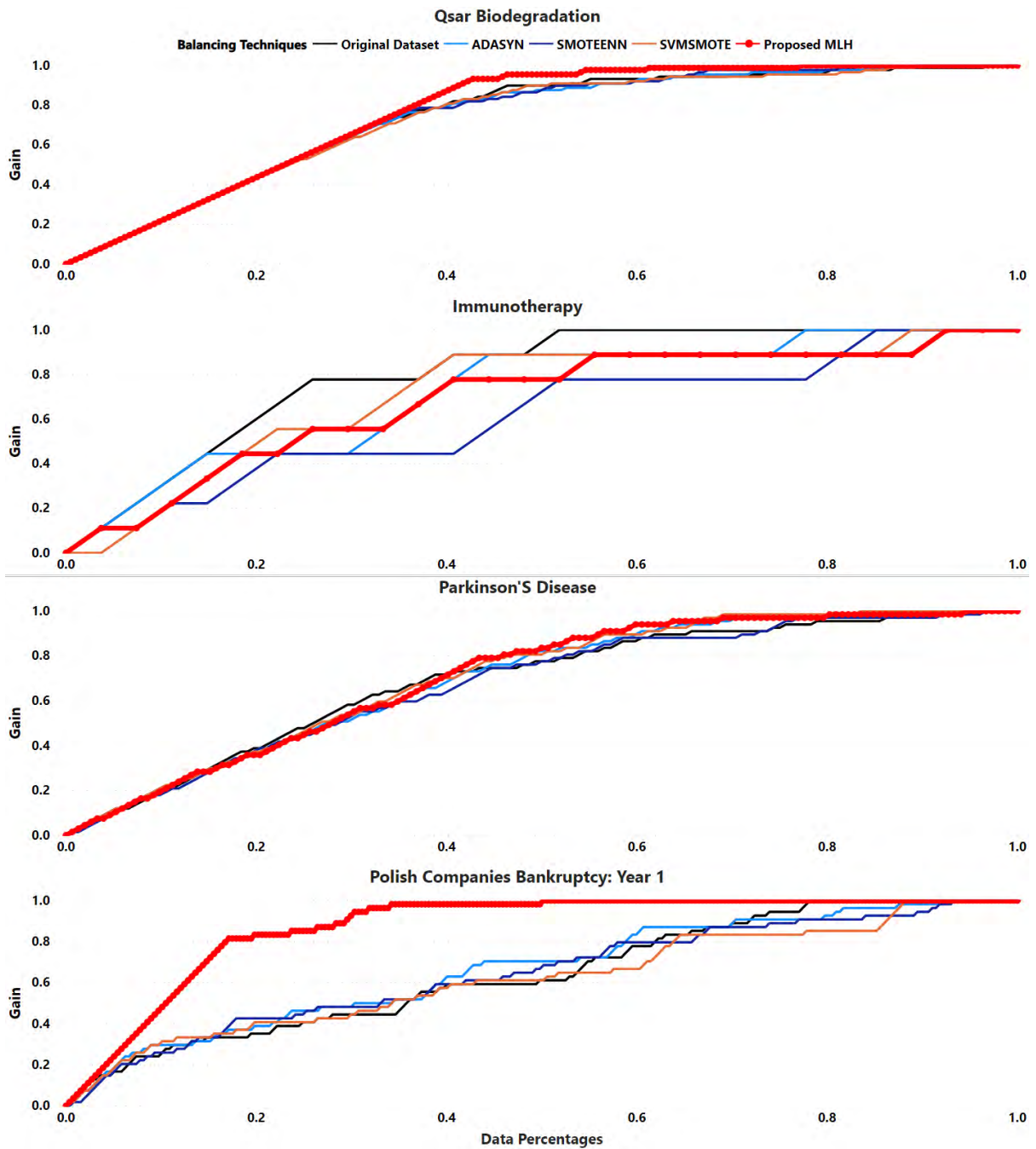


Figure 6.16: Cumulative Gain for QSAR Biodegradation, Immunotherapy, Parkinson's Disease and Polish Companies Bankruptcy (Y1)

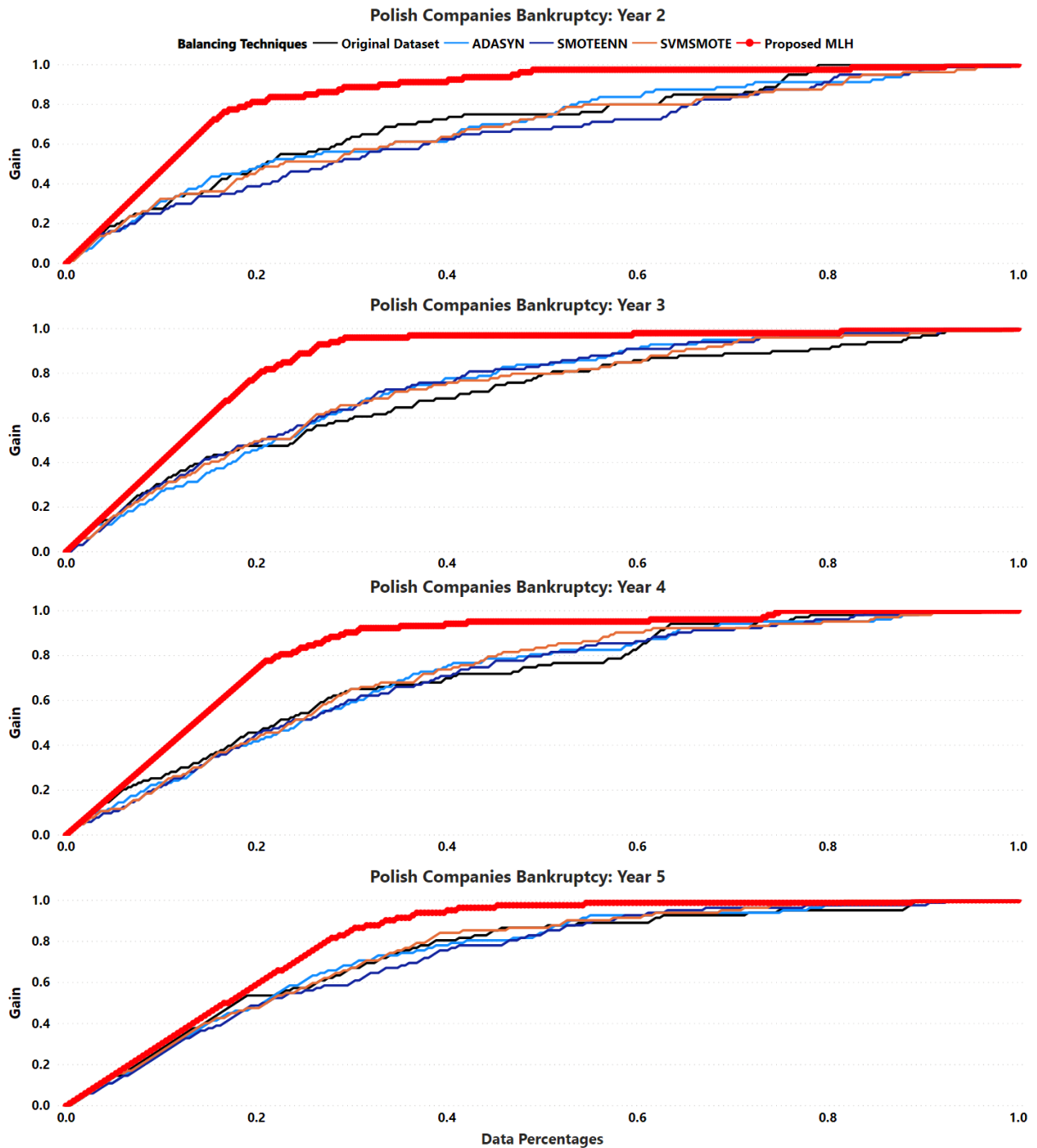


Figure 6.17: Cumulative Gain for Polish Companies Bankruptcy (year 2, year 3, year 4, year 5)

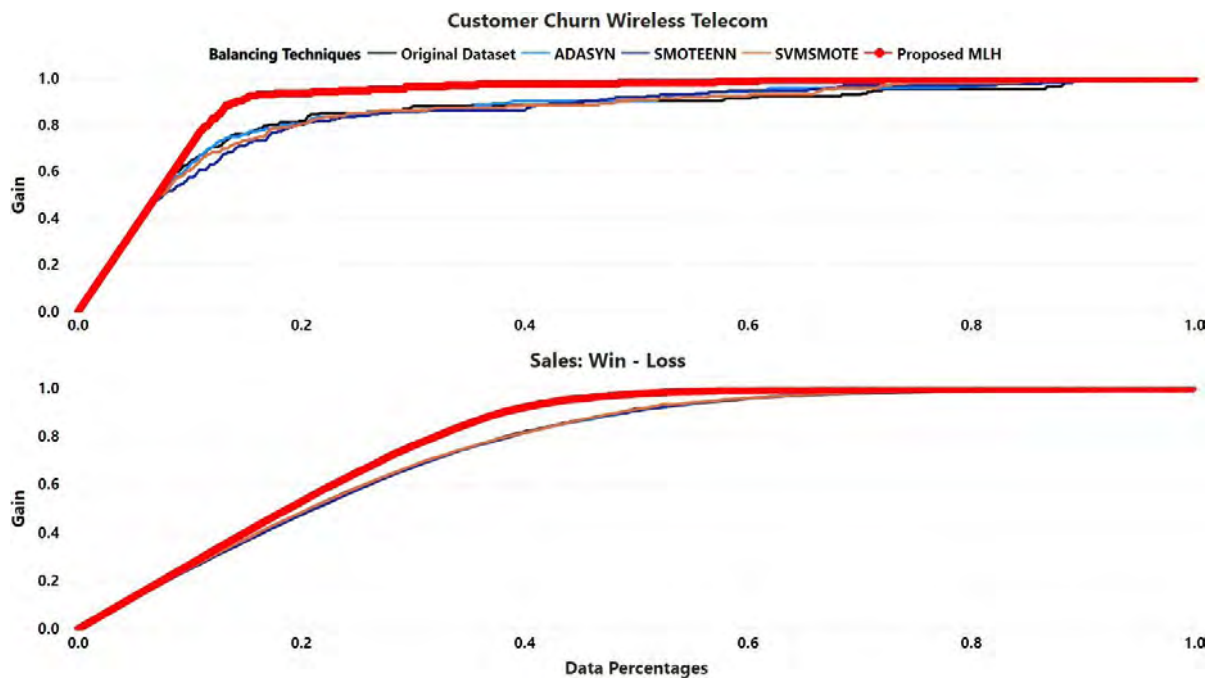


Figure 6.18: Cumulative Gain for Customer Churn Wireless Telecom and Sales: win - loss

Chapter 7

Conclusions

Conclusions

Dataset is very important in our daily life for prediction and analysis in business and scientific research. The Machine-Learning model is the important thing in prediction. But without a proper and balanced dataset, Machine-Learning model is meaningless and cannot work well for analysis. In real life example, most of the useful data is imbalance and have very little information for the minor class attribute. The degree of imbalance dataset can have in a different label and it is difficult to balance the dataset. Some existing data balancing techniques can solve this problem but also creates noisy data points. The existing techniques cannot work well for different dimensions of the dataset.

In this Thesis, we propose a Multi-Layer Hybrid (MLH) balancing scheme to balance different kinds of low, medium, and highly imbalanced datasets. Our proposed scheme can effectively balance datasets as we can observe from our simulation results. We used 34 different dimensions of separate datasets in our proposed approach and by comparing the existing research with our proposed approach, we showed that our proposed MLH balancing scheme works well for different dimensions of datasets and generates a noise-free output. We also observed our dataset group-wise and overall manner and from most of the cases, our proposed approach gives a balanced output from the imbalanced dataset. For the overall result using random forest, we get 91% Accuracy, 83% Precision, 94% Recall, 88% G-Mean and 94% AUC score from the Proposed MLH Balancing technique. Those results are better than the original dataset and existing approach. Also when observing the group data, we see Bio-Life group data average result is 89% Accuracy, 89% Recall, 81% F-Measure and 89% G-Mean; Business group data

average result is 83% Accuracy, 73% Recall, 69% F-Measure and 78% G-Mean; Medical Research group data average result is 77% Accuracy, 79% Recall, 70% F-Measure and 72% G-Mean; Social group data average result is 77% Accuracy, 85% Recall, 66% F-Measure and 78% G-Mean. For each group our proposed approach gives the better result than any other techniques. However, our proposed scheme reduces the most integer-type value, which is the out-range of the original dataset. In that case, some characteristics are also removed from the output resulting reduction in the dimensionality of the dataset. We also observe that our scheme cannot work well on the datasets which contain all integer-type attribute and have less robust dimension. Despite these limitations, we have shown that our proposed scheme can handle the highly imbalanced dataset and outperforms most of the existing techniques.

Future Prospects of Our Work

The limitations of the proposed solution have indicated the following areas as recommendations for future work:

- As our approach cannot work well for integer type dataset and reduce most of the dimension from the output, so our aim to solve these dimension reduction problems in the future.
- For some dataset it takes lots of time to complete process for huge number of data boosting. So, in future, we solve this problem to make our process faster.
- In our current experiment, we only used binary classification dataset for simulation. So, in future we will use multi classification dataset as a new feature for our study.
- In future, we also use the deep-learning approach in our proposed approach to simulate the result.

Reference

- [1]R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine learning with over-sampling and undersampling techniques: overview study and experimental results,” in *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020, pp. 243–248.
- [2]P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–50, 2016.
- [3]F. Provost, “Machine learning from imbalanced data sets 101,” in *Proceedings of the AAAI’2000 workshop on imbalanced data sets*, vol. 68. AAAI Press, 2000, pp. 1–3.
- [4]V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.
- [5]N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [6]A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, “Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study,” *IEEE Access*, vol. 4, pp. 7940–7957, 2016.
- [7]G. E. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.
- [8]H. He, Y. Bai, E. A. Garcia, and S. Li, “Adasyn: Adaptive synthetic sampling approach for imbalanced learning,” in *2008 IEEE international joint conference*

- on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.
- [9]H. M. Nguyen, E. W. Cooper, and K. Kamei, –Borderline over-sampling for imbalanced data classification,” *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011.
- [10]H. Sain and S. W. Purnami, –Combine sampling support vector machine for imbalanced data classification,” *Procedia Computer Science*, vol. 72, pp. 59–66, 2015.
- [11]M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, –Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier],” *ieeE ComputatioNal iTelligeNCe magaziNe*, vol. 13, no. 4, pp. 59–76, 2018.
- [12]W. A. Rivera, A. Goel, and J. P. Kincaid, —Ops: a combined approach using smote and propensity score matching,” in *2014 13th International Conference on Machine Learning and Applications*. IEEE, 2014, pp. 424–427.
- [13]S. Cateni, V. Colla, and M. Vannucci, –A method for resampling imbalanced datasets in binary classification tasks for real-world problems,” *Neurocomputing*, vol. 135, pp. 32–41, 2014.
- [14]M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, –A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients,” *Journal of biomedical informatics*, vol. 58, pp. 49–59, 2015.
- [15]H.-Y. Wang, –Combination approach of smote and biased-svm for imbalanced datasets,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 228–231.
- [16]C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, –Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and lstm recurrent neural networks,” *Neural Computing and Applications*, vol. 31, no. 10, pp. 6893–6908, 2019.
- [17]M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, —A hybrid under-sampling method (husboost) to classify imbalanced data,” in *2018 21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–7.

- [18]H. Al Majzoub, I. Elgedawy, Ö. Akaydın, and M. K. Ulukök, —Hab-smote: A hybrid clustered affinitive borderline smote approach for imbalanced data binary classification,” *Arabian Journal for Science and Engineering*, pp. 1–18, 2020.
- [19]A. Chemchem, F. Alin, and M. Krajecki, —Combining smote sampling and machine learning for forecasting wheat yields in france,” in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, 2019, pp. 9–14.
- [20]B. A. Johnson, R. Tateishi, and N. T. Hoan, —A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees,” *International journal of remote sensing*, vol. 34, no. 20, pp. 6969–6982, 2013.
- [21]M. Zięba, S. K. Tomczak, and J. M. Tomczak, —Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction,” *Expert Systems with Applications*, vol. 58, pp. 93–101, 2016.
- [22]Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, —Svms modeling for highly imbalanced classification,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2008.
- [23]—Keel: A software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on),”<https://sci2s.ugr.es/keel/dataset.php?cod=1325>, (Accessed on 06/16/2020).
- [24]L. Breiman, —Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [25]T. K. Ho, —Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
- [26]I. Barandiaran, —The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 8, pp. 1–22, 1998.
- [27]J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
- [28]M. Pal, —Random forest classifier for remote sensing classification,” *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.
- [29]L. Breiman, J. Friedman, R. Olshen, and C. Stone, —Classification and regression trees. wadsworth int,” *Group*, vol. 37, no. 15, pp. 237–251, 1984.

- [30]N. Murata, S. Yoshizawa, and S.-i. Amari, “Network information criterion-determining the number of hidden units for an artificial neural network model,” *IEEE transactions on neural networks*, vol. 5, no. 6, pp. 865–872, 1994.
- [31]X. Yao, “Evolutionary artificial neural networks,” *International journal of neural systems*, vol. 4, no. 03, pp. 203–222, 1993.
- [32]M. Kumari and S. Godara, “Comparative study of data mining classification methods in cardiovascular disease prediction 1,” 2011.
- [33]J. M. Keller, M. R. Gray, and J. A. Givens, “A fuzzy k-nearest neighbor algorithm,” *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.
- [34]T. Fushiki, “Estimation of prediction error by using k-fold cross-validation,” *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011.
- [35]S. Moro, R. Laureano, and P. Cortez, “Using data mining for bank direct marketing: An application of the crisp-dm methodology,” in *Proceedings of European Simulation and Modelling Conference-ESM’2011*. EUROSIS-ETI, 2011, pp. 117–121.
- [36]I.-C. Yeh and C.-h. Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
- [37]G. Demiröz and H. A. Güvenir, “Classification by voting feature intervals,” in *European Conference on Machine Learning*. Springer, 1997, pp. 85–92.
- [38]M. Sikora *et al.*, “Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines,” *Archives of Mining Sciences*, vol. 55, no. 1, pp. 91–114, 2010.
- [39]A. Asuncion and D. Newman, “Uci machine learning repository,” 2007.
- [40]K. Fernandes, J. S. Cardoso, and J. Fernandes, “Transfer learning with partial observability applied to cervical cancer screening,” in *Iberian conference on pattern recognition and image analysis*. Springer, 2017, pp. 243–250.
- [41]I.-C. Yeh, K.-J. Yang, and T.-M. Ting, “Knowledge discovery on rfm model using bernoulli sequence,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5866–5871, 2009.

- [42] F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *International journal of medical informatics*, vol. 117, pp. 112–124, 2018.
- [43] F. Grisoni, V. Consonni, and D. Ballabio, "Machine learning consensus to predict the binding to the androgen receptor within the compara project," *Journal of chemical information and modeling*, vol. 59, no. 5, pp. 1839–1848, 2019.
- [44] O. Er, A. C. Tanrikulu, A. Abakay, and F. Temurtas, "An approach based on probabilistic neural network for diagnosis of mesothelioma's disease," *Computers & Electrical Engineering*, vol. 38, no. 1, pp. 75–81, 2012.
- [45] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease," *Computer methods and programs in biomedicine*, vol. 111, no. 1, pp. 52–61, 2013.
- [46] D. Ballabio, F. Grisoni, V. Consonni, and R. Todeschini, "Integrated qsar models to predict acute oral systemic toxicity," *Molecular informatics*, vol. 38, no. 8-9, p. 1800124, 2019.
- [47] D. Lucas, R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic, and Y. Zhang, "Failure analysis of parameter-induced simulation crashes in climate models," *Geoscientific Model Development*, vol. 6, no. 4, pp. 1157–1171, 2013.
- [48] V. Arzamasov, K. Böhm, and P. Jochem, "Towards concise models of grid stability," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2018, pp. 1–6.
- [49] B. Schäfer, C. Grabow, S. Auer, J. Kurths, D. Witthaut, and M. Timme, "Damping instabilities in power grid networks by decentralized control," *The European Physical Journal Special Topics*, vol. 225, no. 3, pp. 569–582, 2016.
- [50] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative structure–activity relationship models for ready biodegradability of chemicals," *Journal of chemical information and modeling*, vol. 53, no. 4, pp. 867–878, 2013.
- [51] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," *Computers in biology and medicine*, vol. 81, pp. 167–175, 2017.

- [52]H. Gunduz, “Deep learning-based parkinson’s disease classification using vocal feature sets,” *IEEE Access*, vol. 7, pp. 115 540–115 551, 2019.
- [53]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [54]G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: <http://jmlr.org/papers/v18/16-365.html>
- [55]T. pandas development team, “pandas-dev/pandas: Pandas,” Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>
- [56]R. Alizadehsani, M. H. Zangoeei, M. J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir, F. Khozeimeh, N. Sarrafzadegan, and S. Nahavandi, “Coronary artery disease detection using computational intelligence methods,” *Knowledge-Based Systems*, vol. 109, pp. 187–197, 2016.
- [57]H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” *Expert systems with applications*, vol. 35, no. 1-2, pp. 82–89, 2008.
- [58]S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decision Support Systems*, vol. 62, pp. 22–31, 2014.
- [59]T. Jaffery and S. X. Liu, “Measuring campaign performance by using cumulative gain and lift chart,” in *SAS Global Forum*, 2009, p. 196.