Imbalance Learning and Its Application on Medical Datasets

Dissertation
for the award of the degree

Doctor of Philosophy (Ph.D.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral Program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Yachao Shao

from Henan, China
Göttingen, 2021

Thesis Committee:

Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Jar-der Luo
Sociology Department Social Science School, and Public Administration School, Tsinghua University

Members of the Examination Board:

Reviewer:
Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Second Reviewer:
Prof. Dr. Ulrich Sax
Institut für Medizinische Informatik, Universitätsmedizin Göttingen

Further members of the Examination Board:
Prof. Dr. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Dieter Hogrefe
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Tim Friede
Institut für Medizinische Statistik, Universitätsmedizin Göttingen

Prof. Dr. Dagmar Krefting
Institut für Medizinische Informatik, Universitätsmedizin Göttingen

Date of the oral examination: 24. March 2021

Acknowledgement

To express my opinion throughout my PhD study, I would like to quote one famous sentence of David Hilbert: Wir müssen wissen, wir werden wissen (We must know, we will know)!

It has been a tough but wonderful journey since I started my PhD study. There are numerous pain and joy while reading papers, searching scientific questions, struggling to propose novel solutions, and designing experiments to evaluate the solutions. Thanks to the great spirits I learned from all brilliant scientists, I can finish this dissertation today.

I would like to give my deepest gratitude to my supervisor Prof. Dr. Xiaoming Fu first and foremost. Prof. Dr. Fu not only teaches me the research skills but also guides me thinking critically during my Phd study, which would deeply influence my study and life in the future. Under his patient and strict supervision, I get countless valuable suggestions to build my thesis step by step.

I also wish to express my sincere gratitude to co-supervisor Prof. Dr. Marcus Baum and Prof. Jar-der Luo for their supervision and valuable advises, which are of great importance to finish this dissertation.

I would like to thank Prof. Dr. Ulrich Sax for reviewing my thesis. I also wish to thank Prof. Dr. Marcus Baum, Prof. Dr. Dieter Hogrefe, Prof. Dr. Tim Friede, and Prof. Dr. Dagmar Krefting for serving as the examination board of my dissertation.

I wish to thank all my friends and colleagues who supported and helped me in the past three years. We have shared our knowledge and research experience through discussion. Specifically, I want to thank Tina Bockler, Annette Kadziora, Heike Jachinke and all other staff members who helped me. I would thank Dr. Osamah Barakat for his assistance in my first teaching assistant of Computer Networks. I want to thank Dr. Yali Yuan for her help and suggestions. I would like to sincerely thank Dr. Tao Zhao for his kindness help and support.

Thanks to the international Computer Network Group in the University of Göttingen, I get the chance to know people and culture from different counties. I also would like to thank Prof. Dr. Xiaofeng Zou, Dr. Xiaoning Wang and other collaborators from First Affiliated Hospital of Gannan Medical University for working together to develop applications on the kidney stone datasets.

I am also very gratitude to China Scholarship Council (CSC), who supported my PhD study financially. I would like to thank all my Chinese and German friends who helped and supported me.

Finally, I would like to thank my father Mr. Tiansheng Shao and my mother Mrs. Xiaojuan Ruan for giving birth to me and supporting me unconditionally. I would like to thank my sister Mrs. Huizhen Shao and my brother Mr. Huixin Shao for their support and love to my family while I study abroad.

Abstract

To gain more valuable information from the increasing large amount of data, data mining has been a hot topic that attracts growing attention in this two decades. One of the challenges in data mining is imbalance learning, which refers to leaning from imbalanced datasets. The imbalanced datasets is dominated by some classes (majority) and other under-represented classes (minority). The imbalanced datasets degrade the learning ability of traditional methods, which are designed on the assumption that all classes are balanced and have equal misclassification costs, leading to the poor performance on the minority classes. This phenomenon is usually called the class imbalance problem. However, it is usually the minority classes of more interest and importance, such as sick cases in the medical dataset. Additionally, traditional methods are optimized to achieve maximum accuracy, which is not suitable for evaluating the performance on imbalanced datasets. From the view of data space, class imbalance could be classified as extrinsic imbalance and intrinsic imbalance. Extrinsic imbalance is caused by external factors, such as data transmission or data storage, while intrinsic imbalance means the dataset is inherently imbalanced due to its nature. As extrinsic imbalance could be fixed by collecting more samples, this thesis mainly focus on on two scenarios of the intrinsic imbalance, machine learning for imbalanced structured datasets and deep learning for imbalanced image datasets.

Normally, the solutions for the class imbalance problem are named as imbalance learning methods, which could be grouped into data-level methods (re-sampling), algorithm-level (re-weighting) methods and hybrid methods. Data-level methods modify the class distribution of the training dataset to create balanced training sets, and typical examples are over-sampling and under-sampling. Instead of modifying the data distribution, algorithm-level methods adjust the misclassification cost to alleviate the class imbalance problem, and one typical example is cost sensitive methods. Hybrid methods usually combine data-level methods and algorithm-level methods. However, existing imbalance learning methods encounter different kinds of problems. Over-sampling methods increase the minority samples to create balanced training sets, which might lead the trained model overfit to the minority class. Under-sampling methods create balanced training sets by discarding majority samples, which lead to the information loss and poor performance of the trained model. Cost-sensitive methods usually need assistance from domain expert to define the misclassification costs which are task specified. Thus, the generalization ability of cost-sensitive methods is poor. Especially, when it comes to the deep learning methods under class imbalance, re-sampling methods may introduce large computation cost and existing re-weighting methods could lead to poor performance. The object of this dissertation is to understand features difference under class imbalance, to improve

the classification performance on structured datasets or image datasets. This thesis proposes two machine learning methods for imbalanced structured datasets and one deep learning method for imbalance image datasets. The proposed methods are evaluated on several medical datasets, which are intrinsically imbalanced.

Firstly, we study the feature difference between the majority class and the minority class of an imbalanced medical dataset, which is collected from a Chinese hospital. After data cleaning and structuring, we get 3292 kidney stone cases treated by Percutaneous Nephrolithonomy from 2012 to 2019. There are 651 (19.78%) cases who have postoperative complications, which makes the complication prediction an imbalanced classification task. We propose a sampling-based method SMOTE-XGBoost and implement it to build a postoperative complication prediction model. Experimental results show that the proposed method outperforms classic machine learning methods. Furthermore, traditional prediction models of Percutaneous Nephrolithonomy are designed to predict the kidney stone status and overlook complication related features, which could degrade their prediction performance on complication prediction tasks. To this end, we merge more features into the proposed sampling-based method and further improve the classification performance. Overall, SMOTE-XGBoost achieves an AUC of 0.7077 which is 41.54% higher than that of S.T.O.N.E. nephrolithometry, a traditional prediction model of Percutaneous Nephrolithonomy.

After reviewing the existing machine learning methods under class imbalance, we propose a novel ensemble learning approach called Multiple bAlance Subset Stacking (MASS). MASS first cuts the majority class into multiple subsets by the size of the minority set, and combines each majority subset with the minority set as one balanced subsets. In this way, MASS could overcome the problem of information loss because it does not discard any majority sample. Each balanced subset is used to train one base classifier. Then, the original dataset is feed to all the trained base classifiers, whose output are used to generate the stacking dataset. One stack model is trained by the staking dataset to get the optimal weights for the base classifiers. As the stacking dataset keeps the same labels as the original dataset, which could avoid the overfitting problem. Finally, we can get an ensembled strong model based on the trained base classifiers and the staking model. Extensive experimental results on three medical datasets show that MASS outperforms baseline methods. The robustness of MASS is proved over implementing different base classifiers. We design a parallel version MASS to reduce the training time cost. The speedup analysis proves that Parallel MASS could reduce training time cost greatly when applied on large datasets. Specially, Parallel MASS reduces 101.8% training time compared with MASS at most in our experiments.

When it comes to the class imbalance problem of image datasets, existing imbalance learning methods suffer from the problem of large training cost and poor performance. After introducing the problem of implementing resampling methods on image classification tasks, we demonstrate issues of re-weighting strategy using class frequencies through the experimental result on one medical image dataset. We propose a novel re-weighting method Hardness Aware Dynamic loss to solve the class imbalance problem of image datasets. After each training epoch of deep

neural networks, we compute the classification hardness of each class. We will assign higher class weights to the classes have large classification hardness values and vice versa in the next epoch. In this way, HAD could tune the weight of each sample in the loss function dynamically during the training process. The experimental results prove that HAD significantly outperforms the state-of-the-art methods. Moreover, HAD greatly improves the classification accuracies of minority classes while only making a small compromise of majority class accuracies. Especially, HAD loss improves 10.04% average precision compared with the best baseline, Focal loss, on the HAM10000 dataset.

At last, I conclude this dissertation with our contributions to the imbalance learning, and provide an overview of potential directions for future research, which include extensions of the three proposed methods, development of task-specified algorithms, and fixing the challenges of within-class imbalance.

Keywords: medical datasets, class imbalance, imbalance learning, data mining, machine learning, deep learning.

Contents

Chapter 1

Introduction

First in Section 1.1, we introduce the class imbalance problem and the issues of existing imbalance learning methods, which is the background of this dissertation. In section 1.2, we list the motivations of our research. In Section 1.3, we list our contributions of this dissertation. Lastly, a content guide about this dissertation structure is given in Section 1.4.

Contents

1.1 Class Imbalance Problem

The rapidly growing available large datasets and the fast development of artificial intelligence enable us to investigate the datasets and discover valuable information accordingly. Data mining technologies are crucial in a variety of applications from microscale data analysis to macroscale knowledge discovery, from daily personal life to national security [38]. One important challenge in data mining area is called class imbalance, where the dataset is dominated by some classes (majority) and other under-represented classes (minority). Under the situation of class imbalance, the standard learning methods will generate poor performance on the minority classes since the class distribution is an important element in classification tasks [13]. Many of the existing standard learning algorithms assume that the classes are evenly distributed and their classification errors have the same cost during the training process. However, the class distribution of real world datasets is usually imbalanced and the misclassification costs of different classes are not equal. For example in the task of cancer diagnosis, the number of healthy cases (majority) are much larger than that of the cancer patients (minority). It is obvious that the misclassification cost of diagnosing a cancer patient to be healthy, which might lead to the loss of the patient's life, is much higher than the misclassification cost of diagnosing a healthy case to be sick, which brings mental stress and additional cost to the patient. Therefore, it is important to improve the classification performance on the minority classes under the class imbalance situation. However, traditional machine learning algorithms are trained to achieve the maximum overall accuracy, which will lead to the poor prediction performance on the minority classes as they contribute little [115]. Assuming the cancer dataset includes 99% of healthy cases and only 1% sick cases, a naive solution is to classify all cases as health, and the overall accuracy of the classifier would be 99%, which is pretty good at the first glance. However, the classifier fails detect any sick case from all cases. Therefore, overall accuracy is not suitable to evaluate the prediction performance under the class imbalance situation. In Section 2.2, we introduce four metrics for evaluating the performance under class imbalance, such as F-measure, G-mean, Area Under the ROC Curve (AUC) , Matthews Correlation Coefficient (MCC).

The class imbalance problem attracts growing interest from both academia and industry. The solutions of the imbalance problem are named as imbalance learning. When dealing with imbalanced structured datasets, a variety of machine learning approaches have been proposed, and they could be categorized into three groups, i.e., data-level approaches, algorithm-level approaches and hybrid approaches. Data-level approaches alleviate class imbalance by changing the distribution of training data to decrease the imbalance degree. Most of these approaches could be grouped into three kinds, over-sampling, under-sampling and hybrid sampling, i.e., using over-sampling and under-sampling simultaneously. Over-sampling has been proved with over-fitting problem [16], which occurs if a model is poorly generalized to new data because the model is trained to fit the training data too closely. Under-sampling approaches discard samples from the majority class to generate a balanced training set, which will lead to the information loss. Different from data-level approaches, algorithm-level approaches do not change the distribution of training data. Alternatively, they are developed to fix the

imbalance problem by increasing the misclassification cost of the minority samples during the training process. The main problem of cost-sensitive approaches is that the cost matrix definition needs domain experts' assistance before hand, which is often not available in real world cases. Another problem is that cost-sensitive approaches usually algorithm-specific, which is much harder than sampling approaches. In order to take both advantage of data-level approaches and algorithm-level approaches, a number of studies have been conducted to combine them in different ways to alleviate the class imbalance problem [60]. As there are sampling based approaches or cost-sensitive learning approaches in hybrid approaches, they still suffer similar drawbacks of sampling approaches or cost-sensitive learning approaches. As mentioned previously, imbalance learning are beneficial to a wide range of real-world applications, such as medical diagnosis [129, 110], mortality prediction [9], fraud detection in user behaviour [30], and defect prediction in software engineering [81].

When it comes to deal with image datasets, class imbalance problem will decrease the prediction performance of deep learning methods, which achieve great success in computer vision applications. Similar to the machine learning under class imbalance, solutions for imbalanced image classification could be grouped into re-sampling methods (data-level) [14, 73, 89] or re-weighting methods (algorithm-level) [71, 55, 22]. Re-sampling methods include over-sampling for the minority classes (adding duplicated minority samples), under-sampling for the majority classes (discarding majority samples), or hybrid sampling for both majority and minority classes. In the context of computer vision applications, over-sampling methods introduce large training costs and make the model prone to overfit the minority classes. Under-sampling methods discard important samples that are valuable for deep representation learning. Taking these issues of applying re-sampling methods on image classification tasks into consideration, we focuses on designing a better re-weighting method to improve the prediction performance of the deep neural networks. Existing re-weighting methods usually assign the class weight inversely proportional to its size respectively, which might lead to poor performance as proved in Section 6.1. The main reason is that there exist relative imbalance, i.e., some minority class is well present by its samples as described in Section 2.1.

In this dissertation, we take the intrinsically imbalanced medical datasets as study cases. Medical datasets usually include patients' healthcare information such as demographics, laboratory tests, medical history, radiology images, symptoms, and diagnosis. Medical datasets provide useful information to build risk prediction models, which could help estimate the risk of developing a condition of interest. For instance, as half of the complications are preventable [53], accurate prediction of complications is highly important for clinical decision making, early treatment and counseling patient [109]. More details of data mining applications on medical datasets is described in Section 2.5.

In this thesis, we propose two machine learning methods under class imbalance and one deep learning method under class imbalance. Our work mainly contains three part as following:

- **A Sampling-based Method SMOTE-XGBoost** This work mainly focuses on analyzing the features of patient treated by PCNL and compares their differences between the majority class and the minority class according to the postoperative complication status. We propose a sampling-based method SMOTE-XGBoost, which combine one sample synthetic method SMOTE and one strong classifier XGBoost, to improve the prediction performance of postoperative complications and merge more features into the binary classification model to further improve the performance.

- **A Ensemble Learning Method Multiple Balanced Subsets Stacking** Since most of the exiting imbalance learning approaches have different kinds of issues, such as, the over-fitting problem of over-sampling methods, the information loss problem of under-sampling methods, and poor generalization ability of cost-sensitive methods. This work proposes a novel ensemble method to alleviate the class imbalance problem and avoid those problems of existing methods in a large extent.

- **A Re-weighting Method Hardness Aware Dynamic Loss Function** When dealing with the imbalanced image datasets, over-sampling methods will introduce great computation cost and training time cost, and under-sampling methods might loss important samples. After demonstrating the issues of exiting re-weighting methods, this work proposes a novel loss function which dynamically customizes the class weight by the classification hardness during the training process of the deep neural network.

1.2 Motivation

In this section, we list the the motivations of three works on medical classification under class imbalance in the dissertation.

1.2.1 The Class Imbalance Problem of A Medical Dataset

There are over 300 million operations performed worldwidely each year. Operation poses considerable risk of postoperative complications, which could worsen the quality of patients' life, even incurring prohibitively expensive costs. As mentioned in Chapter 1.1, as half of the complications are preventable, accurate prediction of postoperative complications is highly important for clinical decision making, early treatment and counseling patient. With the abundance of medical datasets, machine learning approaches have been applied to predict postoperative complications of different diseases, such as stroke[56], cancer[46], bleeding, shock, cardiac[73, 126], acute kidney injury and sepsis[106]. These studies mainly focus on feature selection[56, 106], feature sparseness(missing value)[128, 126].

Postoperative complication distribution of most diseases are highly imbalanced, which would cause the prediction models bias towards majority class and ignoring the minority class[38]. Moreover, existing postoperative complication prediction models, such as multivariate logistic

regression and machine learning classifiers, are usually optimized and evaluated using overall accuracy or error rate, which are not suitable for imbalanced datasets[117], thus limiting the performance of respective models. To solve the class imbalance problem of the postoperative complications, we use kidney stone disease as a study case.

Kidney stone disease (also known as nephrolithiasis) is a worldwide public health problem. Studies report that the incidence of kidney stone disease is globally increasing in 5 European countries, Japan, China and the United States. More and more patients with large kidney stones have been treated by Percutaneous Nephrolithonomy (PCNL) since its introduction in 1976. According to a global study of the PCNL [65], there were 1175 of 5724 (20.5%) patients experienced one or more complications after PCNL operation, which makes the postoperative complication prediction a class imbalance problem. Furthermore, in spite of the class imbalance problem, there are some other limitations of previous works on postoperative complication prediction of PCNL.

Postoperative complication of PCNL could worsen the quality of patients' life, even incurring prohibitively expensive costs. Thus, it is of great importance to build a system that could predict the postoperative risk accurately, which would also be precious for clinical decision making and patients counseling. One of the main limitations of the existing prediction models is only using limited features. There are three commonly used score systems, the Guy's stone score [105], the S.T.O.N.E. (stone size, tract length, obstruction, number of involved calices and essence) nephrolithometry [85], and CORES (clinical research office of the endourological society) nomogram [101], that are used as predictors of stone-free status and postoperative complication of PCNL. While the score systems are designed for stone-free status prediction, using kidney stone related features is enough to build a prediction model. However, when such systems are used to predict postoperative complications, the ignorance of other complication related features will degrade the prediction performance [112, 64]. A systematic review and meta-analysis of three score systems conclude that they are equally accurate and feasible for predicting stone-free status after PCNL, however, the results of predicting postoperative complication of PCNL are controversial[50]. Furthermore, although risk factors of the complication are identified by univariate or multivariate analysis with using statistical logistic regression, no prediction model has been built to predict the postoperative complications of PCNL based on these risk factors.

In Chapter 4, we first perform a detailed analysis of PCNL patient's features and compare the feature difference between the two groups under the class imbalance situation. We implement a sampling-based method, proposed in Section 3.2.1, to build a new postoperative complication prediction model which is able to deal with the imbalance problem. More features are added to the proposed model for better performance.

1.2.2 Issues of Existing Machine Learning Methods under Class Imbalance

Achieving accurate medicine and improving the quality of patient care are the overall objective in healthcare area. With the rapid increasing application of electronic health records in many healthcare facilities, it is possible to get enough medical data to achieve this goal more efficiently. Nevertheless, prediction based on medical dataset has been an intriguing and challenging topic because of its inherent imbalanced nature. Medical datasets are mainly composed of "healthy" samples with only a small section of "sick" samples, leading to the so called class imbalance problem. The imbalance problem could bias classification algorithms to majority class, so that classifiers have weak performance on minority class. Such classifiers are not useful in real world tasks, because usually the classification performance of the minority samples is of higher importance for decision making in the healthcare area [8].

A series of imbalance learning methods have been proposed to overcome the imbalance problem and can also be clustered into three main classes: data-level approaches (e.g., sampling), algorithm-level approaches (e.g., cost-sensitive learning) and hybrid approaches (e.g., ensemble learning). Sampling approaches have been proved effective on imbalance classification tasks, such as, chronic kidney disease prediction [129], diabetes and liver disorders prediction [70]. As elaborated in Section 5.1, existing sampling methods suffer from problems, such as information loss, huge computational cost and overfitting. The challenge of cost-sensitive methods is how to determine a cost matrix, but the defined cost matrix may not be generalized to any other tasks. Ensemble learning approaches usually combine sampling approach or cost-sensitive approach with ensemble learning algorithm to address the imbalance problem [52, 33]. However, they inherently suffers from issues of sampling approaches and cost-sensitive approaches. Moreover, some ensemble methods have the problem of high training cost when applied on large real world tasks, as shown in SMOTEBagging [119] and SMOTEBoost [15].

Taking these issues of existing methods into consideration, we propose a novel ensemble learning method called Multiple bAlance Subsets Stacking (MASS) in Chapter 3.2.2 and evaluate it on three structured medical datasets in Chapter 5.

1.2.3 Issues of Existing Deep Learning Methods under Class Imbalance

Deep neural networks (DNNs) have been proved very successful in computer vision domain [66]. In addition to the improved computation ability and various algorithms breakthroughs, the wide availability of labeled image datasets is another key reason for the success. Lots of the labeled image datasets, such as MNIST and CIFAR, are commonly resembled to be nearly balanced. However, class distribution of real-world image datasets is naturally imbalanced, medical image datasets are the typical examples. For instance, the number of healthy cases (majority) usually dominates that of lung cancer cases (minority) for critical

applications like medical diagnosis [131]. As a result, there will be a significant drop when DNNs are applied on such real-world datasets. Trained with imbalanced datasets, conventional DNNs would bias towards the majority classes, which would lead to poor accuracy for the minority samples. Nevertheless, failing to classify a patient might lead to the loss of life. Thus, it is of great importance to improve the classification performance of the DNNs on minority classes.

Previously, researchers usually use data-level methods (re-sampling) or algorithm-level methods (re-weighting) to tackle the imbalance problem. As described in Section 1.1, re-sampling methods include over-sampling for the minority classes, under-sampling for the majority classes, or hybrid sampling for both majority and minority classes; Re-weighting methods assign relatively larger weights to minority samples, which would influence the loss function to focus more on the minority classes. In the context of computer vision applications, over-sampling methods introduce large training costs and make the model prone to overfit the minority classes. Under-sampling methods discard important samples that are valuable for deep representation learning. Taking these issues of applying re-sampling methods on image classification tasks into consideration, our work focuses on designing a better re-weighting method to improve the accuracy of minority classes.

As minority classes are weakly represented with fewer samples [22, 121], re-weighting methods for imbalance problem penalize classifiers more seriously for misclassification of minority samples compared with those of majority samples. Re-weighting methods assign sample weights in inverse proportional to the class frequencies or the square root of class frequencies, which are proved efficient [39]. However, when applying on large real-world imbalanced datasets, re-weighting methods perform poorly [75]. One main reason might be that some minority classes are well represented by a small size of training data. Under this situation, resetting the weights in inverse proportional to the class frequencies (called overweighting) will decrease the overall performance. Thus, it is of great importance to find out the optimal weight for each class to achieve higher classification performance.

In Chapter 6, we introduce the problem of re-sampling methods and demonstrate re-weighting by class frequency is not always a good option to set weights to alleviate the imbalance problem. We come up a novel loss function which re-weight the class weight by classification hardness in Section 3.3. Then the loss function is evaluated on four imbalanced image datasets.

1.3 Contribution

In this section, we list the main contributions of three studies on the class imbalance problem in this dissertation.

1.3.1 A Sampling-based Method SMOTE-XGBoost

To fix the limitations of PCNL prediction models and the class imbalance problem, which are described in Section 1.2.1, this dissertation first conduct a detailed analysis of the patient features and then propose sampling-based method SMOTE-XGBoost to build a novel Postoperative Complication Prediction model on PCNL dataset. SMOTE-XGBoost uses SMOTE[14] to rebalance the training set and then sends the resampled training set to train XGBoost[17] in oder to predict the postoperative complications. Additionally, instead of using accuracy or error rate as evaluation metrics, we use AUC (also called c-statistic) and F1-score to evaluate our prediction model. To the best of our knowledge, this is the first work focusing on the postoperative complication prediction of PCNL with considering the class imbalance problem.

We evaluate the proposed model on a large collection of real PCNL patients' records spanning from January 2012 to July 2019. Experiment results indicate while only using kidney stone related features, our model significantly outperforms the S.T.O.N.E. nephrolithometry and classic machine learning methods over both AUC and F1-score. Furthermore, we add other complication related features to our model, which further improves the prediction performance. Altogether, our model achieves an AUC of 0.7077 to predict postoperative complication, which is 41.54% higher that of S.T.O.N.E. nephrolithometry.

To sum up, the main contributions of this study could be listed as follows:

- A thorough analysis of 3292 patients with large kidney stones treated by Percutaneous nephrolithotomy;
- Compared the features of the patients according to the postoperative complications;
- Propose A sampling-based method SMOTE-XGBoost and implement it to solve the class imbalance problem of the postoperative complication;
- Conducted extensive experiments to verify the effectiveness of SMOTE-XGBoost over baseline methods;
- Merged more related features into the prediction model and further improved its classification performance.

1.3.2 An Ensemble Learning Method Multiple Balanced Subsets Stacking

As mentioned in Section 1.2.2, existing imbalance learning methods may suffer from issues like information loss, overfitting, and high training time cost. To tackle these issues, in Section 3.2.2, we propose a novel ensemble learning method called Multiple bAlance Subsets Stacking (MASS). Other than simply creating a balance training set or defining a cost matrix, MASS first

generates multiple balance subsets to train base classifiers. Then MASS generates a stacking dataset based on the base classifiers, which keeps the same label as original dataset. After that, the stacking dataset is used to train a stack model, which could optimize the weights of the base classifiers to get a strong ensemble classifier. MASS does not reduce majority samples or generate new meaningless samples, thus will avoid the problem of information loss. Futrhermore, MASS does not duplicate any minority samples, thus avoids the issue of overfitting to the minority class. Specially, as the training processes of the base classifiers and the stacking dataset generation are independent, the main part of MASS could run in parallel. Hence, we propose a parallel version of MASS called Parallel MASS to decrease the training time cost, which is of high importance as the scale of healthcare dataset is increasing rapidly.

In Chapter 5, we extract three real-world healthcare datasets, namely acute kidney failure and diabetes from MIMIC (Medical Information Mart for Intensive Care) III dataset and PCNL dataset collected from the First Affiliated Hospital of Gannan Medical University in China. We conduct extensive experiments to evaluate the classification performance of MASS by comparing it with other baseline methods on these three structured datasets. Besides, to validate the robustness of MASS, we apply MASS and other ensemble learning methods with different base classifiers. Finally, we analyze the speedup of Parallel MASS over MASS on different scales of PCNL dataset.

In conclusion, this study mainly has the following contributions:

- Proposed a ensemble learning method Multiple bAlance Subsets Stacking (MASS) to solve the imbalance problem via multiple balance subsets constructing strategy, and improve it to a parallel version (Parallel MASS) to reduce the training time cost.
- Conducted extensive experiments to evaluate the proposed MASS. Experimental results show that MASS greatly outperforms baseline methods on three different real world healthcare datasets. For example, compared with SPEnsemble [74], MASS improves the classification performance 3.22% in AUC, 3.10% in F1score,improves 2.58% in MCC when applied to the diabetes dataset.
- Validated the robustness of MASS by comparing it with other ensemble learning methods with applying different base classifiers, and the experimental results show that MASS always outperforms other baseline ensemble methods.
- Analyzed the speedup of running Parallel MASS over different scales of dataset. The results demonstrate that running MASS in parallel can reduce the training time cost greatly on large datasets, and its speedup would increase as the data size grows.

1.3.3 A Re-weighting Method Hardness Aware Dynamic Loss Function

Although DNNs have achieved great success in image classification tasks with balanced image datasets, they perform poorly on highly imbalanced image datasets. To solve the class imbalance problem, most existing methods leverage class frequency to rebalance the dataset or resize the class weight. However, while some of the minority classes could be well represented by the training data, re-sampling or re-weighting such classes will decrease the overall performance.

In Section 6.1, we first demonstrate the weakness of re-weighting the class weights by class frequencies. To address the challenges described in Section 1.2.3, we consider using class-level classification hardness to decrease the impact of noise samples rather than sample-level hardness. In Section 3.3, we propose a re-weighting method called Hardness Aware Dynamic (HAD) loss to resize the class weight of a sample in the loss function dynamically by the classification hardness of its class during the training process of DNN. After each training epoch of a deep neural network, we could measure the correctly classified probability for each sample. Then we define the classification hardness of this sample as its misclassification probability, which equals 1 minus its correctly classified probability. Next, we compute the average value of classification hardness of different classes. The average classification hardness values are used to update class weights following the rule that increase class weights with larger average classification hardness values and decrease class weights with smaller average classification hardness values. In Chapter 6, this thesis conduct extensive experiments on imbalanced subsets of two standard image datasets (MNIST, CIFAR-10) and two imbalanced medical image datasets (i.e., Breast Cancer dataset and Skin Cancer MNIST:HAM10000). The experimental results indicate that HAD loss can provide a significant improvement to the classification performance of recently proposed loss functions for training deep learning models.

In summary, the main contributions of this work are:

- We introduce a new *class-level* classification hardness, which captures the classification hardness of each class of the model and alleviates the negative effect of noise samples;
- Based on class-level classification hardness, we propose a novel loss function called HAD loss for improving the imbalanced image classification, which updates class weights dynamically during the training process of DNNs and finds optimized weight for each class;
- We show that HAD loss achieves significant improvement compared with baselines over F1-score and G-mean on the imbalanced medical image datasets, and prove its robustness over several datasets of different imbalance degrees. Especially, HAD loss improves macro-precision from 35.26% to 38.80% compared with the best baseline on Skin Cancer MNIST;

- Overall, HAD loss on quantifying the classification hardness of each class and using it to update class weights dynamically can provide helpful guidelines for researchers working on imbalanced image classification tasks.

1.4 Content Guide

This thesis includes contents of one published paper and two submitted papers.

- **Yachao Shao**, Xiaoning Wang, Xiaofeng Zou and Xiaoming Fu. "Postoperative Complication Prediction of Percutaneous Nephrolithonomy via Imbalance Learning." Artificial Intelligence in Medicine 2021 (Under review)[98].

- **Yachao Shao**, Tao Zhao, Xiaoning Wang, Xiaofeng Zou and Xiaoming Fu. "Multiple Balance Subsets Stacking for Imbalanced Healthcare Dataset." In 26th IEEE International Conference on Parallel and Distributed Systems (ICPADS). pp. 300-307. IEEE, 2020 [97].

- **Yachao Shao**, Tao Zhao, Jiaquan Zhang, Shichang Ding and Xiaoming Fu. "Hardness Aware Dynamic Loss on Imbalanced Image Classification." In 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021 (Under review)[96].

Many thanks to the collaborators from Gannan First Affiliated Hospital of Gannan Medical University, they collected the unstructured clinical notes of kidney stone patients who were treated by Percutaneous Nephrolithonomy (PCNL). We name this dataset as PCNL dataset. Based on the PCNL dataset, we extract a structured dataset and conduct a through analysis. I propose a sampling-based method to predict the postoperative complication and finish a paper, which is submitted to Artificial Intelligence in Medicine.

- Chapter 1 introduces the background of this dissertation at the beginning in Section 1.1. Then the motivations of each work respectively are listed in 1.2. Section 1.3 summaries the contributions of each work. At last, Section 1.4 presents the content guide of this dissertation.

- Chapter 2 first describes the nature of class imbalance problem and lists evaluation metrics for imbalance learning methods. Section 2.3 and Section 2.4 review the existing works in machine learning and deep learning to alleviate the class imbalance problem respectively. Section 2.5 briefly introduce the artificial intelligence applications on medical datasets and challenges accordingly.

- Chapter 3 provides an overview of three proposed imbalance learning methods in Section 3.1. A sampling-based method on imbalanced structured datasets in Section 3.2.1. Section 3.2.2 presents details of the proposed MASS and parallel MASS on imbalanced structured dataset. A re-weighting method for imbalanced image datasets is introduced in Section 3.3.

- Chapter 4 mainly focus on analyzing the features of patient treated by PCNL according to the postoperative complications, then fixing the class imbalance problem of postoperative complications. Section 4.1 introduces the importance of accurate prediction models in healthcare area and presents the limitations of existing prediction models of PCNL. Section 4.2 shows the statistical methods for analyzing features of patients and build a postoperative complication prediction model based on the sampling-based method SMOTE-XGBoost to alleviate the class imbalance problem. Section 4.3 presents the statistical results of the patient's features and the results of the comparison between SMOTE-XGBoost and other baselines. Section 4.4 concludes this chapter.

- Chapter 5 focus on evaluating the ensemble learning method MASS on three medical datasets. Section 5.1 introduces the challenges of classification under class imbalance and the contributions. Section 5.2 shows the processing progress of three structured medical datasets. Section 5.3 implements MASS on these three medical datasets, and proves the effectiveness of MASS by the experimental results. Finally, a summary is listed in Section 5.4.

- Chapter 6 focuses on dealing with the class imbalance problem of image datasets via dynamically customizing the class weight during the training process of the deep neural networks. Section 6.1 describes the challenges of applying DNNs in real-world image datasets, which are inherently imbalanced. Section 6.2 presents the process of four different image datasets. Section 6.3 evaluates the performance of HAD loss with both standard image datasets (MNIST, CIFAR-10) and two medical datasets (Breast Cancer, HAM10000). Section 6.4 concludes this chapter.

- Chapter 7 summarizes the contributions of this dissertation and provides plans for future work.

Chapter 2

Background and Related Works

In many real world applications, such as diagnosis of rare disease, fraud detection, image classification, the class distribution is skewed, which lead to the class imbalance problem. Imbalance learning refers to the methods used to deal with the class imbalance problem. Although many datasets have more than two classes, we mainly discuss the binary classification problem in this dissertation, since solving the binary classification problem is the base of multiple classification tasks.

In this Chapter, the nature of class imbalance is described in Section 2.1. Section 2.2 introduces evaluation metrics for imbalance learning methods. Then we introduce related works of imbalance learning methods in machine learning, which is suitable for structured datasets. The existing work could be categorized into three groups: data-level methods, algorithm-level methods and hybrid methods. Next, we introduce related works of imbalance learning methods in deep learning, which are classified similarly to machine learning methods. At last, related works of applying artificial intelligence on medical datasets are introduced in Section 2.5.

Contents

2.1 Nature of The Class Imbalance Problem

In this section we will discuss the nature of the class imbalance problem in three aspects:

- **Intrinsic Imbalance vs. Extrinsic Imbalance** As described previously, in a lot of real world applications, the datasets are inherently imbalanced and such imbalance is named as intrinsic imbalance. On the contrary, the extrinsic imbalance means that observed dataset is imbalanced while the original dataset is balanced, which is caused by the external factors, such as data collection or data storage. For example, if we collect a continuous stream of balanced data, and the received dataset might be imbalanced if the data transmission is not stable during the collection. Extrinsic imbalance usually could be fixed by collecting more samples, which is not suitable for intrinsic imbalance as the intrinsic imbalanced dataset is originally imbalanced. We mainly focus on the intrinsic imbalance and take medical datasets as study cases in this dissertation.

- **Between-class Imbalance vs. Within-class Imbalance** Normally, the class imbalance refers to the between-class imbalance, where the dataset is composed by the majority class and the minority class [38]. The minority class is severely under-represented by less samples compared to the majority class. However, the misclassification cost of a minority sample is usually much lager than that of a majority sample, as explained by the cancer diagnose in Section 1.1. Another type of class imbalance happens within a class, named within-class imbalance, which means that there are several sub-clusters in one class and the distribution of these sub-clusters is skewed. In this dissertation, we clarify that the class imbalance indicates the between-class imbalance for a clear understanding.

- **Absolute Imbalance vs. Relative Imbalance**Absolute imbalance refers to the under-representation of the minority class due to the lack of data. On the other hand, the relative imbalance refers to the minority class is well represented, which could hardly affect the classification performance [38]. Consider a dataset including 100,000 samples, the minority class accounts for 1%. This dataset seems to be severely imbalanced, 1000 minority samples might be able to describe the minority class quite well. Nevertheless, it is difficult to identify whether a dataset is absolute imbalance or relative imbalance. We will discuss the relative imbalance in an image classification task in Chapter 5.1. In that application, one class with fewer samples has higher performance than one class with more samples, which indicates the prior class is relative imbalanced.

2.2 Evaluation Metrics for Imbalance Learning

For binary classification problems, the classification results are composed of four categories based on the ground truth labels and prediction labels. True Positive (TP) is the number of correctly classified positive samples, while False Negative (FN) is the number of misclassified positive samples. True Negatives (TN) is the number of correctly classified negative samples

while False Positive (FP) is the number of misclassified samples. We can store the classified results in a confusion matrix as shown in Table 2.1.

Table 2.1: Confusion matrix for the binary classification Tasks

	Positive prediction	Negative prediction
Positive class	True Positives (TP)	False Negatives (FN)
Negative class	False Positives (FP)	True Negatives (TN)

Traditionally, the most commonly used metrics to evaluate the performance of classifiers are accuracy and error rate. However, they are not suitable when dealing with imbalanced distributed problems.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$ErrorRate = \frac{FN + FP}{TP + TN + FP + FN} = 1 - Accuracy \tag{2.2}$$

For example, if a data set include 99% of majority class samples and only 1% minority class samples, a naive solution is to classify every sample into majority class, and the accuracy would be 99% and error rate would be 1%. It is pretty good at the first glance, however, both accuracy and error rate fail to tell that there are no minority class samples correctly classified. Thus, we need to use other evaluation metrics to access classifiers' performance on the imbalanced problem.

Several metrics have been proposed to evaluate classification performance in imbalance learning, such as Precision, Recall and Specificity:

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

Precision measures the percentage of all positive predicted samples that are correctly classified. Precision is a good evaluation metric for the imbalance problem, because it takes misclassified negative samples (FP). However, Precision alone is not sufficient enough because it neglect the misclassified positive samples (FN). Recall, on the contrary, measure the percentage of all positive samples that are correctly classified. Recall is not sensitive to class imbalance because it only considers positive samples. Specificity represents the percentage of all negative samples that are correctly classified. Thus, to better evaluate the classification under class imbalance,

previous studies propose F-measure and G-mean, two evaluations that combine precision, recall or specificity in different forms.

F-measure is defined as the weighted harmonic mean of the precision and recall, defined as formula 2.6, where β is a coefficient used to tune the relative importance between precision and recall (when β=1, F-measure is the widely used metric F1-score).

$$F - measure = \frac{(1+\beta)^2 \times Precision \times Recall}{\beta^2 \times Recall + Precision} \tag{2.6}$$

G-mean, defined as formula 2.7, considers a balancing between accuracy of positive samples and accuracy of negative samples, and is appropriate to evaluate imbalance learning.

$$G - mean = \sqrt{Recall \times Specificity} \tag{2.7}$$

The Receive Operating Characteristic curve (ROC) plots true positive rate (TPR) over over false positive rate (FPR), which visualizes the trade-off between correctly classified positive samples and misclassified negative samples, i.e., the benefits and costs.

$$TPR = \frac{TP}{TP + FN} = Recall \tag{2.8}$$

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity \tag{2.9}$$

For classifiers which generate continuous predictions, changing the threshold can generate a series of points in the ROC space, for example in Fig. 2.1.

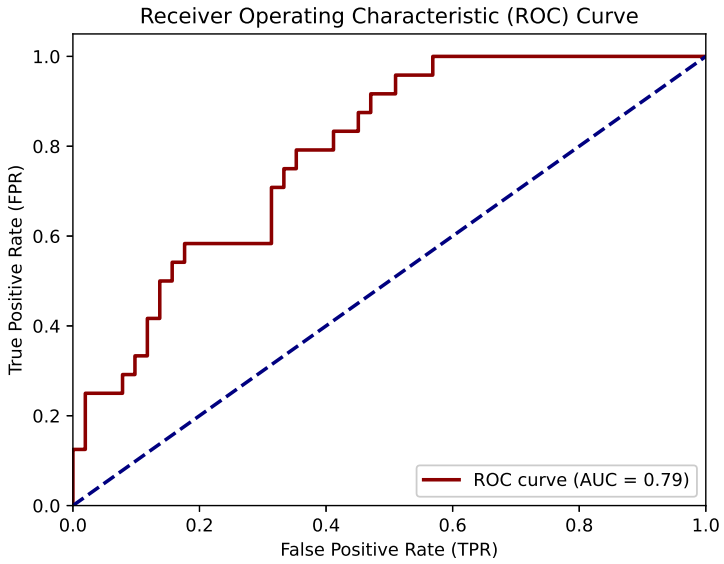

Figure 2.1: An example of Receive Operating Characteristic (ROC) curve and AUC score on one binary classification task

From ROC curve, the ideal scenario is that the TPR is always 1,which means the classifier could perfectly identify positive class, no matter how FPR changes. Hence, classifier performs better when its ROC closes to the left top corner. Therefore, we use the The Area Under the ROC Curve (AUC) to evaluate the classification performance. AUC is a numerical representation and has been proved to be a reliable metric for evaluating classification on imbalanced data set[32].

The Matthews correlation coefficient (MCC) [78], defined as equation 2.10, is proposed by Brian W. Matthews in 1975, is also a good metric to evaluate the performance of classifiers under class imbalance. MCC produces a high score only if a classifier is able to correctly classify most of positive samples and most of negative samples. MCC has been proved more reliable and more informative than F1-score and g-mean in a genomics study [18].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.10}$$

2.3 Machine Learning Methods under Class Imbalance

Through the last two decades, extensive studies have been conducted to address the class imbalance problem using traditional machine learning approaches. As described previously in the first section of Chapter. 1, class imbalance would bias the standard machine learning algorithms to the majority class. This problem could be alleviated by changing the distribution of training data to decrease the imbalance, or by altering the learning or decision process to increase the influence of the minority class. Accordingly, machine learning methods under class imbalance could be categorized into three groups, i.e., data-level methods, algorithm-level methods and hybrid methods. Some popular methods are summarized in this section.

2.3.1 Data-level Methods

Data-level methods alleviate class imbalance by changing the distribution of training data to decrease the imbalance degree. Most of these methods could be grouped into three kinds, over-sampling, under-sampling and hybrid sampling, i.e., using over-sampling and under-sampling simultaneously. Random Over-Sampling (ROS) and Random Under-Sampling (RUS) are two elementary forms of data-level methods. ROS randomly drops majority samples while RUS randomly duplicates minority samples [111].

Over sampling methods increase minority samples, they will increase the training time because of the increased size of training data. Meanwhile, ROS has been proved with over-fitting problem [16], which occurs if a model is poorly generalized to new data because the model is trained to fit the training data too closely. Under-sampling methods discard samples from the majority class to generate a balanced training set, which will lead to the information

loss. To balance these trade-offs, previous works have proposed a variety of intelligent sampling methods.

Intelligent over-sampling methods have been developed to alleviate the over-fitting problem and increase discrimination ability. Synthetic Minority Over-sampling TEchnique (SMOTE) interpolates new samples between minority samples and several nearest minority neighbors [14]. Some variants of SMOTE, such as, Borderline-SMOTE [36] and Certainty Guided Minority Over-Sampling (CGMOS) [130], have been proposed to improve the original SMOTE by taking both majority class and minority class into consideration. Borderline-SMOTE limits the interpolated samples near class borders while CGMOS considers both classification performance of the minority and that of the majority class.

A variety of intelligent under-sampling approaches have also been proposed to alleviate the imbalance degree while keeping the valuable information for training the model. For instance, Near-Miss performs under-sampling based on the distance of the distance between the majority samples and minority samples [76]. Another way to implement intelligent under-sampling is data cleaning. Such methods firstly identify noisy samples and overlapping regions, then remove samples accordingly. One-Sided-Slection (OSS) [63] removes noisy samples from majority class and redundant majority samples which are identified using 1-nearest-neighbor classifier and Tomek Links [107]. The major disadvantage of these intelligent sampling strategies is high computation cost, especially when they are applied to large datasets.

An experimental work has been conducted to compare seven sampling techniques over 11 different machine learning algorithms on 35 imbalanced benchmark datasets [111]. Six evaluation metrics has been used to compare the results. From this study, the performance improvement is highly dependent on the machine learning algorithms and the evaluation metrics. The results reveals that RUS outperforms other six sampling methods in most cases, thus has better overall performance. However, RUS is not always best in all cases, which suggests that no sampling method is guaranteed with best performance in all problem domains. Meanwhile, the performance should be compared with using different evaluation metrics.

2.3.2 Algorithm-level Methods

Different from data-level methods, algorithm-level methods for dealing with class imbalance problem do not change the distribution of training data. Alternatively, they are developed to fix the imbalance problem by increasing the importance of the minority samples during the training process. Most typically, the algorithms are altered to take class weights or misclassification cost into consideration, or shifting the decision threshold to reduce the bias towards the majority class.

Among the algorithm-level methods, cost-sensitive learning is the most typical one. Cost-sensitive learning assumes that the misclassification cost of minority samples are higher that that of the majority samples. A cost matrix is defined to assign misclassification cost to different

classes, and the cost matrix under binary classification situation is shown in Tabel 2.2. In the cost matrix, c_{ij} is the classification cost when the prediction is j while the ground truth label is i. Normally, the cost of correctly classified is set to 0, where $i = j$. The misclassification costs of majority class and minority class could be fine-tuned for desired results. Increasing the misclassifcation cost of one class is identical to increasing its importance, which means that the algorithm would have higher classification performance of this class [60].

Cost-sensitive learning methods can be categorized into two kinds. The first kind of cost-sensitive methods use cost matrix to rearrange the decision threshold and assign different sampling rates to different class. For example, if the prediction result of a cost-insensitive binary classifier is posterior probability, we could reset the decision threshold as θ according to the cost matrix:

$$\theta = \frac{c_{01}}{c_{10} + c_{01}} \tag{2.11}$$

Normally, new threshold θ is used to adjust the output decision threshold when discriminating samples from different classes [72]. For example, researchers proposed to find the optimized classification threshold instead of setting it as 0.5 [132]. Threshold redefinition using Equation 2.11 is one approach that transforms cost-insensitive classifiers to cost-sensitive classifiers.

The other kind of cost-sensitive methods is converting the optimization object function from minimizing the total error to minimizing the total cost. For instance, a cost-sensitive decision tree ensemble method is developed by incorporating the misclassification cost [61]. Likewise, Cost-Sensitive Large margin Distribution Machine (CS-LDM) improves the classification performance by incorporating cost-sensitive margin mean and cost-sensitive penalty.

However, compared with data-level methods (e.g., over-sampling and under-sampling), there is not much attention on cost-sensitive learning methods mainly because it is very challenge to define an effective cost matrix. A common strategy to set the cost matrix is fixing the misclassification cost of majority class at 1 and that of the minority class at the imbalance ratio. The main problem of cost-sensitive methods is that the cost matrix definition needs domain experts' assistance before hand, which is often not available in real world cases. Another problem is that cost-sensitive methods usually need specific modification in the algorithm, which is much harder than sampling methods.

Table 2.2: Cost matrix in binary classification problem

	Positive prediction	Negative prediction
Positive class	$C(1,1) = c_{11}$	$C(1,0) = c_{10}$
Negative class	$C(0,1) = c_{01}$	$C(0,0) = c_{00}$

2.3.3 Hybrid Methods

In order to take both advantage of data-level methods and algorithm-level methods, a number of studies have been conducted to combine them in different ways to alleviate the class

imbalance problem [60]. Typically, hybrid methods firstly perform data sampling to remove noisy samples and decrease imbalance degree, continue with implementing cost-sensitive methods to further improve the overall classification performance. Moreover, data-level methods and algorithm-level methods are usually combined with ensemble methods. Ensemble methods could get better classification performance by ensemble several weak classifier as a strong one. Bagging [10], boosting [15] and stacking [124] are three main forms of ensemble methods.

For Bagging, the training dataset is sampled with replacement by bootstrapping in each iteration. Then the training subset is sent to train one classifier in each iteration. Finally, the prediction result is decided by majority voting from all trained base classifiers. Consequently, the variety of training datasets can help avoid overfitting and reduce variance, thus achieve better classification performance. OverBagging [119] and UnderBagging [6] are two representative bagging methods. OverBagging (UnderBagging) adopt over-sampling (under-sampling) in the bootstrapping step to build balanced training subsets.

Different from bagging, boosting combines weighted weak classifiers generated by training with weighted samples into one strong classifier. The most representative method of boosting methods is Adaptive Boosting (AdaBoost) [39]. In each iteration, if one sample is correctly classified, AdaBoost decreases its weight, and vice versa. The weights of the weak classifiers are assigned by the cost function, which means that the prediction result is decided by weighted majority voting. Three different cost-sensitive version of AdaBoost (AdaC1,AdaC2,AdaC3) are proposed [103]. They methods incorporate misclassification cost into the weight update steps of AdaBoost to increasing the impact of minority samples iteratively. SMOTEBoost [15] uses SMOTE to generate balanced training sets in each boosting iteration, while RUSBoost [95] uses under-sampling.

2.4 Deep Learning Methods under Class Imbalance

Deep learning methods have achieved great success in in areas such as image and speech recognition [66] over the last decade. The effect of class imbalance has been studied in the 1990's [3]. This work proves that the majority class dominates the gradient of shallow neural networks in the backpropagation step, which means the neural network is more sensitive to the error of majority class. Accordingly, the error majority class reduces faster than that of the minority class in the early iterations, which often leads to the neural network bias towards the majority class. Similar to the categorization of machine learning methods under class imbalance, related works of deep learning methods to deal with the imbalance problem are categorized into three classes, data-level methods, algorithm-level methods and hybrid methods.

2.4.1 Data-level Methods

Data-level methods alter the class distribution to generate rebalanced datasets in the pre-processing procedure. This strategy is attractive because it is easy to implement and there is no need to change the deep learning algorithms. Data-level methods include over-sampling minority samples, under-sampling majority samples or using both [45, 68, 127, 89, 82].

ROS has been proved efficient to improve classification performance of the deep Convolutional Neural Networks (CNNs) on imbalanced image datasets[45]. In this study, ten imbalanced image datasets with different imbalance ratios is generated from the CIFAR-10 benchmark data. Then, a variant of the CNN (AlexNet) is trained to classify the images. The experimental results shows that imbalance distribution leads to a loss in classification performance, which verifies that imbalance datasets have negative impact on the classifier. To balance the datasets, ROS randomly duplicates minority samples until the number of each class is equal. The ROS classification results are comparable to the standard dataset, which suggests that ROS is effective to deal with imbalanced image datasets. However, the max imbalance ratio in this study is only 2.3, which is the biggest limitation.

Likewise, RUS is used to remove majority samples in the pre-training phase of a two-phase learning, which improves the performance of minority class while preserving the classification performance on majority class [68]. This work firstly set a threshold of N examples, then randomly under-sampling all large classes to the threshold. At the first phase, a deep CNN is trained by the under-sampled dataset, then it is fine tuned by the original dataset at the second phase. Different from plain RUS, the majority samples are only removed during the pre-training in the first phase. The experiments results shows that the two-phase learning improves the performance of minority classes while keeping that of majority classes.

Hybrid sampling combines over-sampling and under-sampling to alleviate the imbalance problem. By using F1-scores of different classes to adjust the sample number in the next iteration respectively, dynamic sampling assigns a higher sampling rate to classes with lower F1-score, thus the model could focus more on poor classified classes [89]. The dynamic sampling strategy is defined as equation:

$$Sample - size_{i,j} = \frac{1 - f1score_{i,j}}{\sum_{c \in C}(1 - f1score_{i,c})} \times N^* \tag{2.12}$$

$Sample - size_{i,j}$ represents the number of samples for class j on the ith iteration. $f1score_{i,j}$ represents the F1-score of class j on the ith iteration. N^* is the mean value size of all classes. With using dynamic sampling, the transfer learning used in this study is able to self-adjust the sampling rate during the training, thus achieving higher averaged F1-score and better classification performance.

2.4.2 Algorithm-level Methods

These methods modify the learning procedure of deep learning algorithms to improve the classification performance of minority classes. Algorithms level methods could be categorized into three kinds: cost-sensitive learning methods, threshold changing methods and new loss functions.

Cost-sensitive learning methods combine cost matrix with Cross-Entropy (CE) to improve the sensitivity of minority classes. Cost-sensitive deep neural network (CSDNN) incorporates a pre-defined cost matrix to improve the prediction accuracy of hospital readmissions [116]. However, it is time-consuming to find out the best cost parameter in the cost matrix. The cost matrix is task-specific and can not be generalized to other tasks. To overcome these issues, the Cost-Sensitive Convolutional Neural Network (CoSen CNN) can learn weight parameters of neural network and the cost parameters jointly during the training process [55]. Threshold changing is compared with ROS, RUS when dealing with datasets with different imbalance ratios. The misclassification probability of the minority class is reduced effectively when the threshold of one class is divided by its prior estimated probability [12].

To make the model more sensitive to the misclassification of minority samples, new loss functions are introduced, such as Mean False Error (MFE) loss [118], Focal loss [71], Class-Balanced (CB) loss [22] and Class-wise Difficulty Balance (CDB) loss [100].

When the Mean Squared Error (MSE) is used as the loss function, the error of misclassified minority samples is weaken by that of the majority samples. MFE loss has been developed with modifying the loss function to alleviate the imbalance problem. It is composed by the mean False Negative Error (FNE) and the mean False Positive error (FPE). MFE and its improved version Mean Squared False Error (MSFE) have been verified on different datasets that they are not only able to deal with the imbalance problem, but also are easy to implemented and low computation cost [118].

Focal loss is a modified loss function of cross entropy (CE) to reduce weight of the easily classified samples and focus more on samples that are hard to be classified, as defined by equation 2.13, where p is the correctly classified possibility of one sample. The hyper parameter $\gamma \geq 0$ is used to control the lower the importance degree of easily classified samples, whereas α is used to increase the importance degree of the hardly classified samples, most of which are minority samples. Focal loss outperforms several one-stage and two stage deep learning algorithms on the COCO dataset [71].

$$\text{Focal}(p) = -\alpha \left(1-p\right)^{\gamma} \log\left(p\right). \tag{2.13}$$

To further improve the classification performance of deep neural networks under class imbalance, CB Loss is proposed to adjust existing loss function, such as CE loss and Focal loss, based on the inverse of effective sample number [22]. The authors first introduced the

effective number of one group, then merged its inverse value into the existing loss functions, lastly showed the improvement brought by CB loss.

Different from previous re-weighting methods, CDB loss is proposed to change the weight of each class dynamically in every iteration during the training process of the deep neural networks. The authors define the precision of each class as its classification difficulty. Then, after training epoch t, all precision are calculated and the weight of class c is set as equation 2.15.

$$\mathrm{CDB}_{c,t}(p) = -(Difficulty_{c,t})^{\gamma} \log(p) \tag{2.14}$$

$$= -(1 - Precision_{c,t})^{\gamma} \log(p) \tag{2.15}$$

2.4.3 Hybrid Methods

These methods combine data-level methods and algorithm-level methods to deal with the imbalance problem, such as Large Margin Local Embedding (LMLE) [47], Deep Over-Sampling (DOS) [4] and Class Rectification Loss (CRL) and hard sample mining [26].

LMLE method combines a new sampling method (quintuple sampling) and a new novel loss function (triple-header hinge loss) to learn deep feature representations. Quintuple sampling selects one sample and four more samples based on the intra-class and the inter-class distances. These five samples are send to five identical CNNs and triple-header hinge loss is used to compute error and update the parameters. Although LMLE was proved efficient on learning deep representation of class-imbalanced image datasets, it was of highly computational cost and very complex to be implemented.

Deep Over-Sampling (DOS) amplifies the difference between majority class and minority class in the deep feature space by selecting more minority samples and uses the micro-cluster loss to strengthen the inter-class distinction.

The combination of hard sample mining and CRL was proved efficient on large-scale highly imbalanced image datasets. For each mini-batch, the hard sample mining selects more informative minority samples to help the model learning faster with fewer images. CRL can reduce bias towards the majority classes caused by the over-representation. The proposed method was demonstrated more effective than many class imbalance methods on large-scale image datasets.

To sum up, the hybrid methods are more complex and of higher computing cost than data-level and algorithm-level methods. As the hybrid methods combine data-level and algorithm-level method, it is expected that their flexibility will be decrease.

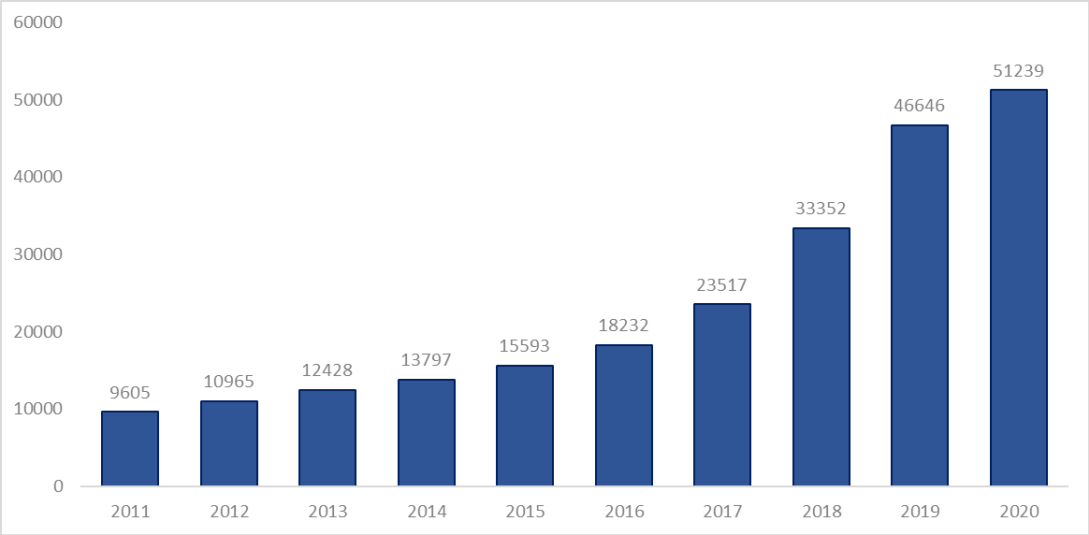

Figure 2.2: Publication Number about Artificial Intelligence (machine learning, deep learning, intelligent, AI) and medical datasets (medical, clinical, healthcare) on Web of Science from 2011 to 2020

2.5 Artificial Intelligence Applications on Medical Datasets

Artificial Intelligence (AI) have achieved lots of success applications with the increasing availability of medical datasets and tremendous increasing computation power. This trend is also reflected by the increasing number of publications about this topic on the literature database Web of Science. We searched publications of last decade with using the following keywords: AI related words (machine learning, deep learning, intelligent and AI) and medical related words (medical, clinical and healthcare). The results is shown in as the Fig. 2.2. In 2011, 9605 articles have been published about the topic of AI application on medical datasets, whereas 51,239 articles were published in 2020. This is a 5-fold increase, which indicates a huge growth of articles during the past decade(see figure 2.2).

The growing application in this topic has multiple reasons. One reason is the increasing implementation of electronic health records (EHRs) by hospitals. According to [44], by 2014 already 75.5% of the hospitals in the US have implemented EHRs successfully and over 95% possessed the technologies to process EHRs . These EHRs are becoming a standard in healthcare and can be used for disease detection [42]. On the other hand, AI has grown rapidly during the last two decades. For example, image recognition, a common technique to detect diseases, has achieved great success. Therefore, applying AI in healthcare seems to be convincing.

AI is a collection of different technologies and algorithms, such as rule-based expert systems, machine learning, deep learning and physical robots. Here is a table of their brief definitions and typical applications, as shown in Table 2.3. Since 1970s, rule-based expert systems has achieved many successes on medical datasets, such as diagnose diseases, clinical reasoning, treatment

suggesting and physician assistance. Nevertheless, rule-based expert systems are difficult to build up with so much decision rules which require human experts updates. Additionally, it is also highly complex to merge different pieces of information from different experts.

Other than relaying on the human-expert knowledge and the decision rules, which is used in the rule-based expert systems, recent AI studies mainly leverage machine learning and deep learning methods to get better performance in healthcare tasks.

Table 2.3: The main categories of AI and their definitions, applications in healthcare area

Method	Brief Definition	Typical Application
Rule-based expert systems	A computer system which emulates the process of human decision-making	Clinical decision support systems
Machine learning	A computational algorithm that could be improved automatically with fitting data	Precision Medicine, i.e., predicting which treatment are useful based on the treatments and patient attributes
Deep learning	The complex forms of machine learning, composed by several levels of neural networks	Medical image analysis, e.g., detecting cancerous lesions in radiology images
Physical robots	Robots perform pre-defined tasks, such as lifting, assembling, repositioning and delivering objects	Surgical robots could be used to improve the ability of surgeons, such as vision, precise incisions

Machine learning applications in healthcare can be grouped into two classes: supervised learning and unsupervised learning. Supervised learning methods mainly predict the output with training a large number of 'training' samples with their labels. Through minimizing the deviations between the ground truth label and the prediction results, supervised methods could approach the optimal parameters to get a generalized model for the new cases, which could be evaluated by the test set. The most widely applications of supervised learning methods are classification and regression tasks. Whereas the unsupervised learning methods are mainly used to find the potential clusters and outlier detection. The most common machine learning applications in healthcare are supervised learning tasks, such as precision medicine and clinical outcome prediction.

Deep learning has achieved great success in voice recognition and image classification tasks [66]. Deep learning applications in healthcare further promote the recent renaissance in AI. Different kinds of Deep Neural Network (DNN) have different application scenarios. For instance, autoencoders are mainly used to reduce dimensions, whereas Recurrent Neural Network (RNN) are mainly used time-sires datasets.

The most successful AI application in healthcare domain is probably the automatic image-based diagnosis, which is crucial to modern medicine. Image-based diagnosis can provide an objective assessment, which are useful to help the doctor achieve a better assessment. Image-based diagnosis have yielded promising results on tasks, such as, interstitial lung

diseases classification based on Computed Tomography (CT) images [5], the breast cancer classification [48] and the skin cancer classification[108].

There are several challenges when implementing AI on medical datasets. Firstly, medical datasets from different healthcare providers usually contain different kinds of bias and noise, which might lead to the poor generalization of the model trained on one source [84]. One possible solution is applying consensus diagnoses to improve the performance and the generalization of the machine learning models [59]. Another idea is enhance the reliability of machine learning models by addressing the idiosyncrasies and noises of various healthcare providers. Additionally, although machine learning models can achieve higher performance in some healthcare tasks, it is challenging to interpret and explain the models. It is hard to extract biological insights from these "black boxes" [49]. Another challenge is the implementation of a computing environment for data curation, data collection, and data sharing. Privacy-preserving approaches are helpful to secure the data communication [83]. Standard representation of diagnosis is also required for communications across healthcare providers [25]. With the development of AI applications on medical datasets, they will lead to new social, economic and legal challenges [23]. For instance, AI applications on medical datasets will inevitably result in legal challenges regarding medical negligence attributed to complex decision support systems. When malpractice cases involving medical AI applications arise, the legal system will need to provide clear guidance on what entity holds the liability[49]. To solve the mentioned challenges, scientists in both AI and healthcare areas should work together develop the applications that deal with crucial needs step by step. This dissertation will focus on dealing with medical classification tasks under class imbalance, which is introduced in Section 1.1.

Chapter 3

Proposed Imbalance Learning Methods

In this chapter, we firstly introduce the framework of three proposed imbalance learning methods, two machine learning methods under class imbalance and one deep learning method under class imbalance. Then each proposed method are described in detail. The sampling-based method is a combination of the sample synthetic method SMOTE and the a strong tree-based classifier XGBoost, and the detail is put in Section 3.2.1. From the insight of the previous sampling-based method, we propose a novel ensemble learning method, named as multiple balanced subset stacking in Section 3.2.2. To improve the performance of deep neural networks on imbalanced image dataset, we propose a novel loss function which could dynamically tune the class weight during the training process and details are set in Section 3.3. Lastly, we briefly conclude the contribution of the proposed imbalance learning methods.

Contents

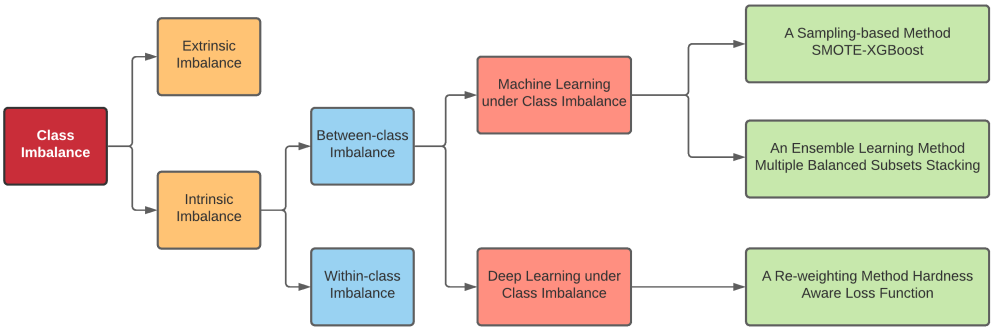

Figure 3.1: The framework of proposed imbalance learning methods

3.1 Framework of Proposed Imbalance Learning Methods

The dissertation aims to fix the class imbalance problem, which is a big challenge in the data mining area. As introduced in Section 2.1, class imbalance is called intrinsic imbalance if the dataset is inherently imbalanced, such as Cancer diagnosis. On the other hand, class imbalance is named as extrinsic imbalance if the dataset is skewed distributed due to extrinsic causes. The extrinsic imbalance could be solved by fixing extrinsic causes accordingly. In this thesis, we focus on the intrinsic imbalance by evaluating the proposed methods on medical datasets, which are class imbalanced in nature.

The class imbalance usually refers to the between-class imbalance, where a dataset is composed by the majority class and the minority class [38]. The minority class is severely under-represented by less samples compared to the majority class. The misclassification cost of a minority sample is usually much larger than that of a majority sample, as explained by the cancer diagnose in Section 1.1. When the imbalance happens within a class, we call it within-class imbalance. A class includes several sub-clusters while some sub-clusters have more sample than others. As declared in Section 2.1, the class imbalance refers to the between-class imbalance in this dissertation.

A variety of imbalance learning methods have been proposed deal with the class imbalance problem of different datasets. We mainly focus on two scenarios, machine learning for imbalanced structured dataset and deep learning for imbalanced image dataset. We propose two machine learning methods for imbalanced structured dataset and a deep learning method for imbalanced image dataset. The framework of our proposed methods is shown as Fig. 3.1.

Many machine learning approaches have been proposed to deal with the imbalanced structured dataset. These approaches be categorized into three groups, i.e., data-level approaches, algorithm-level approaches and hybrid approaches. Data-level approaches try to rebalance the class distribution of the imbalanced dataset. On the other hand, algorithm-level approaches are developed to fix the imbalance problem by assigning higher misclassification cost to the

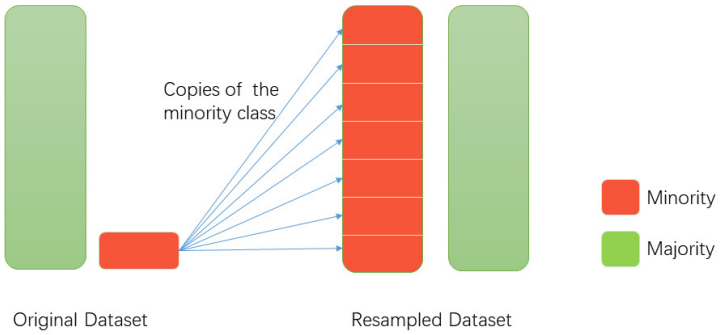

Figure 3.2: The oversampling process on the binary class dataset

minority samples. Whereas hybrid approaches combine data-level approaches and algorithm-level approaches together. We propose one sampling-based method for imbalanced structured datasets in Section 3.2.1. This sampling-based method first use the sample sample synthetic method SMOTE to generate the minority samples, and send the resampled dataset to a strong classifier XGBoost. Experimental results show that the sampling-based method outperforms the classic machine learning methods, which demonstrate the effectiveness of data-level imbalance learning approach on the imbalanced structured dataset.

However, as described in Section 2.3, existing approaches have different kinds of problems. For instance, data-level methods usually encounter the problem of over-fitting or the problem of information loss. Oversampling and undersampling are two representative strategies of data-level methods. Taking an imbalanced dataset with two classes as example, we show the process of oversampling in Fig. 3.2. The minority samples are duplicated to increase the size of the minority class to the same size of the majority class. However, when a classifier is trained on the resampled dataset, one minority sample will be learned multiple times, which makes the model overfit to the minority class. Similarly, the undersampling process on the binary class dataset is shown in Fig 3.3. After the undersampling, the size of the majority class is decreased to the same size of the minority class by discarding majority samples, which will loss lot of valuable information. Meanwhile, it is hard to tune the cost matrix of the cost-sensitive learning method, which is usually task specified and is hard to generalized to other tasks.

To avoid the problems of existing machine learning methods under class imbalance, we propose a ensemble learning method, i.e., Multiple bAlance Subsets Stacking (MASS). MASS first cuts the majority class into multiple subsets by the size of the minority set, and combines each majority subset with the minority set as one balanced subset. In this way, MASS take advantage of each sample, thus could overcome the problem of information loss. The balanced

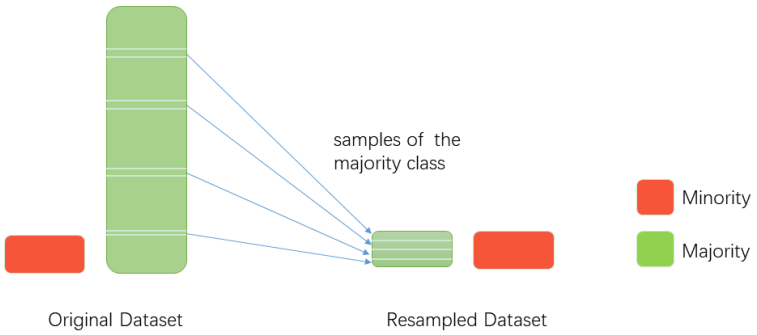

Figure 3.3: The undersampling process on the binary class dataset

subset is sent to train one base classifier of MASS. Then, we can get several trained base classifiers. The original dataset is feed to all the trained base classifiers, whose output are used to generate the stacking dataset. The stacking dataset is sent to train a stacking model, which optimizes the weights of base classifiers. As the stacking dataset keeps the same label as the original dataset, the stacking ensemble process does not encounter any problem of data level methods. Finally, we can get an ensembled strong model based on the trained base classifiers and the staking model. Extensive experimental results on three medical datasets show that MASS outperforms baseline methods. The robustness of MASS is proved over implementing different base classifiers. We design a parallel version MASS to reduce the training time cost. The speedup analysis proves that Parallel MASS could reduce training time cost greatly when applied on large datasets.

When it comes to deal with image datasets, class imbalance problem will decrease the prediction performance of deep learning methods, which achieve great success in computer vision applications. Similar to the machine learning under class imbalance, solutions for imbalanced image classification could be grouped into re-sampling methods or re-weighting methods. In the context of computer vision applications, over-sampling methods introduce large training costs and make the model and under-sampling methods may discard important samples that are valuable for deep representation learning. Taking these issues of applying re-sampling methods on image classification tasks into consideration, we focuses on designing a better re-weighting method to improve the prediction performance of the deep neural networks. Existing re-weighting methods usually assign the class weight inversely proportional to the class frequency respectively, which might lead to poor performance as proved in Section 6.1. After introducing a new definition of classification hardness in Section 3.3, we propose to use it to tune the class weight of loss function dynamically during the training process of deep neural networks and the novel loss function is named as Hardness Aware Dynamic (HAD).

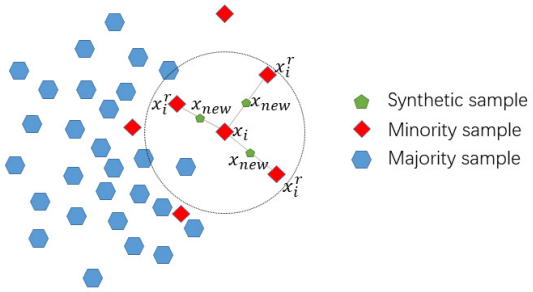

Figure 3.4: The sample synthetic process of SMOTE in the two-dimensional feature space when applied on binary class datasets

3.2 Proposed Machine Learning Methods under Class Imbalance

3.2.1 A Sampling-based Method SMOTE-XGBoost

The Synthetic Minority Over-sampling TEchnique (SMOTE)[14] is an oversampling method to rebalance the original training dataset. The main idea of SMOTE is to generate artificial examples to expand the scope of minority classes. SMOTE generates new minority data points based on the similarities between minority samples from original dataset. The process of sample synthetic process in two dimensional space is shown in Fig. 3.4. Firstly, k-nearest neighbours of a minority sample x_i are identified with the smallest Euclidean distance (k=3 in this figure). Then, one of the nearest neighbours is selected randomly, denoted by x_i^r. Then the new data is defined as formula 3.1, where δ is randomly chosen in range [0,1]. The new data is a point along the line between x_i and x_i^r. By doing so for several rounds, we could obtain a balanced training set. SMOTE is widely used on lots of applications with the class imbalance problem, such as breast cancer detecting[29], network intrusion detecting[19], and histopathology annotation[27], which proved its effectiveness.

$$x_{new} = x_i + \delta(x_i^r - x_i) \tag{3.1}$$

XGBoost is a powerful classification model that assembles weak prediction models (e.g., decision tree) to build a strong prediction model[17]. XGBoost has been proved very successful in lots of applications, such as web text classification, store scale prediction, motion detection, ad click through rate prediction and so on. There are several advantages of XGBoost compared

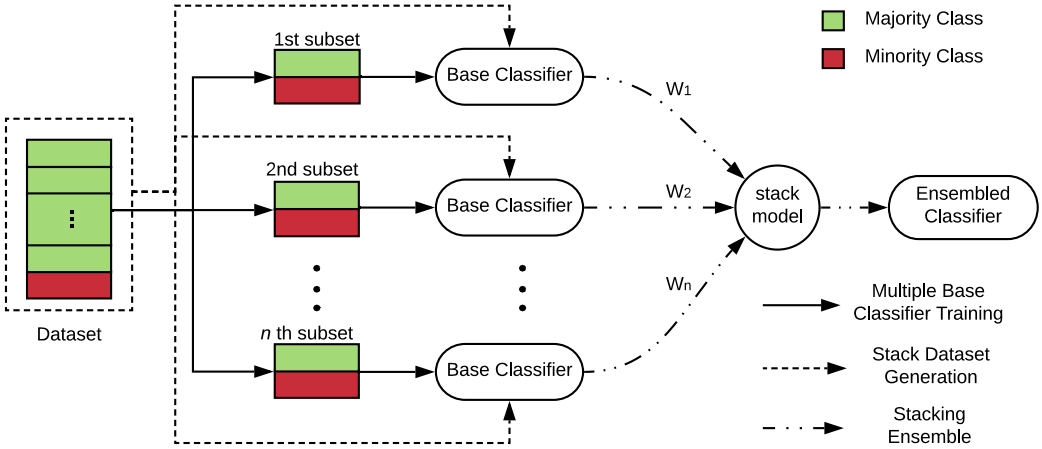

Figure 3.5: Multiple Balance Subsets Stacking Ensemble Process

with other classifiers. Firstly, XGBoost is more flexible and users could define optimization objectives and evaluation metrics. Secondly, XGBoost could overcome over-fitting problem with regularization.

To deal with the imbalance distribution of the imbalanced dataset, we propose to combine SMOTE and XGBoost, which is named as SMOTE-XGBoost. The dataset is randomly split into training set and testing set. Afterwards, we use SMOTE to generate the minority samples to balance the training set. Then the resampled training set is sent to train the XGBoost classifier.

In Chapter 4, we conduct experiments to evaluate the performance of the sampling-based method SMOTE-XGBoost on a medical dataset.

3.2.2 An Ensemble Learning Method Multiple Balance Subsets Stacking

In this section, we describe the novel ensemble learning method called Multiple Balance Subsets Stacking (MASS) in detail. We first define some symbols and show the process of MASS, then we describe the details of each step, and summarize MASS in the pseudo code. Finally, in order to reduce the training time cost and improve efficiency, we optimize MASS with using parallel computing.

From the insight of multiset feature learning [125] and stacking ensemble [124], we propose a novel ensemble method MASS to deal with the imbalance problem, and its construction process is illustrated in Fig. 3.5.

For a given imbalanced dataset $\mathcal{D} = \{(X_i, y_i) : X_i \in \mathbb{R}^N, y_i \in \{0, 1\}\}$, where N is the number of the features. $\mathcal{T}$ is used to denote the minority set, where $y_i = 1$; we use $\mathcal{F}$ to denote majority set, where $y_i = 0$. They are defined as follows:

$$\mathcal{T} = \{(X_i, y_i) : X_i \in \mathbb{R}^N, y_i = 1\} \tag{3.2}$$

$$\mathcal{F} = \{(X_i, y_i) : X_i \in \mathbb{R}^N, y_i = 0\} \tag{3.3}$$

In Fig. 3.5, majority set is presented by green block, and minority is represented by red block.

Our method can be broken into three stages:
1. Generate multiple balance subsets from the original dataset with using multiple balance subsets constructing strategy and send each balance subset to train one base classifier.
2. Every sample from the original dataset is sent to all base classifiers, and the prediction results are collected into a new dataset, which is denoted as stacking dataset.
3. The stacking dataset is used to train a stack model to optimize the weights of base classifiers and get the final strong ensemble classifier.

Multiple Balance Subsets Constructing Strategy

The process of multiple balance subsets constructing method is illustrated as first stage by solid line in Fig. 3.5. Firstly, the majority set $\mathcal{F}$ is randomly partitioned into several majority subsets $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$, and each majority subset $\mathcal{F}_i$ has the same number of samples as the number of samples in the minority set $\mathcal{T}$. If there are some majority samples left after generating multiple majority subsets, we will randomly duplicate adequate samples from previous generated majority subsets to construct another majority subset. In total, the number of majority subsets is n, where $n = \lceil IR \rceil$. Imbalance Ratio (IR) is defined as number of the samples in majority set $\mathcal{F}$ divided by the number of samples in minority set $\mathcal{T}$: $IR = |\mathcal{F}|/|\mathcal{T}|$

Secondly, we combine each majority subset with the minority set into a balance training subset $\mathcal{S}_i$. Thus we will have $\lceil IR \rceil$ balance sets: $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$. These subsets include all samples from the original training dataset $\mathcal{D}$, so there is no information loss in MASS.

Thirdly, balance training subset $\mathcal{S}_i$ will be sent to train base classifier f_i, and we will get $\lceil IR \rceil$ base classifiers. These base classifiers will be used in the next stage to generate staking dataset.

Stacking Dataset Generation

Each base classifier is trained by one of the balance subsets $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$. Thereby, each trained base classifier only captures part information of majority set. In order to get complete information of majority samples, an intuitive way is to integrate these base classifiers as a strong ensemble classifier. From the insight of stacking ensemble, we will first generate a stack

dataset based on the base classifiers, and then use it to get the optimized weights of these base classifiers to get a strong classifier.

As shown in Fig. 3.5, each sample X_j from the training set $\mathcal{D}$ is sent to all base classifiers, which would generate n prediction scores $\{y\_pre_{j1}, y\_pre_{j2}, ..., y\_pre_{jn}\}$. Each prediction score presents the possibility of the sample classified as minority sample. These scores are combined with ground truth$\{y_j\}$ to construct a new dataset, which is denoted as stacking dataset $\mathcal{D}'$:

$$\mathcal{D}' = \begin{pmatrix} y\_pre_{1,1} & y\_pre_{1,2} & \cdots & y\_pre_{1,n} & y_1 \\ y\_pre_{2,1} & y\_pre_{2,2} & \cdots & y\_pre_{2,n} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ y\_pre_{j,1} & y\_pre_{j,2} & \cdots & y\_pre_{j,n} & y_j \\ \vdots & \vdots & & \vdots & \vdots \\ y\_pre_{m,1} & y\_pre_{m,2} & \cdots & y\_pre_{m,n} & y_m \end{pmatrix} \tag{3.4}$$

In other words, the prediction scores are treat as new features of the original sample X_j. The new features are used to train the stack model in the next stage.

Stacking dataset $\mathcal{D}'$ have same labels as the original dataset $\mathcal{D}$, which means MASS does not increase the number of minority samples. Hence, there is no overfitting problem with applying MASS on imbalanced dataset.

Algorithm 1: Multiple Balance Subsets Stacking

Input Dataset $\mathcal{D} = \{(X_1, y_1), (X_2, y_2), ..., (X_m, y_m)\}$, base classifier f, stack model s
Initialization:
Initialize n base classifiers $f_1, f_2, ..., f_n$, $n = \lceil IR \rceil$
Initialize one stack model s.
Stage 1: Multiple Balance Subsets Construction
Generate n balance training subsets $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$ from training set $\mathcal{D}$ with using Multiple balance subsets construction strategy.
for $i = 1$ *to* n **do**
    Train base classifier f_i with using subset $\mathcal{S}_i$
end
Stage 2: Stacking Dataset Generation
$\mathcal{D}' = \Phi$
for $j = 1$ *to* m **do**
    for $i = 1$ *to* n **do**
        $y\_pre_{ji} = f_i(X_j)$
    end
    $\mathcal{D}' = \mathcal{D}' \cup \{((y\_pre_{j1}, y\_pre_{j2}, ..., y\_pre_{jn}), y_j)\}$
end
Stage 3: Stacking Ensemble
Train stack model s with using stacking dataset $\mathcal{D}'$ to optimize the weights of all base classifiers.
return Stacking ensemble classifier $S(X) = s(f_1(X), f_2(X), ..., f_n(X))$

Stacking Ensemble

The stack model intends to find the optimal weights to aggregate all base classifiers in a way that the final ensemble classifier minimize the classification cost. As our task is a binary classification problem, we use logistic regression as our stack model. The prediction results is defined as:

$$\hat{y} = \boldsymbol{\delta}[\sum_{i=1}^{n}(w_i y\_pre_i) + b], \quad \boldsymbol{\delta}(z) = \frac{1}{1+e^{-z}} \tag{3.5}$$

Cross-entropy loss function is commonly used for binary classification:

$$L_{CE}(\hat{y}, y) = -[y \log \hat{y} + (1-y) \log(1-\hat{y})] \tag{3.6}$$

Therefor, we can get the optimal weights by minimize cost function with using gradient descent method:

$$J(W) = \frac{1}{m}\sum_{j=1}^{m} L_{CE}(\hat{y}_j, y_j) \tag{3.7}$$

$$= \frac{1}{m}\sum_{j=1}^{m} L_{CE}\{\delta[\sum_{i=1}^{n}(w_i y\_pre_{ji}) + b], y_j\} \tag{3.8}$$

Finally, the details of MASS are shown in Algorithm 19.

Parallel MASS

In order to obtain a usable prediction model faster, we need to improve the training speed of the model. From the insight of parallel machine learning frameworks [57] and parallel machine learning on distributed data-parallel platforms [35], we will optimize MASS by running it in parallel. As the balance training subsets are independent with each other, we can train all base classifiers in parallel to get n independent base classifiers. Similarly, the loop in stacking dataset generation stage could also be run in parallel. Therefor, we could optimize these steps with parallelism, and the Parallel MASS is shown in Algorithm 18.

Algorithm 2: Parallel Multiple Balance Subsets Stacking

Input Dataset $\mathcal{D} = \{(X_1, y_1), (X_2, y_2), ..., (X_m, y_m)\}$, base classifier f, stack model s
Initialization:
Initialize n base classifiers $f_1, f_2, ..., f_n$, $n = \lceil IR \rceil$
Initialize one stack model s.
Stage 1: Multiple Balance Subsets Construction
Generate n balance training subsets $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$ from training set $\mathcal{D}$ with using Multiple balance subsets construction strategy.
for $i = 1$ *to* n **do**
    Train base classifier f_i with using subset $\mathcal{S}_i$
end
Stage 2: Generating Stacking Dataset in Parallel
$\mathcal{D}' = \Phi$
foreach $i \in \{1, 2, ..., n\}$ ***Parallel*** **do**
    Train base classifier f_i with using subset S_i on process i
    $y\_pre_i = f_i(X)$
    $\mathcal{D}' = \mathcal{D}' \cup y\_pre_i$
end
Stage 3: Stacking Ensemble
Train stack model s with using stacking dataset $\mathcal{D}'$ to optimize the weights of all base classifiers.
return Stacking ensemble classifier $S(X) = s(f_1(X), f_2(X), ..., f_n(X))$

When running MASS without parallelism, the training time includes five part: the time of initialization T_{init}, the time of balance subsets generation T_{gen}, the time of training base classifiers T_{base}, the time of stacking dataset generation T_{data}, the time of training stack model T_{stack}. As a result, the total training time of MASS T_M is:

$$T_M = T_{init} + T_{gen} + T_{base} + T_{data} + T_{stack} \tag{3.9}$$

Parallel MASS still has the same time cost of initialization, balance dataset generation and stack model training as MASS. Meanwhile, n base classifiers will be run on n processes, and the stacking dataset generation is also merged into the respective processes (step 6 in Algorithm 18). Specially, running MASS in parallel has another time cost, which is communication time between processes T_{com}. Thus the total training time of Parallel MASS T_P is:

$$T_P = T_{init} + T_{gen} + \frac{T_{base}}{n} + \frac{T_{data}}{n} + T_{stack} + T_{com} \tag{3.10}$$

Obviously, the cost of initialization and balance subset generation could be neglect compared with that of training base classifiers. The speedup of running MASS in parallel is T_M (the training time of MASS) divided by T_P (the training time of Parallel MASS). Consequently, the theoretical speedup is:

$$Speedup = \frac{T_M}{T_P} \tag{3.11}$$

$$= \frac{n * (T_{init} + T_{gen} + T_{base} + T_{data} + T_{stack})}{n * (T_{init} + T_{gen} + T_{stack} + T_{com}) + T_{base} + T_{data}} \tag{3.12}$$

$$\approx \frac{n * (T_{base} + T_{data} + T_{stack})}{T_{base} + T_{data} + n * (T_{stack} + T_{com})} \tag{3.13}$$

3.3 Proposed Deep learning Method under Class Imbalance

In this section, we will first introduce a new definition: classification hardness. Then we propose to use the classification hardness of each class to tune the loss function, which is defined as Hardness Aware Dynamic (HAD) loss. At last, we summarize the HAD process in Algorithm 12.

3.3.1 A Re-weighting Method Hardness Aware Dynamic Loss Function

The concept of 'classification hardness' is introduced here to represent the hardness of correctly classifying a sample for a trained classifier. For a given dataset $\mathcal{D} = (x_j, y_j)$, $x_j \in \mathbb{R}^N$ are instances and $y_j \in \{1, 2, ..., C\}$ are ground truth class labels, where C is the number of classes.

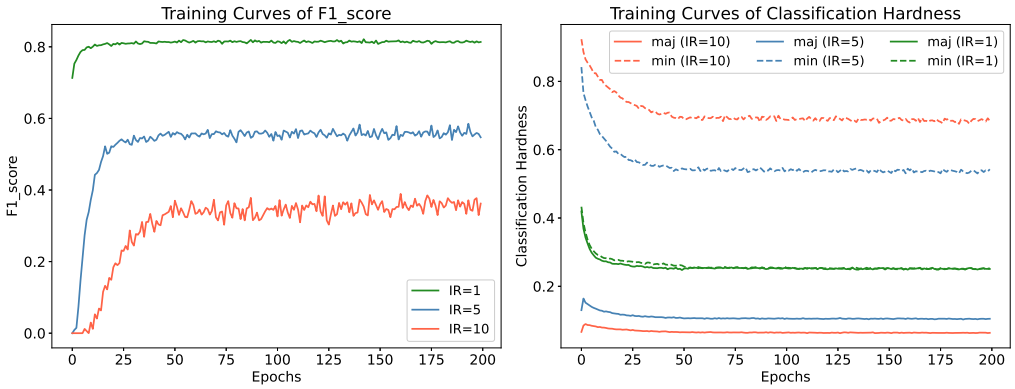

Figure 3.6: Training curves of F1-score and Classification hardness of ResNet-18 on three Breast Cancer subsets with different imbalance ratios(IR), IR=1, IR=5, IR=10

Definition. (Classification Hardness): For any tuple (x, y) in $\mathcal{D}$, the instance x correctly classified probability by classifier f is $p = f(x)$, then the classification hardness of instance x is defined as $1 - p$, which is denoted as $h(x) = 1 - f(x)$.

Accordingly, for the i th class, we have a subset $\mathcal{D}_i = \{(x, y); y = i\}$. The instance number of $\mathcal{D}_i$ is n_i. The hardness of i th class is denoted as H_i:

$$H_i = \frac{\sum_{i=1}^{n_i} h(x)}{n_i} = \frac{\sum_{i=1}^{n_i} (1 - f(x))}{n_i}; y = i \tag{3.14}$$

Classification hardness H_i is the difficulty degree to classify instances belong to i th class. For classifier $f(x)$ trained by dataset $\mathcal{D}$, we can calculate its classification hardness value of each class and put them in a vector $\mathcal{H} = [H_1, H_2, ...H_C]$, which is denoted as the classification hardness vector.

To better explain the classification hardness, for simplicity, we consider the training process of a deep neural network (ResNet-18 [41]) on a binary image classification task, i.e., breast cancer classification, and its detail is in section 6.2.1. Under this situation, we denote patients diagnosed without Invasive Ductal Carcinoma (IDC) as majority samples, and patients diagnosed with IDC is denoted as minority samples. The Imbalance Ratio (IR) is the number of majority samples divided by the number of minority samples. We construct three subsets of breast cancer dataset under different imbalance degrees ($IR \in \{1, 5, 10\}$).

We train ResNet-18 on these three subsets 200 epochs individually and get the training curves of F1-score and classification hardness, which are shown in Fig 3.6. In the left figure, the green line represents the training curve of F1-score on training sets with IR=1, the red line for IR=5, and the blue line for IR=10 respectively. As the left figure shows: (1) When the dataset is balanced (IR=1), F1-score soon arrives nearly 0.8 after a few epochs; (2) the model achieves lower F1-scores when trained with imbalanced subsets compared with the balanced dataset; (3) the larger IR the dataset has, the lower F1-score achieves by the model at the end of training process.

Classification hardness of majority class is presented by the solid lines, minority class by the dashed lines in the right part of Fig. 3.6. We can indicate from the right figure that: (1) when IR=1, the classification hardness of majority (green solid line) is very close to that of minority (green dashed line), which means the model has no bias in predictions while trained with the balanced subset. (2) when IR=5, the classification hardness of majority (blue solid line) is much smaller than that of minority (blue dashed line), which indicates that the trained model is biased towards the majority class. (3) The red and blue dashed lines are always above the green dashed line, which means the classification hardness of majority class under imbalance situations is larger than that under balance situations. It proves that it is harder for the model to classify the minority samples under imbalance situations. (4)The red dashed line (IR=10) is higher than the blue dashed line (IR=5), which indicates that the higher is the imbalance ratio of the training set, the harder is to classify the minority samples for the model.

According to the observations, we propose to use classification hardness to tune the weight dynamically during the training process of DNNs, and we will detail these in the following subsection.

Hardness Aware Weights

Intuitively, to get better classification results, we want to assign higher weights to classes that are harder to be classified during the training process of DNNs.

After each training epoch of a DNN, we could calculate the classification hardness vector $\mathcal{H}$ of the training set at the end of each iteration. Following that, we could calculate the hardness aware dynamic weight $\mathcal{W}_d$. The hardness aware dynamic weight vector of t th iteration is defined as the collection of all the weights of each class $\mathcal{W}_d^t = [w_1^t, w_2^t, ..., w_C^t]$.

To avoid the large fluctuation of the class weights, the hardness aware dynamic weight will be updated by the following equation:

$$\mathcal{W}^{t+1} = (1 + \lambda \times \mathcal{H}_t)\mathcal{W}_d^t; \lambda \in [0, 1] \tag{3.15}$$

$\mathcal{H}_t$ is the classification hardness vector computed at the end of the t th iteration. $\mathcal{W}_d^t$ is the dynamic weight vector used in the training process of the t th epoch, and initial dynamic weight $\mathcal{W}_d^0$ is set as $[1, 1, ..., 1]$ in default. $\mathcal{W}_d^{t+1}$ is the dynamic weight vector that will be used in the training process of the $(t + 1)$ th epoch. λ is a super parameter to control the updating speed of hardness aware weight vector, and it is set to 0.01 by default. Especially, when $\lambda = 0$, the dynamic weight vector will not be updated during the training process. Each weight factor w_i^{t+1} of $\mathcal{W}_d^{t+1}$ is normalized as $\sum_{i=1}^{C} w_i^{t+1} = C$ to keep the total loss roughly same scale with the loss without considering re-weighting.

Hardness Aware Dynamic Loss

The Hardness Aware Dynamic (HAD) Loss is proposed to address the imbalance problem by the dynamic weight vector. For an instance x with a ground truth label $y \in \{1, 2, ..., C\}$, suppose we have a classification model and its prediction probability of instance x are $\mathbf{p} = [p_1, p_2, ..., p_C]$, where $\forall p_i in [0, 1]$ and $\sum_{i=1}^{C} p_i = 1$, The loss of x is denoted as $L(\mathbf{p}, y)$. To merge the hardness aware weight into the loss of instance x, we will assign a weight factor $w_d^t(y)$ to the loss function, where $w_d^t(y)$ is the weight factor of the y th class. The Hardness Aware Dynamic (HAD) loss is defined as:

$$\mathrm{HAD}(\mathbf{p}, y) = w_d^t(y) L(\mathbf{p}, y) \tag{3.16}$$

Dynamic weight vector $\mathcal{W}_d$ depends on the validation process of the classification model after each epoch. Once the classifier and the initial weight is set up, the dynamic weight vector could be used to optimize the classifier, thus it can be used by any deep neural network model and any loss function. To prove the dynamic loss is generic, we apply the dynamic weight vector to three loss functions: sigmoid cross-entropy loss, softmax cross-entropy loss, and the recently proposed focal loss[71].

Hardness Aware Dynamic Softmax Cross-Entropy Loss.

For a given tuple $(x, y) \in \mathcal{D}$, suppose prediction score $\mathcal{S} = [s_1, s_2, ..., s_C]$ by a given classification model f is $\mathcal{S} = f(x)$. The softmax function is used to squash $\mathcal{S}$ into the prediction probability vector $\mathbf{p} = [p_1, p_2, ..., p_C]$, where the prediction probability of x as j th class is p_j:

$$p_j = \frac{\exp(s_j)}{\sum_{i=1}^{C} \exp(s_i)} \tag{3.17}$$

As the ground truth label of instance x is $y \in \{1, 2, ..., C\}$, thus the correctly classification probability is p_y. Accordingly, the softmax cross-entropy (CE) loss for x can be calculated by following equation:

$$\mathrm{CE}_{\text{softmax}}(\mathcal{S}, y) = -\log(\frac{\exp(s_y)}{\sum_{i=1}^{C} \exp(s_i)}) \tag{3.18}$$

The dynamic weight factor of x in the t iteration is $w_d^t(y)$, then its Hardness Aware Dynamic (HAD) softmax cross-entropy loss is:

$$\mathrm{HAD}_{\text{softmax}}(\mathcal{S}, y) = -w_d^t(y) \log(\frac{\exp(s_y)}{\sum_{i=1}^{C} \exp(s_i)}) \tag{3.19}$$

Hardness Aware Dynamic Sigmoid Cross-Entropy Loss.

When sigmoid function is used to calculate the prediction probabilities, it assumes different classes are independent of each other. Accordingly, the multi-class classification task is regarded as a multiple binary classification problem. For simplicity, we use the same notations as the previous part, and we define the prediction probability of x to be i th class p_i^b as:

$$p_i^b = \begin{cases} p_i, & \text{if } i = y \\ 1 - p_i, & \text{otherwise} \end{cases} \quad \text{where } p_i = \text{sigmoid}(s_i) \tag{3.20}$$

The sigmoid cross-entropy loss for x can be calculated by following equation:

$$\begin{aligned} \text{CE}_{\text{sigmoid}}\left(\mathcal{S}, y\right) &= -\sum_{i=1}^{C} \log(p_i^b) \\ &= -\sum_{i=1}^{C} \log\left(\frac{1}{1+\exp\left(-s_i^b\right)}\right) \end{aligned} \tag{3.21}$$

The Hardness Aware Dynamic (HAD) sigmoid cross-entropy loss for x is :

$$\text{HAD}_{\text{sigmoid}}\left(\mathcal{S}, y\right) = -w_d^t(y) \sum_{i=1}^{C} \log\left(\frac{1}{1+\exp\left(-s_i^b\right)}\right) \tag{3.22}$$

Hardness Aware Dynamic Focal Loss.

Focal loss is recently proposed to alleviate the imbalance problem in many computer vision tasks [71]. Focal loss focuses the training process of model on the difficult instances and prevent the model bias towards the well-classified instances. The focal loss is defined as:

$$\text{FL}(\mathcal{S}, y) = -\sum_{i=1}^{C} \left(1 - p_i^b\right)^{\gamma} \log\left(p_i^b\right) \tag{3.23}$$

The Hardness Aware Dynamic (HAD) focal loss is:

$$\text{HAD}_{\text{focal}}(\mathcal{S}, y) = -w_d^t \sum_{i=1}^{C} \left(1 - p_i^b\right)^{\gamma} \log\left(p_i^b\right) \tag{3.24}$$

Algorithm 12 presented the training process of DNNs using HAD loss.

Algorithm 3: Deep Learning with Dynamic Hardness Aware Loss

Input The initial weight $\mathcal{W}_d^0$, training dataset $\mathcal{D}_t = \{X_t, y_t\}$, the super parameter λ, the number of epochs N, the HAD loss HAD.
Output Deep neural network classifier f
Initialization Initialize a deep neural network classifiers f, initialize the Hardness Aware Weight $\mathcal{W}_d$= $\mathcal{W}_d^0$
for $i = 1$ *to* N **do**
    Transfer the dynamic weight $\mathcal{W}_d$ to the dynamic loss function HAD
    Use training dataset $\mathcal{D}_t$ to train classifier f with the goal of minimizing HAD
    Get prediction probability of the training dataset $\mathcal{D}_t$, i.e., **P** = $f(X_t)$
    Compute the classification hardness vector $\mathcal{H} = [H_1, H_2, ...H_C]$
    Update the dynamic weight $\mathcal{W}^{t+1} = (1 + \lambda \times \mathcal{H}_t)\mathcal{W}_d^t$
    Normalize the dynamic weight so that $\sum_{i=1}^{C} w_d^t(i) = C$
end
return Deep neural network classifier f

3.4 Summary

To deal with class imbalance problem of imbalanced structured dataset, we first propose a sampling-based method, and we implement it on one medical dataset in Chapter 4. The experimental results in Chapter 4 verify its effectiveness over classic machine learning methods.

After reviewing the existing machine learning methods on imbalanced structured datasets, we propose a novel ensemble learning approach called Multiple bAlance Subsets Stacking (MASS) to solve the imbalance problem via multiple balance subsets constructing strategy, and improve it to a parallel version to reduce the training time cost. MASS is implemented on three medical datasets in Chapter 5.

To tackle the imbalance problem on imbalanced image dataset, we introduce the definition of class-level classification hardness and use it in a novel loss function HAD loss, which could update class weight dynamically during the training process. In Chapter 6, we evaluate the proposed HAD loss on four different imbalanced image datasets.

Chapter 4

Evaluating SMOTE-XGBoost on A Medical Dataset

Postoperative complications worsen the quality of patients' life, even incurring prohibitively expensive costs. Accurate prediction of postoperative complication is highly important for clinical decision making and early treatment. However, distributions of postoperative complication of most diseases are imbalanced, which results in poor performance of conventional statistical logistic regression model and machine learning methods. Patients with large kidney stones are mostly treated by Percutaneous nephrolithotomy (PCNL). Previous studies on postoperative complication prediction of PCNL mainly focus on stone score systems and risk factors analysis. Stone score systems (e.g., S.T.O.N.E. nephrolithometry) only use kidney stone features to predict the complication, while ignoring lots of complication related features. In this chapter, based on a large clinic dataset spanning over 8 years, we first perform a thorough analysis of 3292 PCNL patients and compare the features between different classes. Then we identify different features that are statistically associated with complications. To solve the problem of imbalance distribution of the complications, we implement the sampling-based method SMOTE-XGBoost proposed in Section 3.2.1 on postoperative complications prediction of the PCNL dataset. SMOTE-XGBoost first uses SMOTE to balance the distribution of training dataset, then a XGBoost classifier is trained on the balanced dataset. Experiment results indicate by using kidney stone related features only, our approach outperforms the S.T.O.N.E. nephrolithometry and machine learning methods in both AUC and F1-score. Additionally, when other PCNL related feature sets are added into our model, the complication prediction performance could be improved further. Overall, SMOTE-XGBoost achieves an AUC of 0.7077 which is 41.54% higher than that of S.T.O.N.E. nephrolithometry. Our method can also be used to predict postoperative complication of other diseases.

Contents

4.1 Introduction

Every year there are over 300 million operations performed worldwide [122]. Operations pose considerable risk of postoperative complications, which could worsen the quality of patients' life, even incurring prohibitively expensive costs [54]. As half of the complications are preventable [53], accurate prediction of postoperative complications is highly important for clinical decision making, early treatment and counseling patient [109]. With the abundance of electronic health records, machine learning approaches have been applied to predict postoperative complications of different diseases, such as stroke[56] ,cancer[46], bleeding, shock, cardiac[73, 126], acute kidney injury and sepsis[106]. These studies mainly focus on feature selection[56, 106], feature sparseness(missing value)[128, 126]. Through postoperative complication distribution of most diseases are highly imbalanced, which would cause prediction models bias towards majority class and ignoring the minority class[38], few works have been done to deal with the imbalance problem in PCNL. Moreover, existing postoperative complication prediction models, such as multivariate logistic regression and machine learning classifiers, are usually optimized and evaluated using accuracy or error rate, which are not suitable for imbalanced datasets[117], thus limiting the performance of respective models. We take kidney stone disease as a study case. Kidney stone disease (also known as nephrolithiasis) is a worldwide public health problem. Studies report that the incidence of kidney stone disease is globally increasing in 5 European countries, Japan, China and the United States [92, 120, 94].

More and more patients with large kidney stones have been treated by PCNL since its introduction in 1976 [31]. PCNL becomes safer and more effective with the development of imaging and endourological instrumentation, but complications of the procedure, such as fever, bleeding, sepsis, adjacent organ injury are still common [113]. According to a global study of the PCNL [65], there were 1175 of 5724 (20.5%) patients experienced one or more complications after PCNL operation. It makes the postoperative complication prediction of PCNL a class-imbalance problem. Furthermore, in spite of class-imbalance problems, there are some limitations of previous works on postoperative complication prediction of PCNL.

One of the main limitations for the prediction is limited features used in exiting prediction models. There are three commonly used score systems, the Guy's stone score [105], the S.T.O.N.E. (stone size, tract length, obstruction, number of involved calices and essence) nephrolithometry [85], and CORES (clinical research office of the endourological society) nomogram [101], that are used as predictors of stone-free status and postoperative complication of PCNL. While the score systems are designed for stone-free status prediction, using kidney stone related features is enough to build a prediction model. However, when such systems are used to predict postoperative complications, the ignorance of other complication related features will degrade the prediction performance [112, 64]. A systematic review and meta-analysis of three score systems conclude that they are equally accurate and feasible for predicting stone-free

status after PCNL, however, the results of predicting postoperative complication of PCNL are controversial[50].

Another limitation of previous works in PCNL is that although risk factors of the complication are identified by univariate or multivariate analysis with using statistical logistic regression[28, 93, 102, 58, 2, 86, 65], no model has been built to predict the postoperative complications of PCNL based on these risk factors.

To address these limitations and the class-imbalance problem, we first perform a detailed analysis of postoperative complication of PCNL, then implement a sampling-based approach SMOTE-XGBoost to predict postoperative complications of PCNL. Patients' demographic characteristics, disease history, laboratory test variables, preoperative variables, kidney stone features and operation outcomes are compared according to complication status. Through the analysis, we identify variables statistically associated with the complications. To overcome the class-imbalance problem, SMOTE-XGBoost uses Synthetic Minority Oversampling TEchnique (SMOTE) [14] to balance the dataset and then uses eXtreme Gradient Boost (XGBoost) [17] to predict the postoperative complications. Additionally, instead of using accuracy or error rate as evaluation metrics, we use *area under the curve* (AUC) (also called c-statistic) and *F1-score* to evaluate our prediction model. To the best of our knowledge, this is the first work focusing on the postoperative complication prediction of PCNL with considering the data imbalance problem. We evaluate our developed system on a large collection of 3292 PCNL patients' records spanning from January 2012 to July 2019. Experimental results indicate that by using kidney stone related features only, our system outperforms baseline methods without considering the imbalance problem and state-of-the-art imbalance learning methods significantly. In addition, with considering four new kinds of PCNL related features, the complication prediction performance of our system achieves further enhancement. We believe that our method will also bring insight to improve the other medical tasks, e.g., heartbeat classification [40].

4.2 PCNL Dataset and Background of PCNL Complication

In this section, we first introduce the data preprocessing methods and statistical analysis methods in this work. The complication classification system and one of the traditional prediction models are introduced in detail.

4.2.1 Statistical Analysis

Ethics committee statement. This study complied with the Declaration of Helsinki and was approved from the ethics committee of First Affiliated Hospital of Gannan Medical University. Single center electronic data was collected consecutively from data warehouse system of the

hospital between January 2012 and July 2019. Protected information was excluded from the dataset.

We extracted 39 features from the unstructured clinical notes, including patients' demographic characteristics, disease history , laboratory test variables, kidney stone related features, and operation outcomes. The extracted features are gender, age, Body Mass Index (BMI), disease history (hypertension, diabetes, heart,liver, brain vessel, lung), laboratory test variables (scr, bun, alt, ast, protein, etc.), kidney anomalies, previous kidney surgical procedures, stone size, stone location, staghorn stone, stone compositions, American Society of Anesthesiologists score (ASA), blood loss, hospitalization, stone-free status and postoperative complication within 30 days. Missing values of continuous variable were imputed by mean values respectively, and missing values of categories variables were imputed with their mode values. The difference in these variables between complication group and complication free group were analyzed. Chi-square test was used to analyze categorical variables and student t test for analyzing continuous variables. Numerical variables were noted as mean $\pm$ Standard Deviation (SD) and categorical variables was noted as number and percentage. P-values were two-tailed and statistical significance was set to 0.05. Statistic analysis was done using Python (Python Software Foundation, https://www.python.org/) and a python package named scipy[114].

4.2.2 Postoperative Complication Classification System

The Clavien classification system has been widely used to asses the surgery complications since proposed nearly three decades ago [20]. In order to promote its applicability and accuracy, the system was changed and refined in 2004, denoted as Clavien-Dindo classification system [24]. Clavien-Dindo classification system has received great recognition during the past decade, and it has been widely used for evaluating postoperative complication of urological procedures[65]. Therefore, we use the Clavien-Dindo classification system, shown as Table4.1, to label complications after PCNL.

Table 4.1: postoperative complication grading of PCNL based on Clavien-Dindo classification system

Grading	**Description**
Grade 0	*no complications*
Grade I	Any deviation from the normal postoperative course *without the need for intervention*
Grade II	*Minor complications* requiring pharmacological treatment, blood transfusion and total parenteral nutrition
Grade III	Severe complications requiring surgical, endoscopic or radiological intervention
Grade IV	Life threatening complications requiring intensive care unit management
Grade V	Death

4.2.3 S.T.O.N.E. Nephrolithometry

S.T.O.N.E. nephrolithometry is a kidney stone score system used to asses and predict PCNL outcomes including stone-free status and postoperative complication. The system includes five most clinically relevant and reproducible variables that have been showed to impact PCNL outcomes. These five variables are stone size, PCNL tract length (skin-to-stone distance), obstruction (presence of hydronephrosis), number of involved calices, and stone essence (stone density), measured from preoperative CT.

Firstly, the stone size is calculated by aggregating the estimated volume of each stone using the formula4.1, where n is number of kidney stones. According to calculated area of 0-399, 400-799, 800-1599, and $\geq 1600mm^2$, stone size is scored to 1 to 4 respectively.

$$stone\_size = \sum_{i=1}^{n}(\frac{1}{4} \times \pi \times Length_i \times Width_i) \tag{4.1}$$

Secondly, the tract length means the average vertical distance from skin to the center of the stone. The score of tract length smaller than 100 mm is 1, others is 2. Thirdly, the obstruction has two levels score, score 1 for patient without obstruction or mild dilation, score 2 for patients with moderate to severe dilation. Fourthly, the number of calices containing stones. Score 1 is assigned if only one calix is involved, score 2 is assigned if two or more calices are affected, score 3 is assigned to patients with full staghorn calculus. Lastly, stone density score is 1 or 2 according to density threshold smaller than 950 or larger than 950 Housfield units respectively. Several studies has validated S.T.O.N.E. nephrolithometry. Several studies confirmed that the model was significantly associated with stone-free status and overall complications[1, 104]. Therefore, we select it as our base line in the prediction part.

4.3 Results

We present the statistical analysis results in this section, and conduct experiments to evaluate the performance of the proposed method.

4.3.1 Statistical Analysis of PCNL Patients

A total of 3,292 adult patients with large renal stones has been cured by PCNL between 2012 and 2019 in First Affiliated Hospital of Gannan Medical University in Jiangxi, China. According to postoperative complication status, we classify patients to two groups, complication free group and complication group. Complication free group includes 2461 patients without any complications, and complication group includes 651 patients with any complications. The overall complication rate is 19.78%. Table 4.2 summarises parts of patients' demographic characteristics, disease history, laboratory test variables, preoperative variables, and operation outcome according to complication status.

Table 4.2: Percutaneous Nephrolithonomy Patients characters (n= 3292)

Characteristic	complication-free	complication	Overall	P-value
Number	2461	651	3292	
Gender,n (%)				**0.001**
Female	1254 (47.48)	358 (50.53)	1612 (48.97)	
Male	1387 (52.51)	293 (49.46)	1680 (51.03)	
Age (years) — Mean (SD)	48.72 (11.99)	48.06 (12.67)	48.59 (12.13)	0.213
BMI (kg/m^2)— Mean (SD)	22.64 (3.19)	22.28 (3.41)	22.58 (3.24)	**0.011**
Disease hisroty, n (%)				
Hypertension	439 (16.62)	78 (11.98)	517 (15.70)	**0.004**
Diabetes	119 (4.51)	21 (3.23)	140 (4.25)	0.180
Cardiac	41 (1.55)	5 (0.77)	46 (1.40)	0.180
Liver	255 (9.65)	33 (5.07)	288 (8.75)	<**0.001**
brain vessel	38 (1.43)	8 (1.23)	46 (1.40)	0.824
Pulmonary	106 (4.01)	13 (2.00)	119 (3.61)	**0.019**
Laboratory test variables				
Scr, Mean (SD)	97.24 (60.14)	97.75 (65.55)	97.34 (61.24)	0.850
Bun, Mean (SD)	5.55 (5.63)	5.32 (2.93)	5.50 (5.21)	0.322
Alt, Mean (SD)	18.68 (15.10)	18.14 (12.89)	18.57 (14.69)	0.400
Ast, Mean (SD)	19.48 (9.83)	19.60 (9.20)	19.50 (9.71)	0.782
Protein, Mean (SD)	40.92 (4.32)	40.76 (5.00)	40.89 (4.46)	0.405
Blood-wbc, Mean (SD)	7.51 (10.22)	7.36 (4.75)	7.48 (9.39)	0.714
Urine-wbc, Mean (SD)	2.11 (1.51)	2.21 (1.54)	2.13 (1.51)	0.153
Urine-culture, n (%)	2165 (81.98)	425 (65.28)	2590 (78.68)	<**0.001**
Heart-cu, Mean (SD)	62.8 (13.04)	62.41 (3.64)	62.72 (11.79)	0.454
Heart-ct, n (%)	721 (27.30)	195 (29.95)	916 (27.86)	0.192
Preoperative variables				
Stone size (mm^2) — Mean (SD)	165.91 (173.76)	196.47 (565.19)	171.96 (295.74)	**0.018**
Staghorn stone, n (%)	209 (7.91)	64 (9.83)	273 (8.29)	0.131
Multiple stones, n (%)	2519 (95.38)	628 (96.47)	3147 (95.60)	0.270
Kidney-anomaly, n (%)	130 (4.92)	39 (5.99)	169 (5.13)	0.314
Kidney puncture, n (%)	248 (9.39%)	95 (14.59)	343 (10.42%)	<**0.001**
Stone composition number, Mean (SD)	1.10 (0.41)	1.09 (0.40)	1.10 (0.40)	0.349
Hydronephrosis, Mean (SD)	1.86 (0.95)	1.89 (0.92)	1.87 (0.94)	0.551
ASA level, Mean (SD)	1.81 (0.48)	1.79 (0.49)	1.81 (0.48)	0.389
Stone location				<**0.001**
Calyx	1401 (53.05)	339 (52.07)	1740 (52.85)	
Pelvis	380 (14.40)	168 (25.81)	548 (16.65)	
Ureter	210 (7.95)	28 (4.30)	238 (7.23)	
Multiple locations	650 (24.61)	116 (17.81)	766 (23.27)	
Operation results				
Blood loss, Mean (SD)	13.83 (11.31)	17.00 (13.17)	14.46 (11.77)	<**0.001**
Hospitalization (day), Mean (SD)	8.24 (3.93)	11.06 (6.33)	8.80 (4.64)	<**0.001**
Operation time (min), Mean (SD)	85.42 (30.59)	89.54 (39.82)	86.24 (30.48)	**0.002**
Residue stone n (%)	1053 (42.79)	289 (44.39)	1342 (40.77)	**0.040**
Operation number, Mean (SD)	1.13 (0.36)	1.25 (0.49)	1.16 (0.39)	<**0.001**

Kidney stone cases and complication rate distribution over demographic characteristics

There were 1612 (48.97%) female patients and overall patients mean age was 48.59 ± 12.13 years. Mean BMI of complication free group was 22.64 ± 3.19, complication group 22.28 ± 3.41. P-value of demographic features Gender, age, BMI were 0.001, 0.213 and 0.011 respectively, which indicates that gender and BMI were statistically associated with the complications.

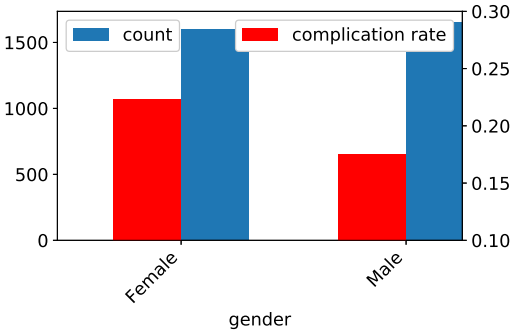

Figure 4.1: Complication distribution on patients' gender (left axis: PCNL patient count represented by the blue bar, right axis: postoperative complication rate of PCNL patient represented by the red bar)

Fig. 4.1 presents kidney stone cases and complication ratio distribution over patient's gender. The figure shows that the number of female cases was nearly equal to the number of male cases. However, the complication rate of female was obviously higher than male, which might because surgery on women is more complicated due to their physiological structure.

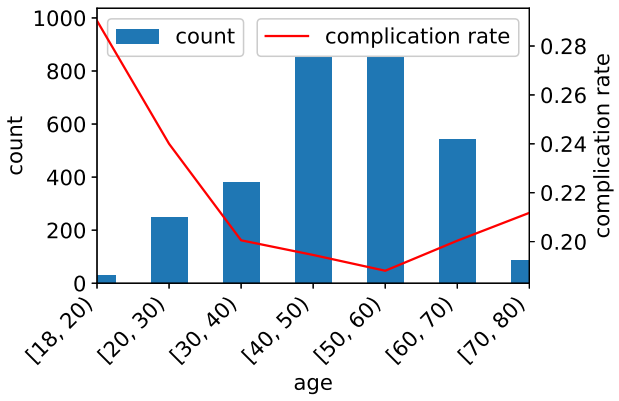

Figure 4.2: Postoperative complication distribution on PCNL patient's age (left axis: PCNL patient count represented by the blue bar, right axis: postoperative complication rate of PCNL patient represented by the red line)

Fig. 4.2 presents complication distribution over patient's age. Kidney stone disease occurs frequently in two age groups (i.e., $[40, 50)$ and $[50, 60)$). However, the complication rates of these two groups are relatively lower than that of others and the complication rate of younger cases group $[18, 20)$ is the highest. The phenomenon indicates that clinicians might be more

proficient in surgical skills in the age group with many cases, thus the complication rate of the corresponding age group will be lower.

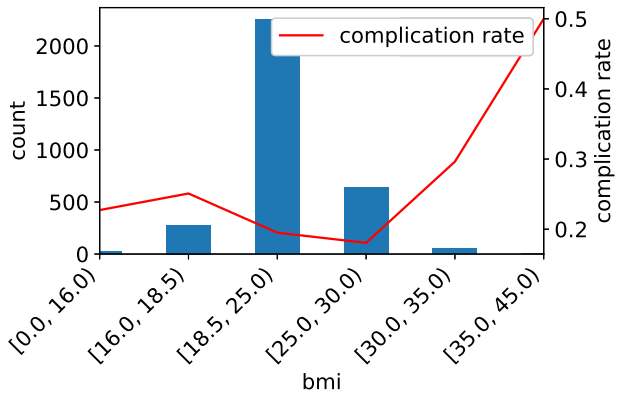

Figure 4.3: Postoperative complication distribution on PCNL patient's BMI (left axis: PCNL patient count represented by the blue bar, right axis: postoperative complication rate of PCNL patient represented by the red bar)

In terms of BMI, we can see from the Fig 4.3 that the BMI of most kidney stone cases is under 30 and obese patients ($BMI > 30$) have obviously higher complication rate than other patients. The complication rate arises quickly when BMI is over 30 and the possible reason might be it is more complicate to conduct PCNL operation on obese cases as the distance between skin and kidney stone is much lager than other cases.

Kidney stone cases and complication rate distribution over disease history

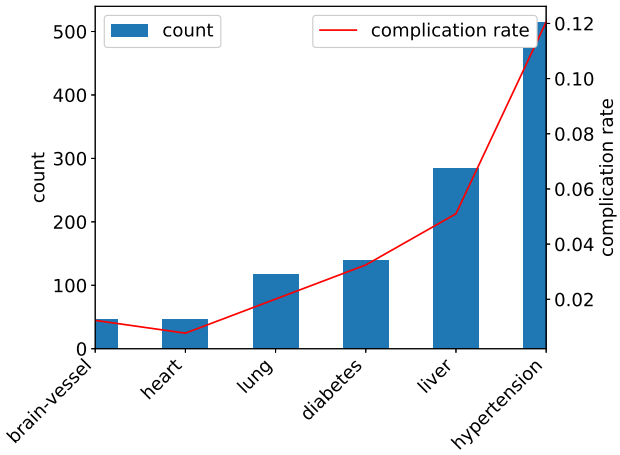

Figure 4.4: Postoperative complication distribution on PCNL patient's disease history variables (left axis: PCNL patient count represented by the blue bar, right axis: postoperative complication rate of PCNL patient represented by the red line)

In this part, we present statistical analysis result of disease history according to complication levels. Fig. 4.4 shows six different disease history frequency and complication rate of PCNL

patients, hypertension, diabetes, heart (cardiac), liver, brain vessel, lung (pulmonary). The top two common diseases are hypertension, liver in 15.70% and 8.75% of patients, respectively. The red line presents the complication rate of kidney stone cases, and all of them are far blow the averaged complication rate 19.78%. Table 4.2 tells that three disease, hypertension ($p=0.004$), liver ($p<0.001$) and pulmonary ($p=0.019$) are associated with the postoperative complication.

Correlation between complications and laboratory test

The hospital did a series test on PCNL patients, such as ast, alt, scr, bun, protein and so on. Only one laboratory test variable urine-culture ($p<0.001$) is significantly associated with postoperative complication. Fig. 4.5 presents the Pearson correlation heat map of laboratory test variables and complication level. Protein and urine culture have slightly negative correlation with complication level, which means that patients with lower protein value or abnormal urine culture are more likely to suffer from complications. There are neither strong Pearson correlation nor statistically significant difference between complication and other variables.

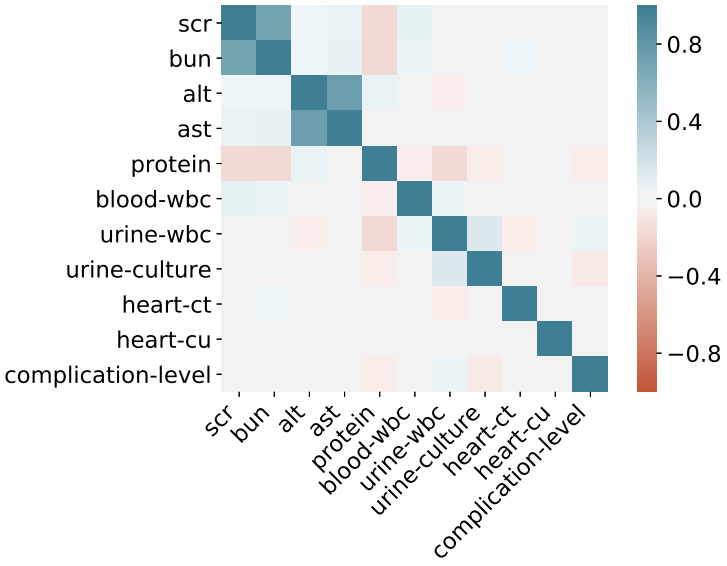

Figure 4.5: Pearson correlation of laboratory test variables and postoperative complication level of PCNL patients; Pearson correlation ranges from -1 to 1

Complication distribution over preoperative features

Table 4.2 indicates that staghorn stone ($p=0.131$), multiple stones ($p=0.270$), stone composition number ($p=0.349$), hydronephrosis ($p=0.551$), abnormal kidney ($p=0.314$) and ASA level ($p=0.626$) are not statistically associated with postoperative complication, while stone size ($p=0.018$), stone location ($p<0.001$) and kidney puncture ($p<0.001$) are. Fig. 4.6, Fig. 4.7 and Fig 4.8 are used to present how differently complication free (Grade 0) patients, low level complication (Grade I-II) patients and high level complication (Grade III-V) patients

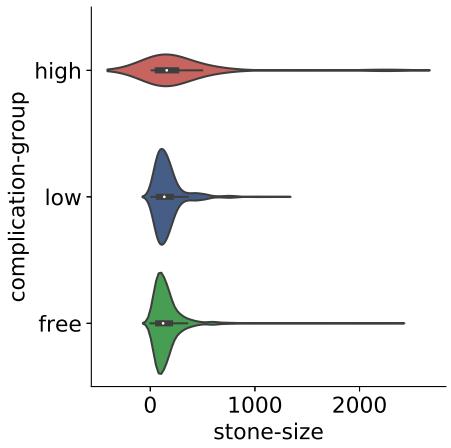

Figure 4.6: Postoperative complication groups distribution over stone size of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

distributed over these stone size, stone position and kidney puncture. Fig. 4.6 indicates that there is no big difference of stone size between complication free group patients and low level complication group patients, while the stone size of high complication patients is larger than other two groups. While stone location ($p<0.001$) is another feature that has great impact on postoperative complication. We categorized stone locations to four kinds, calyx, pelvis, ureter and multiple locations (more than one location). As Fig 4.7 shows patients with kidney stone located at pelvis have the highest complication rate of both low and high level complications, which indicates that kidney stone in pelvis is an important predictor of postoperative complications. In terms of kidney puncture, patients who have had a kidney puncture are more likely to have complications.

Complication distribution over operation results

Table 4.2 shows that all five operation results are statistically associated with the complications. The mean of blood loss of complication group patients was $17.00 \pm 13.17mL$, and that of complication free group patients was $13.83 \pm 11.31mL$. It indicates that patients with lager blood loss had bigger probability to contract complications. Patients of complication group went through longer operation (89.54 ± 39.82min), and that of complication free patients was $85.42 \pm 30.59min$. Average hospitalization (11.06 ± 6.33 days) of complication group was nearly three days longer than that (8.24 ± 3.93 days) of complication free group. Patients with complications have a higher complication residual stone rate than complication free patients (44.39% vs 42.79%; p=0.040). Additionally, complication group patients averagely had operations than the other group (1.25 ± 0.49 vs 1.13 ± 0.36; $p<0.001$).

Fig. 4.9 shows the difference of complication free (Grade 0) patients, low level complication (Grade I-II) patients and high level complication (Grade III-V) patients in blood loss, hospitalization, operation time. Vertical lines were at mean values of each group. As presented in Fig.

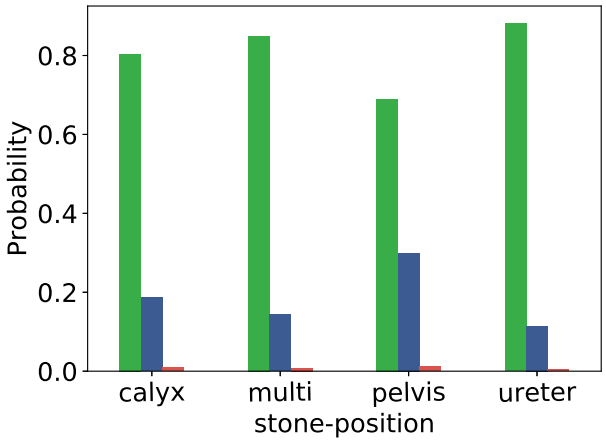

Figure 4.7: Postoperative complication groups (free, low, high) distribution over stone location of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

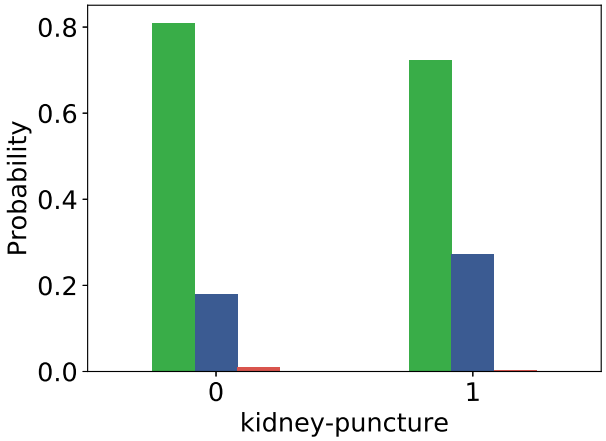

Figure 4.8: Postoperative complication groups (free, low, high) distribution over kidney puncture of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

4.9 left part, the average blood loss of high level complication patients was lager than that of low level complication group patients, which was lager than that of complication free group.

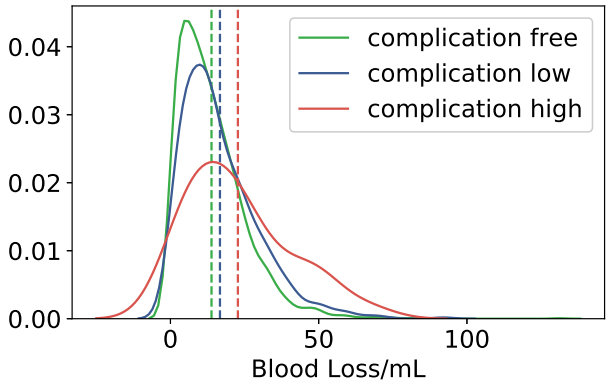

Figure 4.9: Fitted densities of postoperative complication groups (free, low, high) over blood loss of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

Fig. 4.10 shows that hospitalization density estimations of three groups were similar with that of blood loss, which verifies the conclusion we get from statistical analysis. The average hospitalization increases as the complication level increases. The high level complication group has the largest average hospitalization, in other words, they need to stay longer to recover after operations.

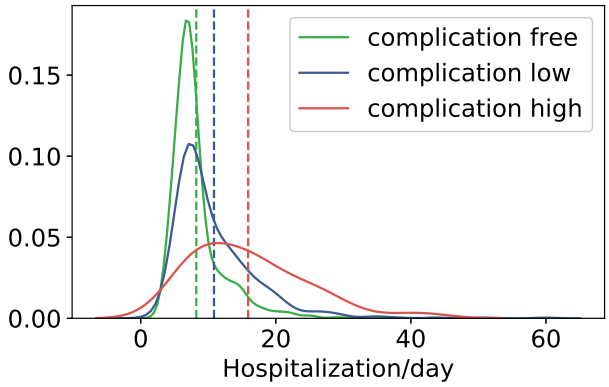

Figure 4.10: Fitted densities of postoperative complication groups (free, low, high) over hospitalization of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

Operation time density distribution was presented at right of Fig. 4.11. Although mean operation time of these three group was almost the same, we could see the density line of high level complication group spanned broader than that of low level complication group on the right side, which means patients with high level complications likely undergo longer surgery.

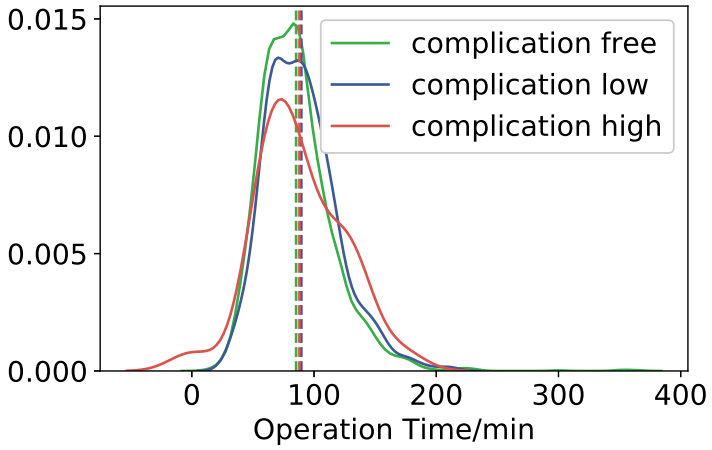

Figure 4.11: Fitted densities of postoperative complication groups (free, low, high) over operation time of PCNL patients (complication free group represented by green color, low level complication group represented by blue color, high level complication group represented by red color)

4.3.2 Prediction Results

Experiment Setup

S.T.O.N.E. nephrolithometry is selected as baseline to compare with our proposed model. As described in previous session, nephrolithometry scores kidney stone patients, then use multivariate logistic regression model to predict postoperative complications. Frequently used machine learning models, such as K-Nearest Neighbors (KNN)[21], Random Forest (RF)[11], Support Vector Machine (SVM)[99], neural network (multi perceptron)[37], were deployed on our dataset for comparison.

After data structuring and missing value imputation, we split the dataset into training set and testing set. Training set contained 80% of the overall dataset and it was used to training prediction models. The remaining 20% of data was testing set, which was used to estimate the models' performance. Then we conducted extensive experiments to evaluate the prediction performance. All experiments were implemented with using python.

S.T.O.N.E. nephrolithometry V.S. machine learning classifiers

S.T.O.N.E. nephrolithometry is set as the baseline in the prediction part. It uses kidney related features and its stone score as input, then uses multivariate logistic regression to predict postoperative complication of PCNL. The nephrolithome is compared to classical classifiers, such as SVM, RF, KNN, neural network (multi perceptron, also noted as MLP) and XGBoost.

The nephrolithometry only used kidney stone features and score based on these features, thus same features except nephrolithometry score were fed to other classifiers. We choose AUC

and F1 score as performance metrics. The result is presented in Table 4.3. The AUC of the nephrolithometry is 0.5, and its F1 score is 0, which means it can barely predict the minority patients. None of other classifiers are able to achieve an AUC larger than 0.51. While, XGBoost achieved the best performance of all models with an AUC of 0.5458.

Table 4.3: Prediction performance of traditional models on postoperative complication with using kidney stone features from the PCNL dataset

Model	AUC	F1 score
S.T.O.N.E.	0.5000	0
SVM	0.5001	0.0269
RF	0.5023	0.0132
KNN	0.5045	0.1499
MLP	0.4989	0
XGBoost	**0.5458**	**0.1949**

Postoperative complication prediction model: SMOTE-XGBoost

A total of 651 postoperative complication were observed in 3292 procedures. Grade I was recorded 467 (14.19%), grade II in 155 (4.71%), grade III in 18 (0.55%), and grade IV in 10 (0.53%). One patient was dead (Grade V) after PCNL operation, accounting 0.03%. The postoperative complication (see Fig. 4.12) shows postoperative complication level distribution of PCNL patients. And we can see that the risk level distribution is imbalanced, the higher complication level, the fewer PCNL patients. The postoperative complication prediction will be formed as a binary classification problem, which are no complication group (complication level 0), complication group (complication level I-V). From Fig. 4.12 we can see that the postoperative complication distribution is highly imbalanced, the complication rate is 19.78%. Imbalanced ratio (the ratio of the majority amount and the minority amount) of our data set is 4.01.

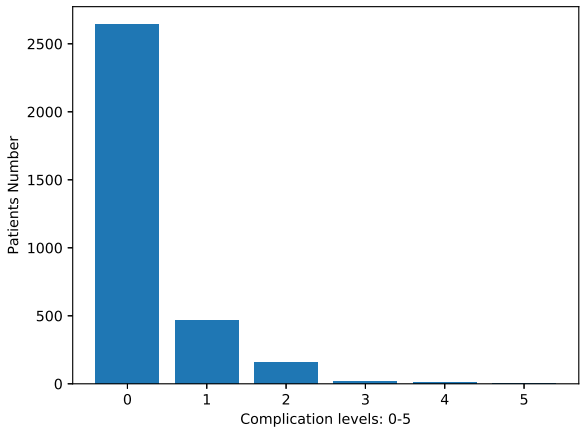

Figure 4.12: Postoperative complication distribution by Clavien-Dindo classification system from level 0 to level 5

Because the distribution of complication is highly imbalanced, we implement the sampling-based method SMOTE-XGBoost to predict the complications. SMOTE-XGBoost only resamples the training set by SMOTE, and put testing set aside. By doing so, we could balance the training set and guarantee the purity of testing set. Then the balanced training set was used to train XGBoost. Another method to deal with the imbalanced problem is cost-sensitive learning. With setting positive class weight as imbalanced ratio 4.01, XGBoost classifier could be transformed to cost-sensitive XGBoost. Other classifiers combined with SMOTE were compared as baselines. In this experiment, we used kidney stone features that S.T.O.N.E. nephrolithometry used except its stone score.

As Table 4.4 presents, all models combined with SMOTE get higher AUC and F1 compared with themselves in previous part, which proves SMOTE is a good solution to our problem. Our model SMOTE-XGBoost achieved an AUC of 0.6140, which is a great improvement from an AUC of 0.5458. Our model outperforms cost-sensitive XGBoost and other models combined with SMOTE.

Table 4.4: Prediction performance of SMOTE-XGBoost and other models with using kidney stone features

Method	AUC	F1
SMOTE-XGBoost	**0.6140**	**0.8874**
cost-sensitive XGBoost	0.5708	0.3382
SMOTE-S.T.O.N.E.	0.5184	0.3049
SMOTE-SVM	0.5006	0.2121
SMOTE-RF	0.5495	0.2817
SMOTE-KNN	0.5134	0.2772
SMOTE-MLP	0.5150	0.2915

Prediction performance of different feature sets

As we can see from the statistical results, there are lots of features statistically significantly associated with postoperative complication other than kidney stone features. We want to find out prediction performance of these features and whether adding these features to our model will improve the prediction performance. AUC and F1 were still used as performance metrics in this experiment.

Features were grouped to Kidney Stone (KS), Laboratory Test (LT), Disease History (DH), OPeration related variables (OP), Stone Composition (SC), similar to analysis part. Firstly, these five feature sets were fed to train the SMOTE-XGBoost. Results is shown in the upper part of Table 4.5. Among five variable groups, the prediction performance of KS achieved highest AUC and F1 score, while disease history had the smallest AUC and F1. The results indicates that kidney stone feature set is the best feature set for predicting complication.

Secondly, feature groups were added to kidney stone features one by one. Their prediction performance metrics were computed respectively. The results is presented in Table 4.5. AUC of SMOTE-XGBoost increased with adding feature sets. Therefore, adding features is beneficial

to increase the performance of our model to predict postoperative complication. Specially, when all features were used, our model achieved an AUC of 0.7077, while its F1 score of 0.8871 was very close to the highest F1 of 0.8882. Overall, AUC of our model SMOTE-XGBoost was improved from 0.6104 to 0.7077 via adding other features of PCNL patients.

At last, we selected features that were statistically significantly associated with postoperative complication, denoted as statistical selected feature set. Then we computed prediction performance of statistical selected feature set. However, the prediction performance is worse than the result when using all features. It indicates that features, which are not statistically associated with complications, are beneficial to improve the prediction performance.

Table 4.5: Prediction performance of different group features by SMOTE-XGBoost

Feature	AUC	F1
Kidney stone (KS)	**0.6104**	**0.8874**
laboratory test variables (LT)	0.5846	0.8789
disease history (DH)	0.5380	0.4241
operation variables (OP)	0.5884	0.8780
stone composition (SC)	0.5812	0.7954
KS+LT	0.6414	0.8860
KS+LT+DH	0.6437	0.8856
KS+LT+DH+OP	0.6599	**0.8882**
KS+LT+DH+OP+SC	**0.7077**	0.8871
Statistical selected features	0.6645	0.8866

4.4 Summary

There is an increasing interest in postoperative complication prediction which could help physicians and hospitals make preparation prior to operation or refer the challenging cases to more experienced centers. With using kidney stone disease as an examplar, we evaluate the sampling-based approach SMOTE-SGBoost to address the class-imbalance problem and limitations of postoperative complication prediction of PCNL.

In the analysis part, we compared demographic characteristics, disease history, laboratory test variables, preoperative variables and operation outcome according to the complication status, which was skewed distributed. Variables statistically associated with the complication were identified. Our analysis represents that female patients, young patients, obese patients has higher complication rate; complication rate of patients with any kind of disease is surprisingly lower than those without disease history; urine culture is the only laboratory test variable that statistically significant associate with the complications. Furthermore, high level complication patients are likely to have larger stone size, loss more blood, experience longer operation time, and stay longer in hospital.

With 19.78% patients who had postoperative complication, the complication distribution was highly imbalanced. It makes the prediction of complications an imbalanced classification

problem. To solve this problem, we implement the sampling-based method SMOTE-XGBoost. Overall, our approach outperforms the S.T.O.N.E. nephrolithometry and current machine learning methods in both AUC and F1-score. To find out prediction performance of different features and whether adding these features to our model could improve the prediction performance, features were grouped to five different sets: kidney stone features, laboratory test variables, disease history, operation variables, and stone compositions. Results show that kidney stone features achieves best AUC compared with other feature sets. In addition, different feature sets of PCNL patients were added to our model. When all features are used, our model achieves an AUC of 0.7077, which is 41.54% higher than S.T.O.N.E. nephrolithometry. Lastly, we compared the prediction performance of feature set statistically associated with complications with that of using all features. The results indicates that adding other features not statistically associated with complications would help improve complication predicting accuracy.

In the future, larger data collected from more hospital could be analyzed and used to improve prediction performance of our method. Although our method outperforms S.T.O.N.E. nephrolithometry, KNN, RF, SVM, neural networks and XGBoost, there is still a quite distance from an AUC of 0.7077 to perfect classification. Additionally, we formed the postoperative complication as a binary classification problem due to the lack of data and imbalanced distribution. With lager sample size of PCNL patients, we could build a multi-class complication prediction model with higher accuracy. Lastly, the structuring of the radiological report provides benefits to improve medical practice and diagnoses according to[91]. We will use deep learning methods to get more precise report of kidney stones, which could be fused in our proposed prediction model to achieve better performance.

Chapter 5

Evaluating Multiple Balance Subsets Stacking on Imbalanced Structured Datasets

Accurate prediction is highly important for clinical decision making and early treatment. In this chapter [1] , we study the imbalanced data problem in prediction, a key challenge existing in the healthcare area. Imbalanced datasets bias classifiers towards the majority class, leading to an unsatisfied classification prediction performance on the minority class, which is known as imbalance problem. Existing imbalance learning methods may suffer from issues like information loss, overfitting, and high training time cost. As described in Section 3.2.2, this dissertation proposes a ensemble method called MASS which avoid the problem of information loss and overfitting. Furthermore, Parallel MASS could reduce the training time cost. This dissertation evaluates MASS on three real-world structured medical datasets, and experimental results demonstrate that its prediction performance outperforms the state-of-art methods in terms of AUC, F1-score and MCC. Through the speedup analysis, Parallel MASS reduces the training time cost greatly on large dataset, and its speedup increases as the data size grows.

Contents

[1] This chapter is based on the paper I published in 26th IEEE International Conference on Parallel and Distributed Systems (ICPADS) [97].

5.1 Introduction

Accurate prediction is of significant importance for clinical decision making, early treatment and patient counseling. Achieving accurate medicine and improving the quality of patient care are the overall objective in healthcare area. With the rapid increasing application of Electronic Health Records (EHR) systems in many healthcare facilities, it is possible to get enough medical data to achieve this goal more efficiently. Nevertheless, prediction based on medical datasets has been an intriguing and challenging topic because of its inherent imbalanced nature. Medical datasets are mainly composed of "healthy" samples with only a small section of "sick" samples, leading to the so-called class imbalance problem [70]. The imbalance problem could bias classification algorithms to majority class, so that classifiers have weak performance on minority class. Such classifiers are not useful in real world tasks, because usually the classification performance of the minority samples is of higher importance for decision making in the healthcare area [8].

A series of imbalance learning methods have been proposed to overcome the imbalance problem in healthcare dataset and can be clustered into three main classes: sampling approaches, cost-sensitive learning approaches and ensemble learning approaches. Sampling approaches have been proved effective to improve the classification performance of classifiers used to predict chronic kidney disease [129], diabetes and liver disorders [70]. As elaborated in Section 2, existing sampling methods suffer from problems, such as information loss, huge computational cost and overfitting. Cost-sensitive learning was used to deal with imbalance problem of healthcare dataset in a fast imbalance classification framework [90]. The challenge of cost-sensitive methods is how to determine a cost matrix, and the defined cost matrix may not be generalized to any other tasks. Ensemble learning approaches usually combine sampling approach or cost-sensitive approach with ensemble learning algorithm to address the imbalance problem [52, 33]. However, they inherently suffers from issues of sampling approaches and cost-sensitive approaches. Moreover, some ensemble methods have the problem of high training cost when applied on large real world tasks, as shown in SMOTEBagging [119] and SMOTEBoost [15].

As introduced in Section 3.2.2, MASS is able to avoid or alleviate these issues of exiting methods. Instead of simply creating a balance training set or defining a cost matrix, MASS first generate multiple balance subsets to train base classifiers. Then MASS generate a stacking dataset based on the base classifiers, which keeps the same label as original dataset. After that, the stacking dataset is used to train a stack model, which could optimize the weights of the base classifiers to get a strong ensemble classifier. MASS does not reduce majority samples or generate new meaningless samples, thus will not suffer from the problem of information loss. MASS does not duplicate any minority samples, thus preventing the issue of overfitting. Specially, as the training processes of the base classifiers and the stacking dataset generation are independent, the main part of MASS could be run in parallel. Parallel MASS could decrease the

training time cost, which is of high importance as the scale of healthcare dataset is increasing rapidly.

To evaluate the proposed MASS, we extract three real-world medical datasets, namely acute kidney failure and diabetes from Medical Information Mart for Intensive Care (MIMIC) III dataset and PCNL dataset collected from the First Affiliated Hospital of Gannan Medical University in China. These three datasets have different degrees of imbalance problems. We conduct extensive experiments to evaluate the classification performance of MASS by comparing it with other imbalance learning baseline methods on these three real datasets. Besides, to validate the robustness of MASS, we apply MASS and other ensemble learning methods with different base classifiers. Finally, we analyze the speedup of Parallel MASS on different scales of PCNL dataset.

In all, we mainly have the following contributions in this work:

- We conduct extensive experiments to evaluate the proposed MASS. Experimental results show that MASS greatly outperforms baseline methods on three different real world healthcare datasets. For example, compared with SPEnsemble [74], MASS improves the classification performance 3.22% in AUC, 3.10% in F1score,improves 2.58% in MCC when applied to the diabetes dataset.
- The robustness of MASS is validated by applying different base classifiers.
- We analyse the speedup of running Parallel MASS over different scales of dataset. The results demonstrate that running MASS in parallel can reduce the training time cost greatly on large datasets, and its speedup would increase as the data size grows. Specially, Parallel MASS reduces 101.8% training time compared with MASS at most in our experiments.

5.2 Imbalanced Structured Medical Datasets

In this section, we extract three real world healthcare datasets, acute kidney failure dataset and diabetes dataset from MIMIC III dataset and one PCNL dataset collected from the First Affiliated Hospital of Gannan Medical University in China.

The MIMIC-III dataset is a freely accessible healthcare dataset, which is collected from the Beth Israel Deaconess Medical Center over 11 years [51]. We select two datasets from MIMIC-III with different imbalance degrees: acute kidney failure dataset and diabetes dataset.

We extract all relevant medical information of adult patients (age$\geq$18) according to ICD-9 codes for these diseases, respectively. All the data come mainly from following tables: patients, admissions, and lab events. Firstly, we need to calculate the age of each patient by using the difference between their date of birth and the date of their first admission. Patients who are 18 years old or older are defined as adults. After that, we extract the corresponding data of each

adult patient according to laboratory parameters of each disease. Typically, we calculate the average of the features because most of the features are measured multiple times.

5.2.1 Acute Kidney Failure

Acute Kidney Failure (AKF) is characterized by a sudden loss of kidney ability resulting in the retention of nitrogen wastes and water-electrolyte and acid-base imbalance. Acute Tubular Necrosis (ATN) is one of the major causes. We select two related diagnoses as classes of an imbalanced dataset. One is “Acute kidney failure with lesion of tubular necrosis" and the other is “Acute kidney failure without lesion of tubular necrosis".

Extracted patient’s laboratory items are listed here: creatinine (in blood), urea nitrogen (in blood), protein (in urine), White Blood Cell (WBC) (in urine), Red Blood Cell (RBC) (in urine), sodium (in urine) and osmolality (in urine).

An important indicator of renal function is the estimated Glomerular Filteration Rate (eGFR) based on the following abbreviated MDRD equation:

$$\begin{aligned} GFR = 175 \times (Scr)^{-1.154} \times (Age)^{0.203} \\ \times (0.742 \text{ if female}) \times (1.212 \text{ if African American}) \end{aligned} \tag{5.1}$$

In total, 8 features from MIMIC-III are collected into this imbalanced dataset. In this dataset, labels 0 and 1 indicate the type of patient’s disease. Label 0 indicates that an AFK patient is diagnosed without ATN. The number of recorded samples with label 0 is 7,300. Label 1 indicates that an AFK patient is diagnosed with ATN. The number of recorded samples with label 1 is 2,182. The number of total samples is 9,482 and IR is 3.35.

5.2.2 Diabetes

Diabetes is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. In some cases, diabetes may result in kidney diseases. Diabetic nephropathy is known as a common cause of kidney failure. Diabetes patients with and without renal manifestation are selected as two classes of an imbalanced dataset.

Extracted patient’s laboratory items: creatinine (in blood), glucose (in blood), albumin (in blood), pH (in urine), protein (in urine), WBC (in urine), and RBC (in urine). A total of 7 extracted laboratory features from MIMIC-III are collected in this imbalanced dataset. In this dataset, labels 0 and 1 indicate the type of patient’s disease. Label 0 indicates that a diabetes patient is diagnosed without renal manifestation. The number of recorded samples is 9,462.

Label 1 indicates that a diabetes patient is diagnosed with renal manifestation. The number of recorded samples is 1,075. The number of total samples is 10,537 and the IR is 8.80.

5.2.3 PCNL

The PCNL dataset used in this work is collected from the First Affiliated Hospital of Gannan Medical University in China. The dataset spans from **January 2012** to **July 2019**, which contains 3,293 PCNL patients' medical records.

As most of the PCNL patient records are recorded by unstructured clinical notes, we first structure the records and then extract features from them. There are three kinds of features in the structured records, i.e., numeric features, category features and clinic notes. Numeric features could be directly used in machine learning algorithms. Category features are transfer into numeric features by one hot encoding. We use keyword matching method to extract different kinds of features from clinic notes, then use one hot encoding to transfer them into numeric features respectively. With the doctors' advice, we set reasonable ranges for features and clean records automatically.

Finally, we extract 39 features from the unstructured clinical notes and there are 3,293 patients with kidney stones treated by PCNL. 2,642 patients have no complications after operation, and they are labeled as 0. 651 patients have different kinds of postoperative complications, and these patients are labeled as 1. Accordingly, the IR of this dataset is 4.01.

5.3 Experimental Results

In this section, we first evaluate MASS on three real world healthcare datasets, then validate the robustness of MASS over different base classifiers. At last, we analyze the speedup performance of Parallel MASS over MASS.

5.3.1 Experimental Setup

The prediction task is formed as a binary classification problem. Therefore, patients without respective disease or postoperative complication is labeled as False and grouped into majority set $\mathcal{F}$ ($y = 0$), while patients with respective disease or any postoperative complications is labeled as True and grouped into minority set $\mathcal{T}$ ($y = 1$).

Stratified k-fold cross-validation is used to evaluate the performance of the models to be compared. In stratified k-fold cross-validation, the data is split into k equally (or nearly equally) sized folds, each fold keep the approximately same scale of each class(i.e., majority class and minority class). Subsequently, the model is trained and validated k times, and each time a different fold is held-out for validation and the remaining k-1 folds are used for training. We

employ stratified 5-fold cross validation and repeat the process 10 times to reduce bias due to random partitioned folds.

Instead of using accuracy or error rate as evaluation metrics, AUC (area under the ROC curve) and F1-score are commonly used for evaluating the performance on minority class, and we also consider Matthews Correlation Coefficient (MCC) to evaluate our prediction model [74].

5.3.2 Prediction Performance of MASS

In this subsection, we conduct experiments to compare MASS with all three kinds of imbalance learning methods: sampling approaches, cost-sensitive learning approaches, ensemble learning approaches.

Comparison with sampling based approaches

MASS is compared with three under-sampling methods, three oversampling methods and one hybrid sampling method:

- Random Under-Sampling(RUS) randomly select $|\mathcal{T}|$ majority samples to get a subset $\mathcal{F}'$, and combine it with minority set $\mathcal{T}$ to get a balance training set.

- Edited Nearest Neighbor (ENN) [123] removes noisy samples from the majority set for which their class is different from one of their nearest-neighbors.

- Near Miss (NM) selects $|\mathcal{T}|$ samples from the majority set $\mathcal{F}$ for which the average distance of the k nearest samples of the minority class is the smallest.

- Random Over-Sampling(ROS) randomly duplicates some minority samples to get a balance dataset for training.

- Synthetic Minority Over-sampling TechniquE (SMOTE) generates new minority data points based on the similarities between minority samples from original dataset to balance the training dataset

- ADAptive SYNthetic over-sampling (ADASYN) [39] generate more synthetic samples for minority class that are harder to learn.

- SMOTE with Edited Nearest Neighbours cleaning (SMOTEENN) [7] uses SMOTE to over sample the minority set and then use ENN to reduce noise and get a cleaner space.

Decision Tree is selected as the base classifier in MASS. Accordingly, it will be used as classifiers of the seven sampling methods for a fair comparison. We use imbalance-learn Python package 0.5.0 [69] to implement all these sampling methods on Python 3.7.

Table 5.1 lists the experimental results (AUC) of the seven sampling methods and MASS applied on the three datasets (acute kidney failure, diabetes, PCNL). Our approach MASS has the best prediction performance over the three evaluation criteria (AUC, F1-score and MCC). The performance improvement of MASS is only slightly better than other sampling methods when applied on AKF dataset. The main reason is that its IR is 3.35. Small IR means undersampling methods do not need to discard too many majority samples, thus its information loss is not that severe. Same situation is it to the problem of overfitting for oversampling methods. When applied on dataset with lager IR, such as diabetes, MASS has lager improvement compared with other sampling methods as shown in Table 5.1.

Table 5.1: Performance (AUC) of MASS VS. sampling methods on three medical datasets (i.e., AKF, Diabetes, PCNL)

Method	AKF	Diabetes	PCNL
RUS	0.710±0.012	0.846±0.014	0.574±0.031
ENN	0.710±0.013	0.828±0.018	0.586±0.017
NM	0.673±0.011	0.729±0.027	0.532±0.025
ROS	0.711±0.012	0.844±0.014	0.585±0.023
SMOTE	0.710±0.025	0.828±0.015	0.580±0.025
ADASYN	0.707±0.012	0.823±0.023	0.583±0.027
SMOTEENN	0.712±0.010	0.832±0.018	0.588±0.030
MASS	**0.714**±0.011	**0.864**±0.008	**0.605**±0.016

Comparison with cost-sensitive learning approaches

MASS is compared with three cost-sensitive learning approaches:

- Cost-sensitive Logistic Regression (LR)
- Cost-sensitive Support Vector Machine (SVM)
- Cost-sensitive Random Forest (RF)

Cost-sensitive learning approaches need to set cost matrix, and its structure in binary classification scenario is shown in Table 5.2. In the cost matrix, the cost of correct classified samples is set as 0 ($C(1,1) = 0$, $C(0,0) = 0$). The cost of misclassification of negative samples is set as 1 ($C(0,1) = 1$). The cost of misclassification of negative samples is set as IR ($C(1,0) = IR$). These three cost-sensitive classifiers are implemented by sklearn Python package [88] with Python 3.7.

Table 5.2: Cost matrix in binary classification scenario

	Positive prediction	Negative prediction
Positive class	$C(1,1)$	$C(1,0)$
Negative class	$C(0,1)$	$C(0,0)$

In parameter list of LR, SVM and RF, we set the parameter "class_weight" as "balanced" to get the cost-sensitive version of respective classifier. With such setting, the misclassification cost weights are inversely proportional to class frequencies in the training dataset.

The performance (AUC) of these cost-sensitive classifiers is shown in table 5.3. MASS achieves the best performance compared with three cost-sensitive learning methods over three datasets.

Table 5.3: Performance (AUC) of MASS VS. Cost-sensitive method on three medical datasets (i.e., AKF, Diabetes, PCNL)

Method	AKF	Diabetes	PCNL
Cost-sensitive LR	0.683±0.012	0.852±0.009	0.555±0.026
Cost-sensitive SVM	0.698±0.013	0.853±0.010	0.580±0.026
Cost-sensitive RF	0.710±0.012	0.845±0.015	0.587±0.029
MASS	**0.714**±0.011	**0.864**±0.008	**0.605**±0.016

Comparison with ensemble learning approaches

In this part, we compare MASS with six ensemble learning based approaches:

- SMOTEBoost [15] : It creates new synthetic minority samples by SMOTE to change the updating weight of in each AdaBoost iteration.

- SMOTEBagging [119] : It creates new synthetic minority samples by SMOTE as a preprocessing step for each base classifier and Bagging .

- RUSBoost [95] : It applies Random Under-Sampling to change the updating weight of within each AdaBoost iteration.

- UnderBagging [6] : It applies Random Under-Sampling as a preprocessing step for each base classifier in bagging .

- Balance Cascade [73] : It trains the learners sequentially, where in each step, the majority class examples that are correctly classified by the current trained learners are removed from further consideration.

- Self-paced Ensemble learning(SPEnsemble) [74] : It uses a self-paced factor in each under-sampling iteration to focus more on majority samples that are hard to classify and generate a final strong ensemble classifier based on base classifiers trained in each iteration.

In this part, we compare MASS with these ensemble learning methods over three evaluation metrics (AUC, F1-score and MCC) on the three datasets, and MASS always performs best. Due to the content limit, here we only show results of the diabetes dataset. For a fair comparison, decision tree is used as the base classifier of the six ensemble learning based approaches and

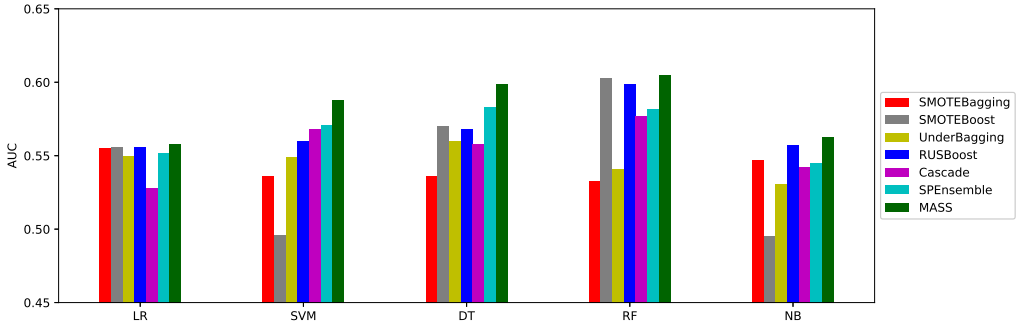

Figure 5.1: Prediction Performance (AUC) of Ensemble Learning Approaches with using different base classifiers on PCNL dataset

the proposed method. They all use the same number of base classifiers, which is $\lceil IR \rceil = 9$. The comparison results are shown in Table 5.4. MASS outperforms the two over-sampling based ensemble methods and the four under-sampling based ensemble methods over all three evaluation metrics.

Table 5.4: Multiple Balance Subsets Stack VS. Ensemble methods on diabetes dataset over three evaluation metrics (AUC, F1-score, MCC)

Model	AUC	F1-score	MCC
SMOTEBagging	0.813 ± 0.016	0.426 ± 0.030	0.364 ± 0.035
SMOTEBoost	0.823 ± 0.020	0.453 ± 0.026	0.382 ± 0.027
UnderBagging	0.828 ± 0.012	0.422 ± 0.012	0.362 ± 0.014
RUSBoost	0.826 ± 0.024	0.436 ± 0.038	0.376 ± 0.039
Cascade	0.837 ± 0.013	0.437 ± 0.018	0.373 ± 0.019
SPEnsemble	0.837 ± 0.014	0.484 ± 0.029	0.427 ± 0.030
MASS	$\mathbf{0.864} \pm 0.008$	$\mathbf{0.499} \pm 0.015$	$\mathbf{0.438} \pm 0.017$

5.3.3 Robustness Analysis of MASS

In order to check the robustness of MASS, we run it on PCNL dataset with five different base classifiers, i.e., Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB). The previous six ensemble learning approaches are compared with MASS. We use AUC as the performance metric here, and the results are shown in Fig. 5.1:

- MASS greatly boosts base classifiers prediction performance except LR, and MASS performs best among the seven ensemble methods over AUC. This proves that MASS has a strong robustness over different base classifiers.
- Most of the ensemble methods combined with LR only have the similar prediction performance with an AUC near 0.55 as LR, which indicates that LR is not suitable used as base classifiers of ensemble methods.

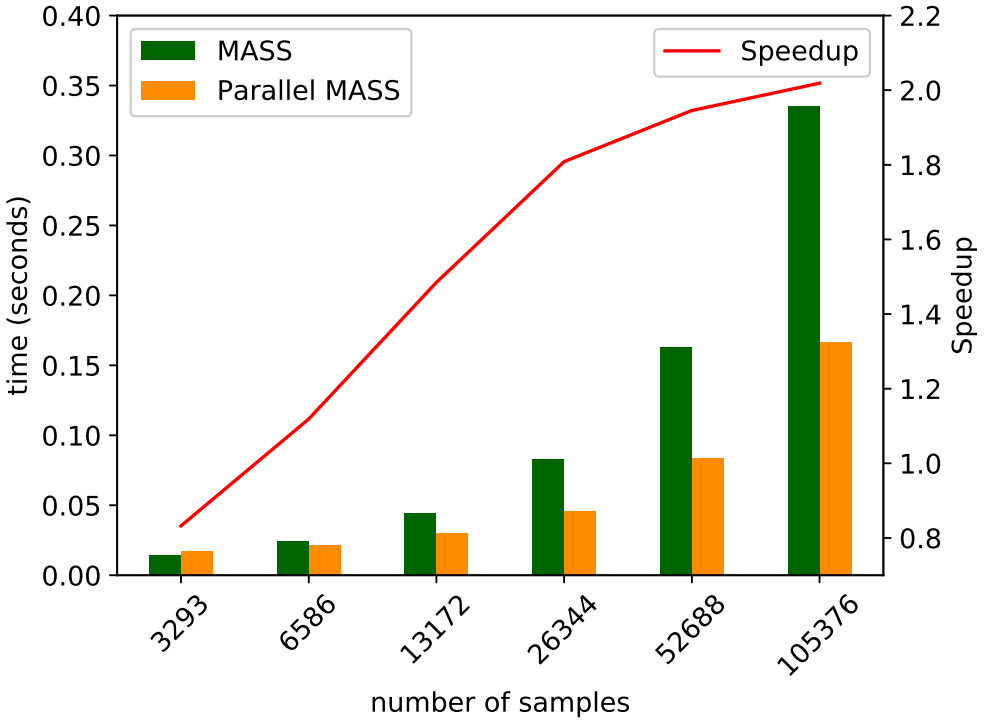

Figure 5.2: Speedup analysis of Parallel MASS over MASS on PCNL datasets (left axis: green bar represents the training time of MASS, orange bar represents the training time of Parallel MASS; right axis: the speedup of Parallel MASS over MASS)

5.3.4 Speedup Analysis of Parallel MASS

In this part, we analyze speedup of Parallel MASS compared with MASS on the different scales of PCNL dataset. We use multiprocessing Python package [43, 80] to implement Parallel MASS on Python 3.7.

In order to show the relationship between data volume and speedup, we firstly run MASS and Parallel MASS with using the original PCNL dataset on the same platform 10 times independently. Then we get the average training time of these two methods respectively. Following that, we repeat the same process but using larger dataset, which is doubled of the dataset used in the previous process. As we only focus on the impact of data volume, we just duplicate the previous dataset, then concatenate the duplicate dataset and the previous dataset. Accordingly, the data volume is doubled of the previous one. At last, we compute the speedup according to its equation. We show the relationship between training time and data volume, and the speedup curve in Fig. 5.2.

Fig. 5.2 shows the training time of MASS keeps nearly linear growth with data volume, while the training time of Parallel MASS in not that sensitive. The speedup of Parallel MASS increases as the data volume grows. However, the training time of Parallel MASS is larger than

that of MASS when applied on the original PCNL dataset, and thus the speedup is smaller than 1. We will checkout this interesting phenomenon by the equation of speedup:

$$Speedup < 1 \tag{5.2}$$

$$\frac{T_{base} + T_{data} + T_{stack}}{(T_{base} + T_{data})/n + T_{com} + T_{stack}} < 1 \tag{5.3}$$

$$(1 - \frac{1}{n})(T_{base} + T_{data}) < T_{com} \tag{5.4}$$

It means the communication cost of parallel processes is larger than the cost decrease of training base classifiers and stacking dataset generation. The communication cost does not grow with the increasing of data size because the process number equals to imbalance ratio ($\lceil IR \rceil = 5$), which does not change. On the contrary, the training cost of base classifiers and stacking dataset generation grow linearly with data size. Thus, the speedup of Parallel MASS will increase along with the data size, which is demonstrated by the experimental result in Fig. 5.2. Specially, the speedup is 2.018 when the data size of 32 times of PCNL dataset, which means that the speed of Parallel MASS increases 101.8% compared with MASS.

5.4 Summary

In this chapter, MASS is evaluated on three real-world medical datasets, which are intrinsically imbalanced. We compare MASS with other imbalance learning methods on three real world healthcare datasets. The experimental results show that MASS outperforms other state-of-the-art methods over AUC, F1-score and MCC. We also prove that MASS is robust over using different base classifiers. In addition, the speedup analysis proves that Parallel MASS could reduce huge training time cost when applied on large datasets. As most of the real world datasets have skewed distributions, in the future, we will test MASS on more datasets with more diverse IRs.

Chapter 6

Evaluating Hardness Aware Dynamic Loss on Imbalanced Image Datasets

Deep neural networks (DNNs) have achieved great success in image classification tasks with balanced image datasets, such as Modified National Institute of Standards and Technology database (MNIST) and Canadian Institute For Advanced Research (CIFAR). However, most of real-world image datasets are inherently highly imbalanced, which are dominated by a few classes (majority classes) and the rest classes (minority classes) are weakly presented. When training with such imbalanced image datasets, DNNs perform poorly on minority classes. To solve the imbalance problem, most existing methods leverage class frequency to assign higher weights to the minority classes. However, while some of the minority classes could be well represented by the training data, overweighting such classes will decrease the overall performance. Proposed in Section 3.3, HAD loss is designed to improve the prediction performance with using the classification hardness (i.e., misclassified probability of each sample) of each class to tune class weights dynamically during the training process of DNNs. HAD can find the optimized weights for both majority and minority classes, thus significantly improving the classification accuracy of minority classes. Extensive experimental results on real-world imbalanced image datasets show that HAD loss significantly outperforms the baselines. Especially, HAD loss improves 10.04% average precision compared with the best baseline, Focal loss, on the HAM10000 dataset.

Contents

6.1 Introduction

Deep neural networks (DNNs) have been proved very successful in computer vision domain [66]. In addition to the improved computation ability and various algorithms breakthroughs, the wide availability of labeled image datasets is another key reason for the success. Lots of the labeled image datasets, such as MNIST and CIFAR, are commonly resembled to be nearly balanced. However, class distribution of real-world image datasets is naturally imbalanced and medical image datasets are the typical examples. For instance, the number of healthy persons (majority classes) usually dominates that of lung cancer patients (minority classes) for critical applications like medical diagnosis [131]. As a result, there will be a significant drop when DNNs are applied on real-world datasets. Trained with imbalanced datasets, conventional DNNs would bias towards the majority classes, which would lead to poor accuracy for the minority samples. Nevertheless, failing to classify a patient might lead to the loss of life. Thus, it is of great importance to improve the classification performance of the DNNs on minority classes.

Previously, researchers usually use data-level methods (re-sampling) [14, 73, 89] or algorithm-level methods (re-weighting) [71, 55, 22] to tackle the imbalance problem. Re-sampling methods include over-sampling for the minority classes (adding duplicated minority samples), under-sampling for the majority classes (discarding majority samples), or hybrid sampling for both majority and minority classes. Re-weighting methods assign relatively larger weights to minority samples, which would influence the loss function to focus more on the minority classes. In the context of computer vision applications, over-sampling methods introduce large training costs and make the model prone to overfit the minority classes. Under-sampling methods discard important samples that are valuable for deep representation learning. Taking these issues of applying re-sampling methods on image classification tasks into consideration, our work focuses on designing a better re-weighting method to improve the accuracy of minority classes.

As minority classes are weakly represented with fewer samples [22, 121], re-weighting methods for imbalance problem penalize classifiers more seriously for misclassification of minority samples compared with those of majority samples. Re-weighting methods assign sample weights in inverse proportional to the class frequencies or the square root of class frequencies, which are proved efficient [39]. However, when applying on large real-world imbalanced datasets, re-weighting methods perform poorly [75]. One main reason might be that the minority classes are well represented by a small size of training data. Under this situation, resetting the weights in inverse proportional to the class frequencies (called overweighting) will decrease the overall performance. Thus, it is of great importance to find out the optimized weight for each class to achieve higher classification performance.

To better illustrate the previous problem of using class frequency, we run a DNN (ResNet-32) on a real-world medical image dataset HAM10000 (details in section 6.2.2) and measure the classification accuracy of each class. The class distribution of HAM10000 and classification

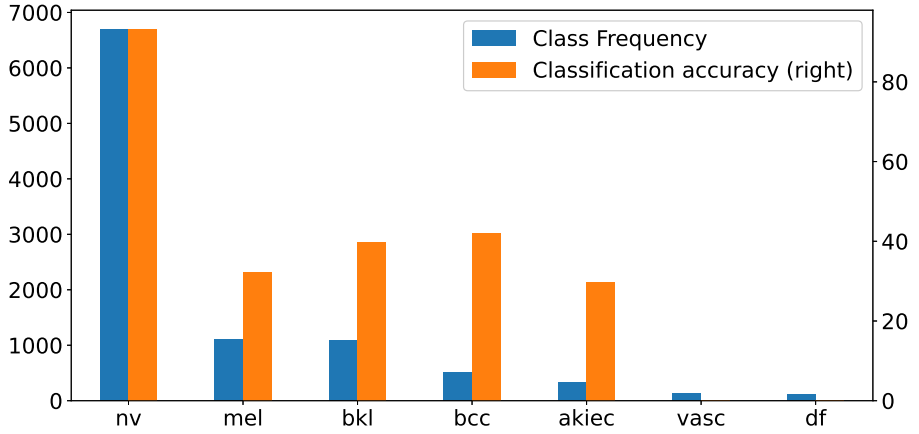

Figure 6.1: Class frequency of HAM10000 and Classification accuracy of each class of ResNet-32 trained on the HAM10000; left axis is class frequency (blue), right axis is classification accuracy

accuracy of each class are shown in Fig. 6.1. Class 'nv' dominates the dataset in terms of its sample size and it has an accuracy of 93.31% of the trained ResNet-32, while class 'df' has an accuracy of 0 due to its small sample size. It shows that the DNN model biases towards the majority class. However, class 'bcc' has higher classification accuracy than class 'mel' even though the class frequency of 'bcc' is smaller than that of 'mel'. If we assign weights in inverse proportional to the class frequencies, class 'mel' will have a smaller weight than class 'bcc', which will make the classification accuracy of class 'mel' worse. Thus, class frequency is not always a good option to set weights to alleviate the imbalance problem, which may lead to overweighting.

Instead of using class frequency, we aim to tackle the imbalance problem from the perspective of the classification hardness of classes during the training process. The concept of classification hardness has been previously used in self-paced ensemble (SPEnsemble), an under-sampling ensemble method for majority class samples [74] and focal loss, a sample-level weight assignment method [71]. Self-paced Ensemble learning (SPEnsemble) focuses on using classification hardness to undersample the majority samples by removing most of the easy majority samples. Focal loss individually increases weights for samples with large classification hardness, decreases weights for samples with small classification hardness. However, noise samples usually have larger classification hardness, which could lead to the poor performance of deep learning models.

In this work, we consider using class-level classification hardness to decrease the impact of noise samples rather than sample-level hardness. The propose loss function called Hardness Aware Dynamic (HAD) loss is introduced in Section 3.3. As describe in Section 3.3, HAD reweights each sample weight in the loss function dynamically by the classification hardness of its class during the training process of DNN. After each training epoch of a deep learning

model, we could measure the correctly classified probability for each sample. Then we define the classification hardness of this sample as its misclassification probability, which equals 1 minus its correctly classified probability. Next, we compute the average value of classification hardness of different classes. The average classification hardness values are used to update class weights following the rule that increase class weights with larger average classification hardness values and decrease class weights with smaller average classification hardness values. Extensive experiments have been conducted on two real-world imbalanced medical image datasets (Breast Cancer dataset, Skin Cancer MNIST:HAM10000) and two standard datasets (MNIST and CIFAR-10). The experimental results indicate that HAD loss can provide a significant improvement to the classification performance of recently proposed loss functions for training deep learning models.

In summary, the main contributions of this work are: (1) HAD loss achieves significant improvement compared with baselines over F1-score and G-mean on the real-world medical datasets. Especially, HAD loss improves 10.04% average precision compared with the best baseline, Focal loss, on HAM10000 dataset. (2) We prove the robustness of HAD loss over several datasets of different imbalance degrees. Overall, HAD loss on quantifying the classification hardness of each class and using it to update class weights dynamically can provide helpful guidelines for researchers working on imbalanced image classification tasks.

6.2 Imbalanced Image Datasets

To evaluate the effectiveness of our method, we conduct extensive experiments under both binary classification and multiple classification scenarios. The binary classification tasks includes three datasets: a breast cancer dataset [48], a binary subset of the skin cancer dataset and a binary subset of MNIST [67]. The multiple classification task includes two datasets: the CIFAR-10 dataset [62] and the skin cancer dataset [108].

6.2.1 Breast Cancer Dataset

Breast cancer is the most common cancer in women, and Invasive Ductal Carcinoma (IDC) is the most common subtype of breast cancer. There are 162 whole mount slide images of breast cancer scanned images and 277,524 IDC patches of size 50×50 are extracted. The breast cancer dataset is composed of 198,738 negative IDC negative samples and 78,786 IDC positive samples. We first split the dataset in stratifying with ground truth labels into two subsets with a ratio of 4:1. Then we generate four datasets with different IR:

- *Balanced training set and test set (IR=1)*we select 50,000 IDC positive samples and 50,000 IDC negative samples from the bigger subset as a balanced training set, select 10,000 IDC positive samples and 10,000 IDC negative samples from the smaller subset as a balanced test set;

- *Imbalanced training set(IR=5)* we decrease IDC positive patients of the balanced training set to 10,000, then combine them with the 50,000 IDC negative samples as an imbalance training set with IR=5;

- *Imbalanced training set(IR=10)* IDC positive patients of the balanced training set are decreased to 5,000 and are combined with 50,000 IDC negative patients to form an imbalanced training set with IR=10;

- *Imbalanced training set(IR=20)* Similar to previous training set generation, an imbalanced training set with IR=20 are composed of 2,500 IDC positive patients and 50,000 patients;

6.2.2 Skin Cancer MNIST: HAM10000 Dataset

HAM10000 (Human Against Machine with 10,000 training images) dataset, also called Skin Cancer MNIST, consists of 10,015 dermatoscopic images. HAM10000 includes a representative collection of 7 important diagnostic classes in the realm of pigmented lesions. It is composed of 6705 melanocytic nevi (nv) cases, 1113 melanoma (mel) cases, 1099 benign keratosis-like lesions (bkl) cases, 514 basal cell carcinoma (bcc), 327 cases diagnose with ctinic keratoses or intraepithelial carcinoma (akiec), 142 vascular lesions (vasc) and 115 dermatofibroma (df) cases.

Binary Classification Task To further validate the effectiveness of HAD Loss on another binary classification task with a natural IR, we select 'nv' class as majority class and 'bkl' class as minority class to form a binary class dataset with $IR = 6705/1099 \approx 6.10$.

Multiple Classification Task We conduct experiments on HAM10000 to measure the performance of HAD loss. As there are only around 100 samples from 'vasc' class and 'df' class, to keep enough samples in test set, we will split the original dataset into training set and test set at a ratio of 2:1.

6.2.3 MNIST

MNIST is the standard dataset used for handwritten recognition tasks. We extract all the images of '4' and '9' from the training set and the testing set. As there are 5,949 images of '9' and 5,842 images of '4' in the training set, 1,009 images of '9' and '982' images of '4' in the testing set. Thus, we set '9' as the majority class and set '4' as the minority class. We randomly under-sample the majority class to the same size of the minority size to generate a balanced training set and a balanced testing set. Similar to the subsets generation of Breast Cancer Dataset, we generate 4 subsets with different IR:

- *Imbalanced training set(IR=10)* we decrease class '4' in training set to 584, then combine them with the 5,842 class '9' images as an imbalance training set with IR=10;

- *Imbalanced training set(IR=20)* Class '4' in training set is decreased to 292. It is combined with 5,842 class '9' images to form an imbalanced training set with IR=20;
- *Imbalanced training set(IR=40)* Similar to previous training set generation, an imbalanced training set with IR=20 are composed of 146 class '4' images and 5,842 class '9' images ;
- *Imbalanced training set(IR=80)* Similarly , an imbalanced training set with IR=80 are composed of of 73 class '4' images and 5,842 class '9' images.

6.2.4 CIFAR-10

CIFAR-10 is a class-balanced dataset with 10 classes used for image classification tasks. To compare our method with other methods, we generate a class-imbalanced CIFAR-10 by reducing the training size of each class according to an exponential function. For multiple classification tasks, the IR of a imbalanced dataset is defined as the size of largest class divided by that of the smallest class. Here, we set IR of long-tailed CIFAR-10 as 100. We use the original balanced test set to test the performance of the algorithms.

6.3 Experiments

In this section, we first introduce the baseline methods. Following that, we describe the implementation details of our experiments and the evaluation metrics under imbalance situations. At last, we show experimental results on both binary classification and multiple classification tasks.

6.3.1 Baseline Methods

We compare HAD loss with several state-of-the-art approaches that have been used to deal with the imbalance problem: (1) Cross-Entropy Loss (CE): each sample has the same weight; we use sigmoid cross-entropy for binary classification tasks, softmax cross-entropy for multiclass classification tasks. (2) Inverse-weight (IW): we set the sample weight inversely proportionally to its class frequency. (3) Random Over-Sampling (ROS): we select each sample with probability inversely proportional to its class frequency. (4) SMOTE : we use SMOTE to synthetic more minority samples to create a balanced dataset. (5) Focal loss (Focal): assign higher weights to relative hard samples to improve the minority classification performance. (6) Class-balanced loss (CB): reweight each sample by the inverse of the effective number of samples for its class, defined as $(1-\beta^{n_i})/(1-\beta)$ (7) Class-wise Difficulty-Balance loss (CDB) dynamically assign class weight by the class-wise difficulty measured during the training process.

6.3.2 Experiments Setup

Pytorch [87] is used to implement and train all the neural networks on 2 NVIDIA 1080Ti GPUs. We detail the experimental settings and the definitions of evaluation metrics in the following parts.

Binary classification task. For binary classification tasks, we conduct all experiments using ResNet-18 [41] for 60 epochs, and batch size is set as 32. Stochastic Gradient Descent (SGD) is used as the optimizer. Its momentum is 0.9 and weight decay is 0.0005. The initial learning rate is set to 0.0001 and it will decay by 0.1 after 10, 20, 30, 40, and 50 epochs.

Multiple classification task. For the multiple classification task on Skin Cancer MNIST, ResNet-32 is trained on the training set for 100 epochs using a batch size of 64. We use Stochastic Gradient Descent (SGD) with a momentum of 0.9 as the optimizer, and its weight decay is set to 0.0005. The initial learning rate is set to 0.1 and we adopt the 'linear warm-up' learning rate adjusting approach [34] in the first 5 epochs. After 60 and 80 epochs, the learning rate is decayed by 0.01.

Evaluation metrics. Under the imbalance scenario, we usually focus on the classification performance on the minority classes of the model, where accuracy is not appropriate. For example, if a data set includes 99% of majority class samples and only 1% minority class samples, a naive solution is to classify every sample into majority class, and the accuracy would be 99% and error rate would be 1%. It is pretty good at the first glance, however, accuracy fails to tell that there is no minority class sample correctly identified. Thus, to take the classification performance of minority class into consideration, we will compare classification models over evaluation metrics such as F1-score and G-mean.

F1-score is defined as the harmonic mean of the precision and recall:

$$F1 - score = \frac{2 \times Precision \times Recall}{Recall + Precision} \tag{6.1}$$

G-mean is defined as the geometric mean of precision and recall :

$$G - mean = \sqrt{Recall \times Precision} \tag{6.2}$$

6.3.3 Experimental Results on Binary Classification Tasks

We conduct an experiment to evaluate the effectiveness of HAD Loss to improve the performance of commonly used loss functions, such as Class-Balanced loss and Sigmoid Cross-Entropy loss which is also known as Binary Cross-Entropy (BCE) Loss. Firstly, we train ResNet-18 on the imbalance training set (IR=5) using BCE loss and HAD loss. Then, we train ResNet-18 using CB loss ($\beta \in \{0.99, 0.999, 0.9999\}$) and HAD loss with setting initial class

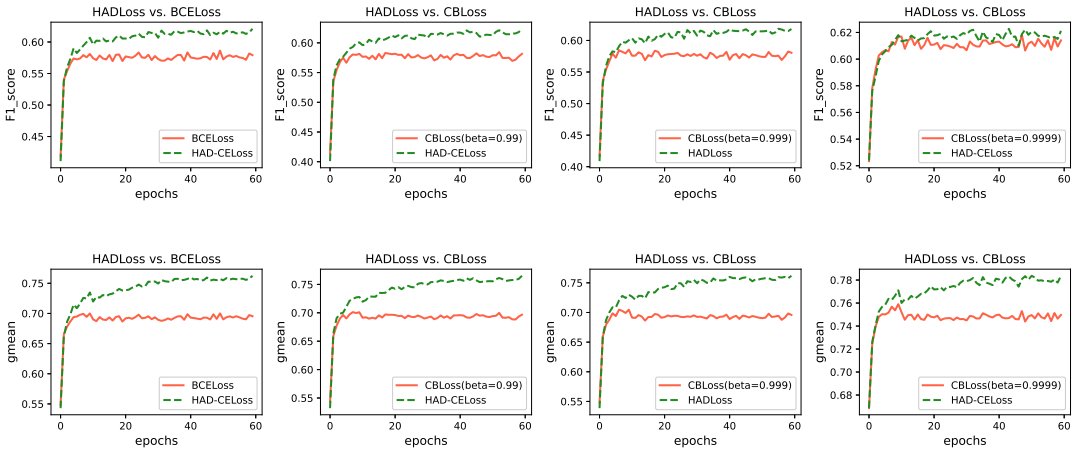

Figure 6.2: (a) F1-score training curves of ResNet-18 on breast cancer imbalanced subset (IR=5) (b) G-mean training curves of ResNet-18 on breast cancer imbalanced subset (IR=5)

weights computed from CB loss. The training curves over F1-score and G-mean of ResNet-18 by using different loss function on the Breast Cancer test set is shown in Fig. 6.2. From these training curves, we can indicate that:

- HAD curve is higher than BCE curve both in the first sub-figure in Fig. 6.2(a) and the that in Fig. 6.2(b), which proves HAD outperforms BCE both over F1-score and G-mean;
- When $\beta = 0.9999$, CB loss gets the best f1-score and g-mean, according to this, we select CB loss with setting $\beta = 0.9999$ in the following experiments;
- Both in the last three sub-figures of F1-score training curve and G-mean training curve, HAD always performs better than CB, which proves that HAD could improve the performance of CB over different initial weights.

Robustness of HAD loss over imbalance ratio. In this part, we test the robustness of HAD over different degrees of data imbalance. We retrain ResNet-18 with different commonly used loss functions on three imbalanced training sets ($IR \in \{5, 10, 20\}$) of the breast cancer dataset. After the training, we measure the G-mean of each trained model. To better show the robustness of our method, we compare G-mean of our method with that of Cross-Entropy loss (CE), Class-Balance loss (CB), Focal loss (Focal), and Class-wise Difficulty-Balance loss (CDB) in Fig. 6.3. As can be seen from Fig. 6.3 that:

- With the increasing imbalance degree from 5 to 20, G-mean of CE decreases from 0.69 to 0.32, which indicates that the deep neural network performs worse when trained on the dataset with higher imbalance ratios;
- HAD loss performs better than other commonly used loss functions over different IR, which illustrates the robustness of HAD loss over IR.

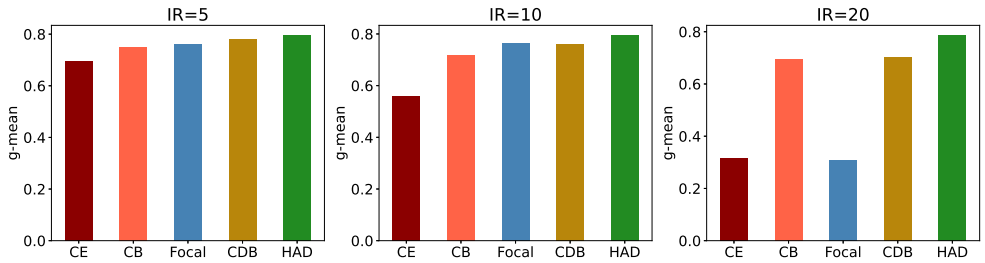

Figure 6.3: G-mean of ResNet-18 with using HAD loss and commonly used loss functions on breast cancer subsets under different imbalance degrees, i.e., IR=5, IR=10, IR=20

- Compared with best baseline CDB, HAD respectively improves 1.92% when IR=5, 4.33% when IR=10, and 11.86% when IR=20, which indicates HAD can bring more improvement when the dataset is of higher IR.

To further validate the robustness of HAD loss over IR, a another experiment is conducted to evaluate the effectiveness of HAD Loss over the commonly used loss CE loss, and state-of-the-art loss functions, such as, CB loss, Focal loss and CDB loss. We train LeNet-5 on the binary subsets of MNIST with different IR. The experimental results are listed in Table 6.1 and Table 6.2.

Table 6.1: F1-score of LeNet-5 trained by four different binary MNIST subsets composed of class '4' and class '9' using different loss functions

Method	CE	CB	Focal	CDB	HAD
IR=10	0.953	0.970	0.958	0.898	**0.976**
IR=20	0.899	0.955	0.000	0.784	**0.963**
IR=40	0.745	0.898	0.000	0.792	**0.935**
IR=80	0.000	0.463	0.000	0.000	**0.865**

Table 6.2: G-mean of LeNet-5 trained by four different binary MNIST subsets composed of class '4' and class '9' using different loss functions

Method	CE	CB	Focal	CDB	HAD
IR=10	0.954	0.970	0.958	0.902	**0.976**
IR=20	0.904	0.955	0.000	0.803	**0.963**
IR=40	0.771	0.902	0.000	0.810	**0.933**
IR=80	0.000	0.549	0.000	0.000	**0.857**

- Under different IR, HAD always outperforms all other loss functions both over F1-score and G-mean. The larger is IR, the larger improvement could HAD get over other loss functions;

- The performance decreases as the imbalance degree increases, whereas the decreasing speed of our method HAD is the slowest;

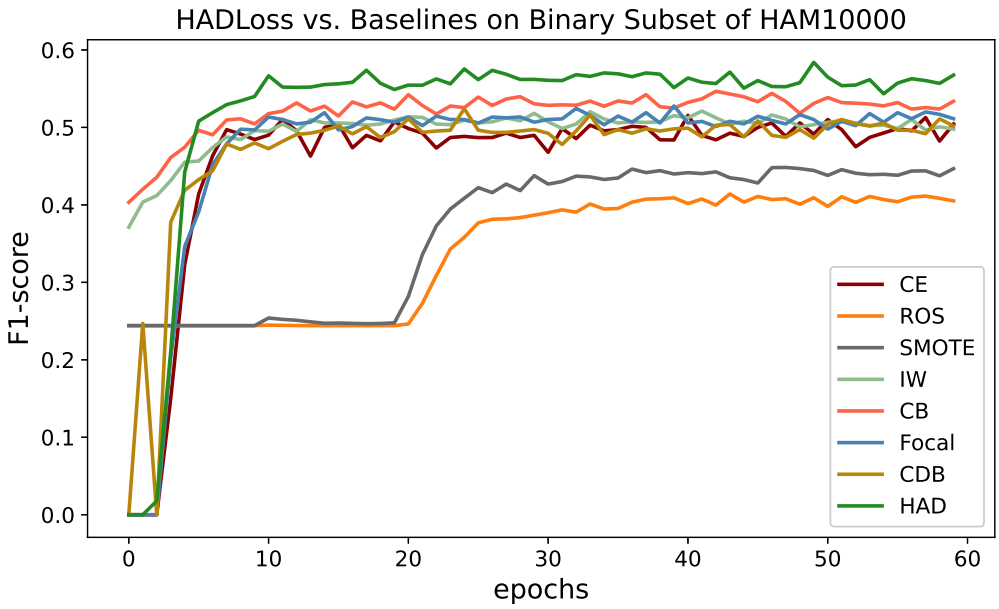

Figure 6.4: Training curves of f1-score of ResNet-18 on binary subsets of HAM10000

- When IR=20,40,80 the F1-score and the g-mean of Focal loss drop to 0, which indicates that Focal loss performs poorly on highly imbalanced binary classification tasks;

- When IR=80, the F1-score and g-mean of CE loss are down to 0, and it is same to Focal loss and CDB loss. However, only CB loss and HAD loss get evaluations lager than 0. Compared with CB loss, our method HAD loss improves F1-score from 0.463 to 0.865 and g-mean from 0.549 to 0.857.

6.3.4 Experimental Results on Multiple Classification Tasks

In this subsection, we first train ResNet-18 on the binary subset of HAM10000 with its natural imbalance ratio to further validate the effectiveness of our method. Then, we compare the classification performance of ResNet-32 with using HAD loss and baseline methods on the original HAM10000 dataset, which is composed of 7 different classes. For the binary classification task, we first randomly split binary subset into training set and testing set at a ratio of 4:1. Secondly, we train ResNet-18 with using different loss functions (i.e., CE loss, CB loss, Focal loss, CDB loss and HAD loss), Inverse-Weight (IW) and resampling methods (i.e., Random-Over-Sampling (ROS) and SMOTE). We use F1-score as the evaluation metric, and show the F1-score curves of ResNet-18 through the training process in Fig. 6.4.

As seen from Fig. 6.4, we can observe that:

- Though resampling methods (ROS and SMOTE) have relatively good F1-scores at the beginning, their F1-score only starts to increase after 20 epochs. SMOTE performs better

than ROS at the end, but they both perform worse than re-weighting methods (IW and CB) and loss functions such as (Focal loss, CDB loss and HAD loss);

- Re-weighting methods (IW and CB) have higher F1-score than resampling methods from beginning to end, which indicates the re-weighting methods are better solutions in this task;

- HAD loss outperforms all other baseline methods and it improves 6.34% in terms of F1-score compared with the best baseline method, i.e., CB loss.

For the multiple classification task, the original Skin Cancer MNIST: HAM10000 is split into training set and testing set at a ratio of 2:1. Then, we train ResNet-32 with using different loss functions, such as Softmax CE loss, CB loss, Focal loss, CDB loss and HAD loss ($\lambda = 0.01$). We list the results of ResNet-32 on HAM10000 in Table 6.4. From Table 6.4, we can see that HAD loss outperforms all other baseline methods over F1-score and G-mean, which indicates that our method also works on multiple classification tasks.

Another experiment was implemented on the long-tailed CIFAR-10 dataset to compare HAD loss with other baseline methods. ResNet-32 is trained on the training set for 100 epochs using a batch size of 64. All other experiment setup is same as the multiple classification task. We use F1-score and G-mean to evaluate the experimental results, which are listed in Table 6.3.

- HAD loss outperforms all baseline methods both over F1-score and G-mean;

- All methods except CDB loss get improvements over CE loss, which indicates that CDB loss may not suitable for this task;

- Meanwhile Meanwhile, compared with the best baseline method IW, HAD loss improves F1-score from 0.710 to 0.753 and improves G-mean from 0.829 to 0.855.

Table 6.3: F1-score and G-mean trained on the long-tailed CIFAR-10 under the IR=100

Method	CE	ROS	SMOTE	IW	CB	Focal	CDB	HAD
F1-score	0.677	0.707	0.689	0.707	0.710	0.704	0.406	**0.753**
G-mean	0.816	0.827	0.817	0.828	0.829	0.829	0.654	**0.855**

Table 6.4: F1-score and G-mean of ResNet-32 on HAM10000

Method	CE	IW	CB	Focal	CDB	HAD
F1-score	0.707	0.595	0.609	0.713	0.591	**0.722**
G-mean	0.726	0.705	0.710	0.729	0.674	**0.756**

At last, we compare our method the best baseline method Focal loss over the accuracy of each class in Fig. 6.5. As the Fig.6.5 shows, compared with Focal loss, HAD loss improves accuracies of five classes : class 'mel' from 33.52% to 35.15%, class 'bkl' from 42.97% to

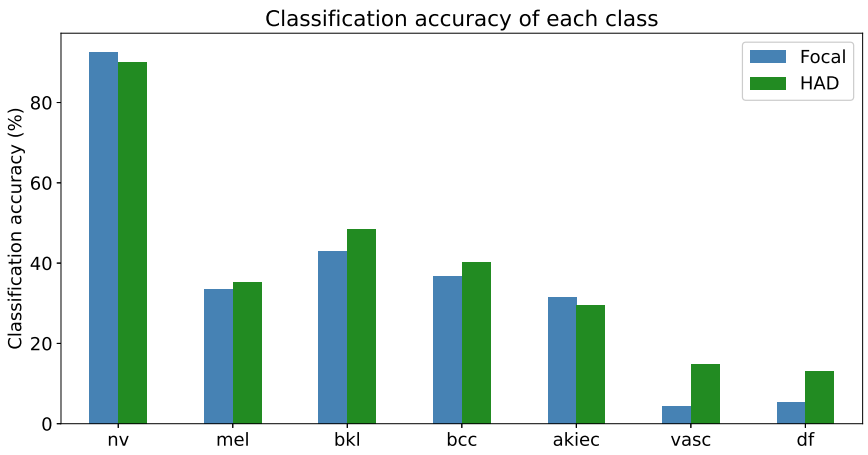

Figure 6.5: Class-wise classification accuracy comparison between focal loss and HAD loss on HAM10000

48.48%, class 'bcc' from 36.69% to 40.23%, class 'vasc' from 4.26 to 14.89% and class 'df' from 5.26 to 13.16%. Compared with the accuracy improvement of the five classes, HAD only makes small compromises in accuracies of the other classes: class 'nv' from 93.31% to 89.92% and class 'akiec' from 31.48% to 29.63%. In all, HAD loss improves average precision from 35.26% to 38.80% compared with Focal loss.

6.4 Summary

In this chapter, we conduct extensive experiments to test the performance of HAD on MNIST, CIFAR10 and two medical image datasets. The experimental results indicate that HAD can provide a significant improvement to the classification performance of state-of-the-art methods. Moreover, HAD significantly improves the classification accuracies of minority classes while making a small compromise of majority class accuracies. In summary, we believe that we have proposed a novel paradigm of leveraging classification hardness into the imbalanced image classification when using DNNs.

Chapter 7

Conclusion and Future Work

This chapter summaries this dissertation and provides an outlook for the future work.

7.1 Conclusion

This dissertation focuses on solving the class imbalance problem which is common to see in the real world datasets. As the conventional methods are proposed based on the assumption that the datasets are statistically balanced, the class imbalance problem could bias the conventional methods to the majority class, in other words, the methods have weak performance on the minority class. Thus, such methods are not helpful in classification tasks on a lot of real-world applications, such as fraud detection and disease diagnose, since the minority class is of higher interest in these applications. Therefore, it is very important to adopt additional methods to tackle the class imbalance problem for building better prediction models. Our objectives are to understand feature difference between the majority class and the minority class, to propose novel solutions for the class imbalance problem. We focus on the topic of imbalance learning and evaluate our proposed methods on several medical datasets, which are intrinsically imbalanced.

Accurate risk prediction models could help physicians and hospitals make preparation prior to operation or refer the challenging cases to more experienced centers. After data preprocessing, 3292 cases treated by PCNL from 2012 to 2019 are collected. With 19.78% patients who have different kinds of postoperative complications, the class distribution is highly imbalanced, which makes the makes the prediction of complications an imbalance problem. However, traditional postoperative complication prediction models of PCNL, such as S.T.O.N.E. nephrolithometry, CORES and Guy's score system, take no consideration of the class imbalance distribution problem. Furthermore, traditional models are designed to predict the kidney stone status and do not consider complication related features, which degrade their prediction performance on complication prediction. To this end, we compare patients' demographic characteristics, disease history, laboratory test variables, preoperative variables and operation outcome between complication free patients and patients with complications. Through the analysis, we identify

features statistically associated with the postoperative complications. The analysis results represents that female patients, young patients, obese patients has higher complication rate after operation; urine culture is the only laboratory test variable that statistically significant associate with the complications. Furthermore, high level complication patients are likely to have larger stone size, loss more blood, experience longer operation time, and stay longer in hospital. To achieve better classification performance, we propose a sampling-based method named SMOTE-XGBoost, which combines the sample synthetic method (SMOTE) and the strong classifier (XGBoost). SMOTE-XGBoost is implemented to build a postoperative complication model to deal with the class imbalance problem. Experimental results verify the proposed method outperforms classic machine learning methods and S.T.O.N.E. nephrolithometry, a traditional PCNL model. More features are merged into the proposed sampling-based method and further improve the prediction performance of the proposed postoperative complication method.

After analyzing the advantages and disadvantages of the existing machine learning methods under class imbalance, we propose a ensemble learning approach called Multiple bAlance Subset Stacking (MASS). MASS first cuts the majority class into multiple subsets which have the same size of the minority set, and combines each majority subset with the minority set as one balance subsets. We name this approach as Multiple Balance Subsets Constructing Strategy, which overcomes the problem of information loss because it does not discard any majority sample. These generated balanced subsets are used to train base classifiers. Then the original dataset are feed to all the trained base classifiers and their outcome are used to generate the stacking dataset. One stack model is trained by the staking dataset to get the optimal weights for the base classifiers. As the stacking dataset keeps the same labels as the original dataset, which could avoid the overfitting problem of base classifiers. Finally, we can get an ensembled strong model based on the trained base classifiers. Extensive experimental results on three medical datasets show that MASS outperforms other state-of-the-art methods over AUC, F1-score and MCC. We also prove that MASS is robust over using different base classifiers. Additionally, with the increasing size of datasets, it is of great importance to reduce the training time cost. Thus, we design a parallel version MASS. The speedup analysis proves that Parallel MASS could reduce huge training time cost when applied on large datasets.

In the third study, we propose a re-weighting method Hardness Aware Dynamic loss for imbalanced image classification when using DNNs. We first introduce the problem of implementing resampling methods in image classification tasks. Then we demonstrate the issues of re-weighting strategy using class frequencies through the classification results on one medical image dataset (HAM-10000). To come up a novel strategy, we introduce the definition of classification hardness, which is the average of misclassification possibilities. After each training epoch of DNN, we compute the classification hardness of each class. In the next training epoch, we will increase the class weights of classes that have large classification hardness values and vice versa. In this way, HAD reweights each sample weight in the loss function dynamically during the training process of DNNs. The experimental results indicate

that HAD can provide a significant improvement to the classification performance of state-of-the-art methods. Moreover, HAD significantly improves the classification accuracies of minority classes while making a small compromise of majority class accuracies. In summary, we believe that we have proposed a novel paradigm of leveraging classification hardness into the imbalanced image classification when using DNNs.

7.2 Future Work

Although our methods have achieved promising performance compared with baseline methods, there are some potential directions to improve them in the future. Additionally, our proposed algorithms are general methods that can be used for all class imbalance tasks. Nevertheless, they might not suitable for the specified application, which need learn more domain knowledge and find the unique property of the task. At last, we will discuss the challenge of within-class imbalance.

Firstly, the potential directions of the three proposed methods are listed as following:

- **The sampling-based method SMOTE-XGBoost** This method is proposed for better predicting the postoperative complication, we could collect more cases from different hospitals in the future. Although the proposed method outperforms S.T.O.N.E. nephrolithometry, KNN, RF, SVM, MLP and XGBoost, there is still a quite distance from an AUC of 0.7077 to perfect classification. With lager sample size of PCNL patients, we could build better prediction models with higher AUC. On the other hand, we formed the postoperative complication as a binary classification problem due to the lack of data and imbalanced distribution. With lager sample size of PCNL patients, we could build a multi-class complication prediction model with higher accuracy. Lastly, the structuring of the radiological report provides benefits to improve medical practice and diagnoses. We will use deep learning methods to get more precise report of kidney stones, which could be fused in our proposed prediction model to achieve better performance.

- **The ensemble method Multiple Balanced Subsets Stacking (MASS)** This method is designed for the imbalanced structured datasets, we could take sampling ratio and cost matrix into consideration. Sampling methods (data-level) and cost-sensitive methods (algorithm-level) are two major categories of imbalance learning as each category has its own superiority over the other one[79, 79]. For sampling methods, most of the study implement them to balance the training sets and few of them discuss the effect of sampling ratio. For cost-sensitive methods, most of the work assign misclassification cost by the ratio between the majority class and minority class while the rest tune the misclassification cost as a free parameter. Hence, there are three potential works that are beneficial to further improve MASS: 1) adjusting the sampling ratio in during the subsets generation step; 2) tuning the misclassification cost of the stacking model; 3)adjusting

the sampling ratio and misclassication cost at the same time. We could verify these ideas by conducting experiments on more medical datasets with different imbalance ratios.

- **The re-weighting method Hardness Aware Dynamic (HAD) loss function** HAD loss is proposed for imbalance image classification tasks, we could combine it with Generative Adversarial Network (GAN). GAN has been proved very effective dealing with the image classification tasks under class imbalance [77]. Therefore, GAN could be used as the data augmentation tool to generate images from the minority class, which could alleviate the imbalance problem in a certain degree. HAD loss could still be used in the training process of the DNNs classifiers. Additionally, we would also implement HAD loss on the radiological reports of kidney stone patients to build a more accurate postoperative complication prediction model, which is mentioned in the future work of our first work. Lastly, HAD loss could be verified on video datasets to show its generalization ability.

As mentioned previously, all those three proposed methods are general ones. When it comes to specific applications, domain knowledge are needed to achieve best performance in practice. For instance, we could work with domain expert to find unique properties from the applications, then merge them into the algorithms to improve the performance and minimize the misclassification cost. For example, we usually set the misclassification cost reversely proportional to the class frequencies, which is not always a good solution as we demonstrate in Chapter 6. Therefore, when the proposed methods are applied on real-world tasks, we should take the advantages of their unique properties.

Lastly, this thesis mainly focuses on the between-class imbalance whereas the within-class imbalance is also very challenging. For within-class imbalance, there are small data disjuncts in the same class, which are even difficult to be observed. Severe within-class imbalance will jeopardize the classification performance. Usually, complex and high-dimensional data sets possess both within-class imbalance and between-class imbalance. Therefore, we could work on within-class imbalance to further improve the performance of the imbalance learning methods in the future.

Bibliography

[1] Arash Akhavein, Carl Henriksen, Jamil Syed, and Vincent G Bird. "Prediction of single procedure success rate using STONE nephrolithometry surgical classification system with strict criteria for surgical outcome". In: *Urology* 85.1 (2015), pp. 69–73.

[2] Tolga Akman, Murat Binbay, Erhan Sari, et al. "Factors affecting bleeding during percutaneous nephrolithotomy: single surgeon experience". In: *Journal of endourology* 25.2 (2011), pp. 327–333.

[3] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. "An improved algorithm for neural network classification of imbalanced training sets". In: *IEEE Transactions on Neural Networks* 4.6 (1993), pp. 962–969.

[4] Shin Ando and Chun Yuan Huang. "Deep over-sampling framework for classifying imbalanced data". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2017, pp. 770–785.

[5] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1207–1216.

[6] Ricardo Barandela, Rosa Maria Valdovinos, and José Salvador Sánchez. "New applications of ensembles of classifiers". In: *Pattern Analysis & Applications* 6.3 (2003), pp. 245–256.

[7] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 20–29.

[8] Richard Bauder and Taghi Khoshgoftaar. "Medicare fraud detection using random forest with class imbalanced big data". In: *2018 IEEE international conference on information reuse and integration (IRI)*. IEEE. 2018, pp. 80–87.

[9] Sakyajit Bhattacharya, Vaibhav Rajan, and Harsh Shrivastava. "ICU mortality prediction: a classification algorithm for imbalanced datasets". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[10] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[11] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[12] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[13] Nitesh V Chawla. "Data mining for imbalanced datasets: An overview". In: *Data mining and knowledge discovery handbook* (2009), pp. 875–886.

[14] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[15] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. "SMOTEBoost: Improving prediction of the minority class in boosting". In: *European conference on principles of data mining and knowledge discovery*. Springer. 2003, pp. 107–119.

[16] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 1–6.

[17] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[18] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), p. 6.

[19] David A Cieslak, Nitesh V Chawla, and Aaron Striegel. "Combating imbalance in network intrusion datasets". In: *IEEE International Conference on Granular Computing*. 2006, pp. 732–737.

[20] Pierre-Alain Clavien, Juan R Sanabria, and Steven M Strasberg. "Proposed classification of complications of surgery with examples of utility in cholecystectomy." In: *Surgery* 111.5 (1992), pp. 518–526.

[21] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[22] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. "Class-balanced loss based on effective number of samples". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9268–9277.

[23] Virginia Dignum. *Ethics in artificial intelligence: introduction to the special issue*. 2018.

[24] Daniel Dindo, Nicolas Demartines, and Pierre-Alain Clavien. "Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey". In: *Annals of surgery* 240.2 (2004), p. 205.

[25] Robert H Dolin, Liora Alschuler, Sandy Boyer, et al. "HL7 clinical document architecture, release 2". In: *Journal of the American Medical Informatics Association* 13.1 (2006), pp. 30–39.

[26] Qi Dong, Shaogang Gong, and Xiatian Zhu. "Imbalanced deep learning by minority class incremental rectification". In: *IEEE transactions on pattern analysis and machine intelligence* 41.6 (2018), pp. 1367–1381.

[27] Scott Doyle, James Monaco, Michael Feldman, John Tomaszewski, and Anant Madabhushi. "An active learning based classification strategy for the minority class problem: application to histopathology annotation". In: *BMC bioinformatics* 12.1 (2011), p. 424.

[28] Ahmed R El-Nahas, Ahmed A Shokeir, Ahmed M El-Assmy, et al. "Post-percutaneous nephrolithotomy extensive hemorrhage: a study of risk factors". In: *The Journal of urology* 177.2 (2007), pp. 576–579.

[29] Amir Fallahi and Shahram Jafari. "An expert system for detection of breast cancer using data preprocessing and bayesian network". In: *International Journal of Advanced Science and Technology* 34 (2011), pp. 65–70.

[30] Tom Fawcett and Foster Provost. "Adaptive fraud detection". In: *Data mining and knowledge discovery* 1.3 (1997), pp. 291–316.

[31] I Fernström and B Johansson. "Percutaneous pyelolithotomy: a new extraction technique". In: *Scandinavian journal of urology and nephrology* 10.3 (1976), pp. 257–259.

[32] Andres Folleco, Taghi M Khoshgoftaar, and Amri Napolitano. “Comparison of four performance metrics for evaluating sampling techniques for low quality class-imbalanced data”. In: *2008 Seventh International Conference on Machine Learning and Applications*. IEEE. 2008, pp. 153–158.

[33] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. “EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling”. In: *Pattern Recognition* 46.12 (2013), pp. 3460–3471.

[34] Priya Goyal, Piotr Dollár, Ross Girshick, et al. “Accurate, large minibatch sgd: Training imagenet in 1 hour”. In: *arXiv preprint arXiv:1706.02677* (2017).

[35] Rong Gu, Shiqing Fan, Qiu Hu, Chunfeng Yuan, and Yihua Huang. “Parallelizing Machine Learning Optimization Algorithms on Distributed Data-Parallel Platforms with Parameter Server”. In: *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE. 2018, pp. 126–133.

[36] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning”. In: *International conference on intelligent computing*. Springer. 2005, pp. 878–887.

[37] Simon Haykin and N Network. “A comprehensive foundation”. In: *Neural networks* 2.2004 (2004), p. 41.

[38] Haibo He and Edwardo A Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[39] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, pp. 1322–1328.

[40] Jinyuan He, Le Sun, Jia Rong, Hua Wang, and Yanchun Zhang. “A pyramid-like model for heartbeat classification from ECG recordings”. In: *PloS one* 13.11 (2018), e0206593.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 770–778.

[42] Jeff Hecht. “The future of electronic health records.” In: *Nature* 573.7775 (2019), S114–S114.

[43] Christian Heimes. “Multiprocessing”. In: *http://code.google.com/p/python-multiprocessing* (2009 (accessed July 15, 2020)).

[44] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. “Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015”. In: *ONC data brief* 35 (2016), pp. 1–9.

[45] Paulina Hensman and David Masko. “The impact of imbalanced training data for convolutional neural networks”. In: *Degree Project in Computer Science, KTH Royal Institute of Technology* (2015).

[46] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. “Batch mode active learning and its application to medical image classification”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 417–424.

[47] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. “Learning deep representation for imbalanced classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 5375–5384.

[48] Andrew Janowczyk and Anant Madabhushi. “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases”. In: *Journal of pathology informatics* 7 (2016).

[49] Fei Jiang, Yong Jiang, Hui Zhi, et al. “Artificial intelligence in healthcare: past, present and future”. In: *Stroke and vascular neurology* 2.4 (2017).

[50] Kehua Jiang, Fa Sun, Jianguo Zhu, et al. “Evaluation of three stone-scoring systems for predicting SFR and complications after percutaneous nephrolithotomy: a systematic review and meta-analysis”. In: *BMC urology* 19.1 (2019), p. 57.

[51] Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3 (2016), p. 160035.

[52] Alistair EW Johnson, Nic Dunkley, Louis Mayaud, et al. “Patient specific predictions in the intensive care unit using a Bayesian ensemble”. In: *2012 Computing in Cardiology*. IEEE. 2012, pp. 249–252.

[53] AK Kable, RW Gibberd, and AD Spigelman. “Adverse events in surgical patients in Australia”. In: *International Journal for Quality in Health Care* 14.4 (2002), pp. 269–276.

[54] Nadia A Khan, Hude Quan, Jennifer M Bugar, et al. “Association of postoperative complications with hospital costs and length of stay in a tertiary care center”. In: *Journal of general internal medicine* 21.2 (2006), pp. 177–180.

[55] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. “Cost-sensitive learning of deep feature representations from imbalanced data”. In: *IEEE transactions on neural networks and learning systems* 29.8 (2017), pp. 3573–3587.

[56] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, et al. “An integrated machine learning approach to stroke prediction”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, pp. 183–192.

[57] Jin Kyu Kim, Qirong Ho, Seunghak Lee, et al. “STRADS: a distributed framework for scheduled model parallel machine learning”. In: *Proceedings of the Eleventh European Conference on Computer Systems*. 2016, pp. 1–16.

[58] Omer Koras, Ibrahim Halil Bozkurt, Tarik Yonguc, et al. “Risk factors for postoperative infectious complications following percutaneous nephrolithotomy: a prospective clinical study”. In: *Urolithiasis* 43.1 (2015), pp. 55–60.

[59] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, et al. “Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy”. In: *Ophthalmology* 125.8 (2018), pp. 1264–1272.

[60] Bartosz Krawczyk. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232.

[61] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. “Cost-sensitive decision tree ensembles for effective imbalanced classification”. In: *Applied Soft Computing* 14 (2014), pp. 554–562.

[62] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning multiple layers of features from tiny images”. In: (2009).

[63] Miroslav Kubat, Stan Matwin, et al. “Addressing the curse of imbalanced training sets: one-sided selection”. In: *International Conference on Machine Learning (ICML)*. Vol. 97. Citeseer. 1997, pp. 179–186.

[64] Kevin Labadie, Zhamshid Okhunov, Arash Akhavein, et al. “Evaluation and comparison of urolithiasis scoring systems used in percutaneous kidney stone surgery”. In: *The Journal of urology* 193.1 (2015), pp. 154–159.

[65] Gaston Labate, Pranjal Modi, Anthony Timoney, et al. “The percutaneous nephrolithotomy global study: classification of complications”. In: *Journal of endourology* 25.8 (2011), pp. 1275–1280.

[66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[67] Yann LeCun, Corinna Cortes, and CJ Burges. “MNIST handwritten digit database”. In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[68] Hansang Lee, Minseok Park, and Junmo Kim. “Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3713–3717.

[69] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 559–563.

[70] Der-Chiang Li, Chiao-Wen Liu, and Susan C Hu. “A learning method for the class imbalance problem with medical data sets”. In: *Computers in biology and medicine* 40.5 (2010), pp. 509–518.

[71] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. “Focal Loss for Dense Object Detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[72] Charles X Ling and Victor S Sheng. “Cost-sensitive learning and the class imbalance problem”. In: *Encyclopedia of machine learning* 2011 (2008), pp. 231–235.

[73] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. “Exploratory undersampling for class-imbalance learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 539–550.

[74] Zhining Liu, Wei Cao, Zhifeng Gao, et al. “Self-paced Ensemble for Highly Imbalanced Massive Data Classification”. In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020.

[75] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, et al. “Exploring the limits of weakly supervised pretraining”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 181–196.

[76] Inderjeet Mani and I Zhang. “kNN approach to unbalanced data distributions: a case study involving information extraction”. In: *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. 2003.

[77] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. “Bagan: Data augmentation with balancing gan”. In: *arXiv preprint arXiv:1803.09655* (2018).

[78] Brian W Matthews. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[79] Kate McCarthy, Bibi Zabar, and Gary Weiss. “Does cost-sensitive learning beat sampling for classifying rare classes?” In: *Proceedings of the 1st international workshop on Utility-based data mining*. 2005, pp. 69–77.

[80] MM McKerns and M Aivazis. “Pathos: a framework for heterogeneous computing”. In: *See http://trac. mystic. cacr. caltech. edu/project/pathos* (2010).

[81] Tim Menzies, Jeremy Greenwald, and Art Frank. “Data mining static code attributes to learn defect predictors”. In: *IEEE transactions on software engineering* 33.1 (2006), pp. 2–13.

[82] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. “Generative Adversarial Minority Oversampling”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2019.

[83] Shivaramakrishnan Narayan, Martin Gagné, and Reihaneh Safavi-Naini. “Privacy preserving EHR system using attribute-based infrastructure”. In: *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*. 2010, pp. 47–52.

[84] Ziad Obermeyer and Ezekiel J Emanuel. “Predicting the future—big data, machine learning, and clinical medicine”. In: *The New England journal of medicine* 375.13 (2016), p. 1216.

[85] Zhamshid Okhunov, Justin I Friedlander, Arvin K George, et al. “STONE nephrolithometry: novel surgical classification system for kidney calculi”. In: *Urology* 81.6 (2013), pp. 1154–1160.

[86] Daniel Olvera-Posada, Thomas Tailly, Husain Alenezi, et al. “Risk factors for postoperative complications of percutaneous nephrolithotomy at a tertiary referral center”. In: *The Journal of urology* 194.6 (2015), pp. 1646–1651.

[87] Adam Paszke, Sam Gross, Francisco Massa, et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 8026–8037.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python ”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[89] Samira Pouyanfar, Yudong Tao, Anup Mohan, et al. “Dynamic sampling in convolutional neural networks for imbalanced data classification”. In: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE. 2018, pp. 112–117.

[90] Talayeh Razzaghi, Oleg Roderick, Ilya Safro, and Nick Marko. “Fast imbalanced classification of healthcare data with missing values”. In: *2015 18th International Conference on Information Fusion (Fusion)*. IEEE. 2015, pp. 774–781.

[91] Douglas M Rocha, Lourdes M Brasil, Janice M Lamas, Glécia VS Luz, and Simônides S Bacelar. “Evidence of the benefits, advantages and potentialities of the structured radiological report: An integrative review”. In: *Artificial intelligence in medicine* 102 (2020), p. 101770.

[92] Victoriano Romero, Haluk Akpinar, and Dean G Assimos. “Kidney stones: a global picture of prevalence, incidence, and associated risk factors”. In: *Reviews in urology* 12.2-3 (2010), e86.

[93] JJMCH De la Rosette, J Rioja Zuazu, P Tsakiris, et al. “Prognostic factors and percutaneous nephrolithotomy morbidity: a multivariate analysis of a contemporary series using the Clavien classification”. In: *The Journal of urology* 180.6 (2008), pp. 2489–2493.

[94] Charles D Scales Jr, Alexandria C Smith, Janet M Hanley, Christopher S Saigal, Urologic Diseases in America Project, et al. “Prevalence of kidney stones in the United States”. In: *European urology* 62.1 (2012), pp. 160–165.

[95] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. “RUSBoost: A hybrid approach to alleviating class imbalance”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40.1 (2009), pp. 185–197.

[96] Yachao Shao, Tao Zhao, xiaoning Wang, Xiaofeng Zou, and Xiaoming Fu. “Hardness aware dynamic loss on imbalanced image classification”. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2021, Under review.

[97] Yachao Shao, Tao Zhao, xiaoning Wang, Xiaofeng Zou, and Xiaoming Fu. “Multiple Balance Subsets Stacking for Imbalanced Healthcare Dataset”. In: *Proceedings of the IEEE International Conference on Parallel and Distributed Systems (ICPADS)*. 2020, pp. 300–307.

[98] Yachao Shao, xiaoning Wang, Xiaofeng Zou, and Xiaoming Fu. “Postoperative Complication Prediction of Percutaneous Nephrolithonomy via Imbalance Learning”. In: *Artificial Intelligence in Medicine* (2021), Under review.

[99] John Shawe-Taylor and Nello Cristianini. “Support vector machines”. In: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (2000), pp. 93–112.

[100] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. “Class-Wise Difficulty-Balanced Loss for Solving Class-Imbalance”. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2020.

[101] Arthur Smith, Timothy D Averch, Khaled Shahrour, et al. “A nephrolithometric nomogram to predict treatment success of percutaneous nephrolithotomy”. In: *The Journal of urology* 190.1 (2013), pp. 149–156.

[102] Toru Sugihara, Hideo Yasunaga, Hiromasa Horiguchi, et al. "Longer operative time is associated with higher risk of severe complications after percutaneous nephrolithotomy: Analysis of 1511 cases from a J apanese nationwide database". In: *International Journal of Urology* 20.12 (2013), pp. 1193–1198.

[103] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. "Cost-sensitive boosting for classification of imbalanced data". In: *Pattern Recognition* 40.12 (2007), pp. 3358–3378.

[104] Thomas O Tailly, Zhamshid Okhunov, Brandon R Nadeau, et al. "Multicenter external validation and comparison of stone scoring systems in predicting outcomes after percutaneous nephrolithotomy". In: *Journal of endourology* 30.5 (2016), pp. 594–601.

[105] Kay Thomas, Naomi C Smith, Nicholas Hegarty, and Jonathan M Glass. "The Guy's stone score—grading the complexity of percutaneous nephrolithotomy procedures". In: *Urology* 78.2 (2011), pp. 277–281.

[106] Paul Thottakkara, Tezcan Ozrazgat-Baslanti, Bradley B Hupf, et al. "Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications". In: *PloS one* 11.5 (2016), e0155705.

[107] Ivan Tomek et al. "Two modifications of CNN." In: (1976).

[108] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5 (2018), p. 180161.

[109] Mehmet Mazhar Utangac, Abdulkadir Tepeler, Mansur Daggulli, et al. "Comparison of scoring systems in pediatric mini-percutaneous nephrolithotomy". In: *Urology* 93 (2016), pp. 40–44.

[110] Rosa Maria Valdovinos and José Salvador Sánchez. "Class-dependant resampling for medical applications". In: *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*. IEEE. 2005, 6–pp.

[111] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. "Experimental perspectives on learning from imbalanced data". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 935–942.

[112] Simone L Vernez, Zhamshid Okhunov, Piruz Motamedinia, et al. "Nephrolithometric scoring systems to predict outcomes of percutaneous nephrolithotomy". In: *Reviews in urology* 18.1 (2016), p. 15.

[113] Philippe D Violette and John D Denstedt. "Standardizing the reporting of percutaneous nephrolithotomy complications". In: *Indian journal of urology: IJU: journal of the Urological Society of India* 30.1 (2014), p. 84.

[114] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3 (2020), pp. 261–272.

[115] Sofia Visa and Anca Ralescu. "Issues in mining imbalanced data sets-a review paper". In: *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*. Vol. 2005. sn. 2005, pp. 67–73.

[116] Haishuai Wang, Zhicheng Cui, Yixin Chen, et al. "Predicting hospital readmission via cost-sensitive deep learning". In: *IEEE/ACM transactions on computational biology and bioinformatics* 15.6 (2018), pp. 1968–1978.

[117] Jialei Wang, Peilin Zhao, and Steven CH Hoi. "Cost-sensitive online classification". In: *IEEE Transactions on Knowledge and Data Engineering* 26.10 (2013), pp. 2425–2438.

[118] Shoujin Wang, Wei Liu, Jia Wu, et al. "Training deep neural networks on imbalanced data sets". In: *2016 international joint conference on neural networks (IJCNN)*. IEEE. 2016, pp. 4368–4374.

[119] Shuo Wang and Xin Yao. “Diversity analysis on imbalanced data sets by using ensemble models”. In: *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE. 2009, pp. 324–331.

[120] Wenying Wang, Jingyuan Fan, Guifeng Huang, et al. “Prevalence of kidney stones in mainland China: A systematic review”. In: *Scientific reports* 7 (2017), p. 41630.

[121] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. “Learning to model the tail”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 7029–7039.

[122] Thomas G Weiser, Alex B Haynes, George Molina, et al. “Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes”. In: *The Lancet* 385 (2015), S11.

[123] Dennis L Wilson. “Asymptotic properties of nearest neighbor rules using edited data”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), pp. 408–421.

[124] David H Wolpert. “Stacked generalization”. In: *Neural networks* 5.2 (1992), pp. 241–259.

[125] Fei Wu, Xiao-Yuan Jing, Shiguang Shan, Wangmeng Zuo, and Jing-Yu Yang. “Multiset feature learning for highly imbalanced data classification”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[126] Hang Wu, Chihwen Cheng, Xiaoning Han, et al. “Post-surgical complication prediction in the presence of low-rank missing data”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 6808–6811.

[127] Zhipeng Xie, Liyang Jiang, Tengju Ye, and Xiaoli Li. “A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning”. In: *International Conference on Database Systems for Advanced Applications (DASFAA)*. Springer. 2015, pp. 3–18.

[128] Yang Yang, Walter Luyten, Lu Liu, et al. “Forecasting potential diabetes complications”. In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

[129] Pinar Yildirim. “Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction”. In: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. IEEE. 2017, pp. 193–198.

[130] Xi Zhang, Di Ma, Lin Gan, Shanshan Jiang, and Gady Agam. “Cgmos: Certainty guided minority oversampling”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2016, pp. 1623–1631.

[131] Maciej Zięba, Jakub M Tomczak, Marek Lubicz, and Jerzy Świątek. “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients”. In: *Applied soft computing* 14 (2014), pp. 99–108.

[132] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. “Finding the best classification threshold in imbalanced classification”. In: *Big Data Research* 5 (2016), pp. 2–8.

List of Acronyms

AI Artificial Intelligence 24

RNN Recurrent Neural Network 25

List of Figures

List of Tables

Yachao Shao

Curriculum Vitae

Education

2018–2021 **Ph.D.**, *Major: Computer Science*, Faculty of Mathematics and Computer Science, University of Goettingen, Germany, Expected.

2013–2015 **M.Sc.**, *Major: Computer Science*, School of Computer, National University of Defense Technology, China.

2009–2013 **B.S.**, *Major: Automation*, Department of Automation, University of Science and Technology of China (USTC).

Research Interests

Data mining, machine learning, deep learning, medical data science, image analysis, data management, imbalanced learning, parallel computing

Teaching Experience

2018-2021 Teaching Assistant of Computer Networks

2019 Teaching Assistant of Practical Course Networking Lab

2019-2020 Teaching Assistant of Seminar on Internet Technologies (SIT)

Academic Experience

2018–2021 **Ph.D.**, *University of Goettingen*, Institute of Computer Science, Goettingen.

Main achievements:

- **Postoperative Complication Prediction of Percutaneous Nephrolithonomy via Imbalance Learning.** This work mainly focus on analyzing the features of patient treated by PCNL according to the postoperative complications, then fixing the imbalanced classification problem of postoperative complications.
- **Multiple Balanced Subsets Stacking for Imbalanced Healthcare Datasets** This work proposes a novel ensemble method to alleviate the imbalance classification problem of healthcare datasets composed by feature vectors.
- **Hardness Aware Dynamic Loss for Imbalanced Image Classification** The work focuses on dealing with imbalanced classification problem of image datasets via dynamically customizing the class weight during the training process of the neural network.

Computer Networks Group, Goldschmidt Str.7 – Göttigen, 37077 – Germany

+86 18570393801 • *yachao.shao@cs.uni-goettingen.de*

2016-2018 **Visit Scientist**, *Jülich Research Centre*, Simulation Lab Neuroscience, Jülich.
Main achievements:

- **Arbor, the multi-compartment neural network simulation library.** Joined the Simulation Lab of Jülich supercomputing center and worked on Arbor, a high-performance library for computational neuroscience simulations with multi-compartment, morphologically-detailed cells, from single cell models to very large networks.
- **Quantitative three-dimensional reconstructions of excitatory synaptic boutons in layer 5 of the adult human temporal lobe neocortex: a fine-scale electron microscopic analysis** Made contribution in the synaptic boutons analysis part with using unsupervised learning methods; collabrated with other neural scientist to finish the fine-scale analysis.

2013-2015 **M.Sc.**, *National University of Defense Technology*, School of Computer, Changsha.
Main achievements:

- **Large-scale Brain Network Simulation Technology Study on Super Computer TH-1A** Analyzing of neural network structure, neural network models; Using analysis tools to study the memory usage and scalability of the NEST.
- **RAM analysis project of supercomputer TH-1A** Conducted experiments to collect RAM and CPU usage of TH-1A; Assembling the Tianhe-1A Supercomputer.

Selected Honors and Awards

2016–2020 CSC Scholarship, China Scholarship Council (CSC)
2011–2012 Outstanding Student Scholarship (USTC)

Languages and Computer skills

C	Basic	*2009 - Present, GCC*
C++	Basic	*2010 - Present, G++, Clion*
Python	Skillful	*2014 - Present, PyCharm, Jupyter Notebook*
git	Basic	*2017 - Present, Github, Gitlab*
Graphing	Skillful	*2011 - Present, MATLAB, Origin*
Documentation	Skillful	*2009 - Present, MS Office, WPS Office, Latex*

Publications

Conference Paper

Submitted **Yachao Shao**, Tao Zhao, Jiaquan Zhang, Shichang Ding and Xiaoming Fu. "Hardness Aware Dynamic Loss on Imbalanced Image Classification." In 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021.

Published **Yachao Shao**, Tao Zhao, Xiaoning Wang, Xiaofeng Zou and Xiaoming Fu. "Multiple Balance Subsets Stacking for Imbalanced Healthcare Dataset." In 26th IEEE International Conference on Parallel and Distributed Systems (ICPADS). pp. 300-307. IEEE, 2020.

Published Liu, Xin, Yutong Lu, Chunjia Wu, Jieting Wu, and **Yachao Shao**. "UGSD: scalable and efficient metadata management for EB-scale file systems." In Proceedings of the International Conference on Compute and Data Analysis, pp. 81-90. 2017.

Published Wei, Li, Liu Guangming, **Shao Yachao**, Liu Junlong, and Zuo You. "Optimization and application in medical big document-data of Apriori algorithm based on MapReduce." In 2016 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-5. IEEE, 2016.

Journal Paper

Submitted **Yachao Shao**, Xiaoning Wang, Xiaofeng Zou and Xiaoming Fu. "Postoperative Complication Prediction of Percutaneous Nephrolithonomy via Imbalance Learning." Artificial Intelligence in Medicine 2021.

Published Yakoubi, Rachida, Astrid Rollenhagen, Marec von Lehe, **Yachao Shao**, Kurt Sätzler, and Joachim HR Lübke. "Quantitative three-dimensional reconstructions of excitatory synaptic boutons in layer 5 of the adult human temporal lobe neocortex: a fine-scale electron microscopic analysis." Cerebral cortex 29, no. 7 (2019). pp. 2797-2814.

Published **Yachao Shao**, Guangming Liu, Si Wu and Wenrui Dong. "Brain network simulation technology on high-performance computing platform." Journal of Beijing Normal University(Natural Science), 2015 (Chinese)