공학박사 학위논문

# Imbalanced Data Learning: Advances in Techniques and Applications

## 비균등 데이터 학습 기법 및 응용 연구

2020년 2월

서울대학교 대학원

전기·컴퓨터공학부

최 현 수

# Imbalanced Data Learning: Advances in Techniques and Applications

지도교수 윤 성 로

이 논문을 공학박사 학위논문으로 제출함

2020년 2월

서울대학교 대학원

전기·컴퓨터공학부

최 현 수

최현수의 공학박사 학위논문을 인준함

2020년 2월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

## Abstract

Machine learning techniques, including deep learning, are renewing state-of-arts across many disciplines. However, many issues exist that need to be addressed for the application of these techniques to actual problems, such as medical diagnosis. A typical issue is imbalanced data, which refers to a state in which the distribution of a specific class among the accumulated data is much larger or smaller than that of the other classes. In the case of learning with imbalanced data, there is a risk of deterioration of the performance of the minority class because the learning is biased toward a majority class.

In this paper, we discuss the existing methods that address the issue of imbalanced data, and propose a new methodology using a generative adversarial neural network. The key idea of this method is a cooperative training loop of the generator and classifier, wherein the generator and classifier are trained alternately to gradually expand the decision region of the minority class. Additionally, three application studies in the biomedical field are conducted to discuss the effects and solutions of the imbalanced data, along with the significance of each study. Each applied study corresponds to the early diagnosis of dementia using neuropsychological assessment, extreme drowsiness detection based on brain waves, and electrocardiogram based biometric authentication. In summary, this paper examines the difficulties of learning caused by imbalanced data through practical application studies, and explores methodologies to solve them.



**Keywords**: imbalanced data, machine learning, deep learning, generative adversarial network, dementia diagnosis, drowsiness detection, biometric authentication, electroencephalography, electrocardiogram



**Student Number**: 2013-23144

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

Machine learning techniques, including deep learning, have significantly improved the state-of-the-art in a variety of fields [1]. However, an imbalanced data problem is one of the typical issues that causes severe degradation in performance of the machine learning techniques. An imbalanced data learning problem can be defined as a learning problem in a binary or multiple-class dataset, where the number of instances for one of the classes, called the majority class, is significantly higher than the number of instances for the rest of the classes, called the minority classes [2]. The imbalance ratio (IR) can been defined as the ratio between the majority and the minority classes.

Imbalanced data are inherent characteristics of diverse real-world applications, including medical diagnosis and bioinformatics. Especially, the medical field is one of the representative fields that faces the imbalanced data problem [3, 4, 5, 6, 7, 8, 8, 9]. Domains, such as biology, networks intrusion, and fraud detection, also suffer from the same phenomenon [9, 10, 11, 5]. An important observation is that in many of these applications, the misclassification cost of the minority classes is often higher than that of the majority class [12, 13].

Standard learning methods perform poorly in imbalanced data sets as they

induce a bias in favor of the majority class. Specifically, the minority classes contribute less to the minimization of the objective function during the training of a standard classification method. Moreover, the distinction between noisy and minority class instances is often difficult. An important observation is that in many of these applications the misclassification cost of the minority classes is often higher than that of the majority class [12, 13]. Therefore, the methods that address the class imbalance problem aim to increase the classification accuracy for the minority classes.

Various methods have been proposed to overcome the imbalanced data problem [14]. Among the existing methods, over-sampling methods [2, 15, 16, 17] have been widely used to generate minority samples. However, these methods do not work successfully for high-dimensional data, such as images [18]. In recent years, generative adversarial networks (GAN) [19] have been used to generate high-dimensional synthetic samples in the minority class [20, 21, 18, 22, 23]. However, existing GAN-based methods do not consider the knowledge that samples near the decision boundary have an important role in expanding the decision region of the minority class [24], that is, because the generator is trained independent of the classifier, the generated samples are not related to the decision boundary of the classifier. Hence, it is unclear as how the generated data are beneficial to expand the decision boundary of the minority class.

## 1.2 Contents of Dissertation

The contents of this dissertation are organized as follows: Chapter 2 covers the background knowledge regarding imbalanced data, including problem definitions, related approaches, and evaluation metrics. Chapter 3 presents a novel methodology addressing imbalanced data learning using GAN and verifies its effectiveness. Chapters 4 - 6 deal with machine learning studies on the real problem with imbalanced data and confirm the usefulness of the methodology presented in Chapter 3. Finally, Chapter 7 concludes the dissertation and discusses its contribution toward resolving research issues in imbalanced data learning with deep neural networks.

Chapter 3 addresses the imbalanced data problem by proposing a novel three-player structure (a classifier, discriminator, and generator), in a cooperative relationship between the generator and classifier. A novel regularization term is embedded to expand the decision boundary of the minority class in a cooperative interaction between the generator and classifier. Furthermore, we develop an alternating optimization strategy and regularization decay scheme, in which the generator and classifier are trained alternately to learn a desirable distribution.

Additionally, three application studies on imbalanced data learning are presented, and the results are applied to the biomedical domain using in-house real data to examine the impact of imbalance data issues on real-world problems. In Chapter 4, the first application topic is the neuropsychological assessment based early diagnosis of dementia. Because the dementia group has a smaller distribution than the normal group, the problem also has an imbalanced data issue. To achieve low-cost high-accuracy diagnosis performance for dementia using a neuropsychological battery, a novel framework is proposed using the response profiles of 2,666 cognitively normal elderly individuals and 435 dementia patients who participated in the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD). The key idea of the proposed

framework is a deep learning-based cost-effective and precise two-stage classification procedure that employs the mini mental status examination as a screening test and the neuropsychological assessment battery as a diagnostic test. The contents of given topic are based on the following research:

- **Hyun-Soo Choi**, Jin Yeong Choe, Hanjoo Kim, Ji Won Han, Yeon Kyung Chi, Kayoung Kim, Jongwoo Hong, Taehyun Kim, Tae Hui Kim, Sungroh Yoon*, Ki Woong Kim*, "Deep Learning Based Low-cost High-accuracy Diagnostic Framework for Dementia Using Comprehensive Neuropsychological Assessment Profiles." *BMC Geriatrics*, vol. 18, no. 1, p. 234, October 2018.

In Chapter 5, the second application topic is extreme drowsiness detection using brain-wave signals. Because extreme drowsiness occurs intermittently during daytime activities, its detection can also be defined as imbalanced data learning. Currently, the socioeconomic losses caused by extreme daytime drowsiness are enormous. Hence, it is necessary to build a virtuous cycle system that can be used in any environment, to improve work efficiency and safety by monitoring instantaneous drowsiness. This study proposes a novel framework to detect extreme drowsiness using a short time segment ($\sim$ 2 s) of EEG, which well represents the immediate activity changes depending on a person's arousal, drowsiness, and sleep state. To develop the framework, we used multitaper power spectral density for feature extraction along with extreme gradient boosting as a machine learning classifier. The contents of given topic are based on the following research:

- **Hyun-Soo Choi**, Seonwoo Min, Siwon Kim, Ho Bae, Jee-Eun Yoon, Inha Hwang, Dana Oh, Chang-Ho Yoon, and Sungroh Yoon, "Learning-based Instantaneous Drowsiness Detection Using Wired and Wireless Electroencephalography." *IEEE Access*, pp. 1-13, October 2019.

- **Hyun-Soo Choi**, Siwon Kim, Jung Eun Oh, Jee Eun Yoon, Jung Ah

Park, Chang-Ho Yun*, and Sungroh Yoon*, "XGBoost-Based Instantaneous Drowsiness Detection Framework Using Multitaper Spectral Information of Electroencephalography." in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, Washington DC, USA, August 2018.

In Chapter 6, the last application topic is electrocardiogram (ECG)-based biometric authentication. Biometric authentication techniques can be defined as classification problems with authorized user and others. Because the person who needs to succeed is unique and multiple people can attempt false authentication, the distribution of target classes is imbalanced. ECG signals from mobile sensors are expected to increase the availability of authentication parameters in the emerging wearable device industry. However, mobile sensors provide a relatively lower-quality signal in comparison with conventional medical devices. This study proposes a practical authentication procedure for ECG signals obtained through one-chip-solution mobile sensors. We achieved 4.61% of the equal error rate (EER) on a single heart-beat and 1.87% of EER in 15 s of testing time on 175 subjects. Despite the noisy ECG signals in the mobile sensors, the proposed method demonstrates a reasonable result and supports the usability of low-cost ECGs for biometric authentication. The contents of given topic are based on the following research:

- **Hyun-Soo Choi**, Byunghan Lee, and Sungroh Yoon*, "Biometric Authentication Using Noisy Electrocardiograms Acquired by Mobile Sensors." *IEEE Access*, vol. 4, pp. 1266-1273, March 2016.

# Chapter 2

# Background

## 2.1 Definition of Imbalanced Data

The imbalanced data problem typically occurs when there are considerably more instances of some classes than others in a classification problem. Technically, any dataset that exhibits an unequal distribution between its classes can be considered as imbalanced. However, the common understanding in the community is that imbalanced data corresponds to datasets exhibiting significant, and in some cases, extreme imbalances.

A simple example from Baptiste Rocca's blog [25] provides better understanding on the definition of imbalanced data and its difficulties, in machine learning problems. Let's suppose that we have two classes, $C_0$ and $C_1$, samples from the class $C_0$ follow the distribution of $P(x|C_i) \sim \mathcal{N}(0, 4)$, which is a one-dimensional Gaussian distribution of mean 0 and variance 4. Samples from the class $C_1$ follow the distribution of $P(x|C_i) \sim \mathcal{N}(2, 1)$, which is a one-dimensional Gaussian distribution of mean 2 and variance 1. Next, we assume the probability of class $C_0$, $P(C_0) = 0.9$ and the probability of the class $C_1$, $P(C_1) = 0.1$. In the following Figure 2.1, we depict a representative dataset containing 50 points along with the empirical distributions of both classes in the given proportions.

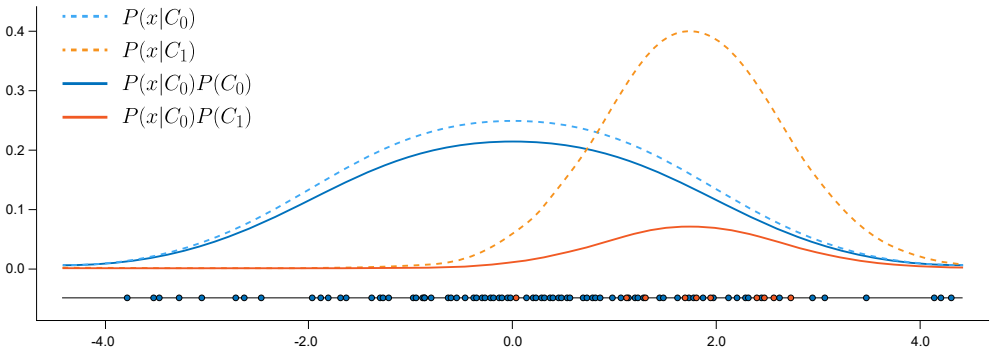



Figure 2.1: Illustration of imbalanced data example

In this example, we can see that the dotted curve of $C_0$ is always above that of $C_1$. Hence, for any given point, the probability that this point is drawn from $C_0$ C0 is always greater than the probability that it is drawn from $C_1$. Mathematically, using the basic Bayes rule, we can define

$$P(C_0|x) = \frac{P(x|C_0)}{P(C_0)}P(x) > \frac{P(x|C_1)P(C_1)}{P(x)} = P(C_1|x) \tag{2.1}$$

where we can clearly see the effect of the priors and how it can lead to a situation where one class is always more likely (have a higher probability) than the other classes. Theoretically, it is implied that if we train a classification model with these datasets, the accuracy of the classifier will always be maximal for $C_0$. However, all $C_1$ samples are incorrectly classified. If the goal is to train a classifier to achieve the best possible accuracy regardless of class, then it should not be a problem. However, misclassification cost of the minority classes is much higher than that of the majority class. Therefore, the classification methods should consider the increase of accuracy for the minority classes $C_1$.

## 2.2 Approaches for Imbalanced Data Learning

### 2.2.1 Conventional Data-level Balancing Approach

Conventional data-level approaches that are adopted to address the imbalance issues can be categorized into the following: over-sampling, under-sampling, and hybrid. The over-sampling methods generate synthetic minor samples to balance the training data. SMOTE [2] is the most well-known over-sampling method that generates a new sample on the line interpolating the k-nearest neighbors among the minority class samples. Borderline-SMOTE [15] and an adaptive synthetic sampling approach [16] improve the over-generalization issue in SMOTE by generating samples through more advanced methods. The under-sampling methods balance the training data by removing the majority samples. The clustering centroid method substitutes a cluster of the majority class with the cluster centroid, and the condensed nearest neighbor method effectively removes the majority class samples that are remote from the decision boundary [17]. SMOTEENN [26] is one of the representative hybrid methods. It combines the over-sampling algorithm, SMOTE, and under-sampling algorithm by editing the nearest neighbor decision rules (ENN).

However, all these conventional methods consider only the local information, indicating that they cannot reflect the entire data distribution. Furthermore, these methods are based on a simple distance metric (*e.g.,* Euclidean). Therefore, they do not successfully consider the high-dimensional data [18].

### 2.2.2 Cost-sensitive Loss-based Balancing Approach

The cost-sensitive loss-based approach modifies the existing classification loss function (*i.e.,* cross entropy loss) to give extra considerations to minority class samples. The advantage of the cost-sensitive approaches is that they are easy to apply, especially in deep learning. The vanilla scheme of giving extra consideration assigns (adaptive) weights to each class. Furthermore, recent works

assigned weights to each sample based on its individual properties based on the concept of hard sample mining. Hard negative samples are more informative than easy samples because they violate the boundary of the classifier by not only being on the wrong side but also far away from it. Therefore, hard negative mining improves a model quickly and effectively with less training data. Representative methods are class rectification loss (CRL) [27], max-pooling loss (MPL) [28], and focal loss [29]. However, as illustrated in our experiments in Chapter 3, the performance improvements of loss-based approaches are limited.

### 2.2.3 GAN-based Balancing Approaches

In recent years, GAN has been used as a method for over-sampling to reflect the actual distribution of a minority class [20, 21, 18, 22, 23]. Various works exploit the GAN models, such as deep convolutional GAN (DCGAN) [20], conditional GAN (cGAN) [18], or cycleGAN [22], to restore the distribution by synthesizing data. Balancing GAN (BAGAN) [21] is the most recently proposed model that is a slightly modified version of an auxiliary classifier GAN [30] to specialize in the generation of minority class samples. Different from conventional approaches, GAN-based methods can generate samples of the minority class by considering the true distribution of class data. In all these studies, the process of generating samples using the GAN and the process of learning a classifier with the generated samples are independent of each other. However, if we use the game theory in training a generator and classifier, the cooperative training between the classifier and generator will contribute to generating effective samples for training the classifier.

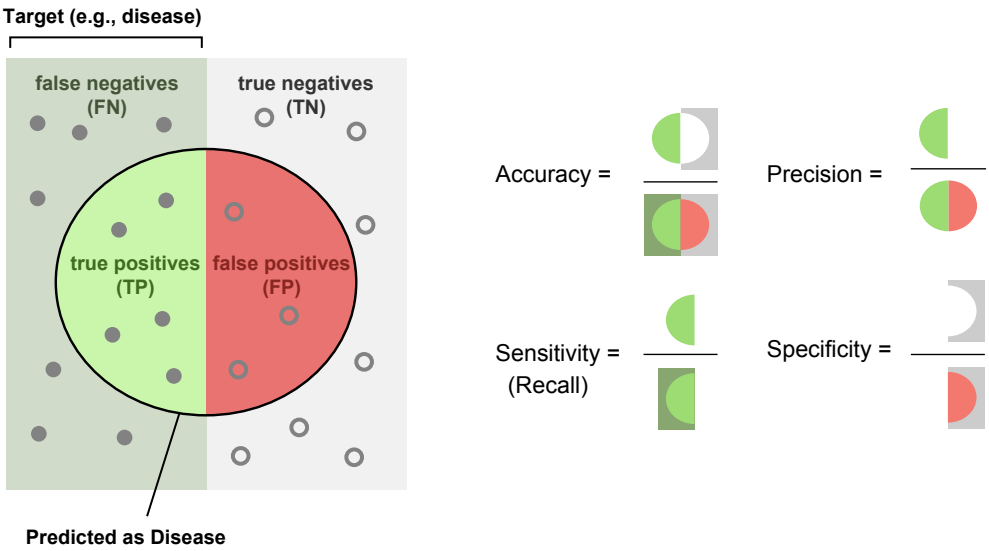



Figure 2.2: Illustration of classification metrics.

## 2.3 Evaluation Metrics

The most common metric for the classification task is accuracy, which is defined as below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2.2}$$

where TP, FP, FN, and TN are respectively true positive, false positive, false negative, and true negative. Several previous studies dealing with imbalanced data [21, 22, 20] have used accuracy as an evaluation metric. However, the accuracy metric cannot precisely evaluate the classification of the minority class for extremely imbalanced data. This is because high accuracy can be achieved by a simple zero-rule classifier, which decides that all samples belong to the majority class. If we use a zero-rule classifier for our imbalanced data, the accuracy will be 95% or more.

Other metrics should be utilized to specifically evaluate a classification of the imbalanced data. The precision expresses how trustworthy the result is when the model answers that a point belongs to a class, and is defined as

below:

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

The recall expresses the ability of the model to detect that class, and is defined as below:

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

The combinations of recall and precision have the following meanings:

- high recall, high precision: the model handles the target class perfectly.
- low recall, high precision: the model seldom detects the target class, but is highly trustworthy when detection occurs.
- high recall, low precision: the model well detects the target class, but it also includes other classes as the target class.
- low recall, low precision: the target class is poorly handled by the model.

The F1 score of a class is given by the harmonic mean of precision and recall, and it combines the precision and recall of a class in one metric as

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.5}$$

Sensitivity refers to the ability of correctly detecting the target class, which does have a certain condition. In the example of a medical diagnosis to identify a certain disease, the sensitivity, also known as detection rate in a clinical domain, is the proportion of people who test positive for the disease among those who have the disease. Sensitivity is defined as

$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.6}$$

Specificity relates to the ability to correctly reject the other class (healthy control) that is without a condition. Consider the example of a medical test

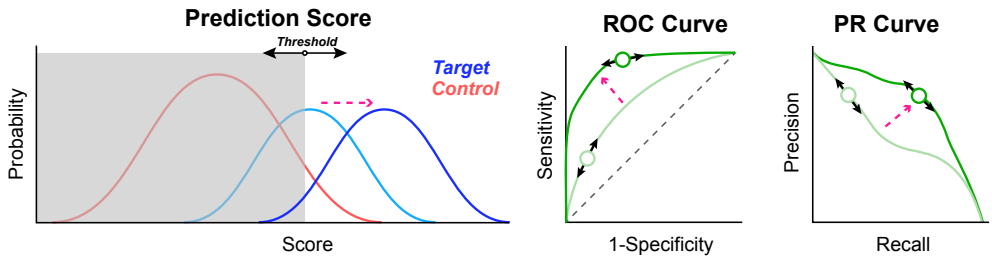



Figure 2.3: Relation between prediction score and curves (*i.e.,* receiver operating characteristic and precision-recall).

for diagnosing a disease. Specificity is the proportion of healthy controls, which do not have the disease and diagnose negative for it. Specificity is defined as

$$\text{Specificity (SPE)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.7}$$

The scheme of each metric till the prediction of the results is depicted in Figure 2.2.

The above-mentioned metrics assume the direct predictions of each class. However, classification results enable the calibration of the threshold, given the probability for each class. This is depicted by black arrows in Figure 2.3, which illustrates the interpretation of the probabilities. To address this problem, we adopt the area under the receiver operating characteristic (AUROC) curve [31] and the area under the precision-recall (AUPR) curve [32]. To calculate AUROC, the ROC curve is obtained by plotting the true positive rate (sensitivity) against the false positive rate ($1 -$ specificity) at various decision threshold settings. To calculate AUPR, the PR curve is created by plotting the positive predictive value (precision) against the true positive rate (recall) at various decision threshold settings. Because AUROC and AUPR are not specific to a particular decision threshold, these metrics can provide a more valid evaluation of the performance on the trained model than the accuracy metric. Higher AUROC and AUPR can be achieved through more separable distribution of the score and is represented by the pink arrows in Figure 2.3.

AUROC is more common than AUPR; however, AUPR is more sensitive than AUROC in terms of highly imbalanced datasets [32]. The AUPR value is measured to be very low for the imbalanced data, and as the analytical power of a classifier for the minority class increases, the AUPR value also increases by a large amount.

# Chapter 3

# GAN-based Imbalanced Data Learning Technique

## 3.1 Introduction

The imbalanced data problem is a phenomenon in which the number of samples in minority and majority classes indicates a large gap in training data. The medical field is representative fields of the imbalanced data problem [3]. Domains including biology, network intrusion, and fraud detection also suffer from the same phenomenon [9, 10, 11, 5]. The imbalance ratio (IR) between minority and majority classes varies depending on the application, and in serious cases, the IR may be as high as 100,000 [2, 33]. For many applications, it is more costly and important to classify the minority than the majority class [12, 13].

Since imbalanced data causes severe performance degradation in machine learning, it is an important research topic in both academia and industry [5]. The decision boundary learned by standard machine learning with imbalanced data can be strongly biased by the majority class, causing low precision of the minority class. Ultimately, the goal of addressing the imbalanced data problem is to increase the classification performance of the minority class.

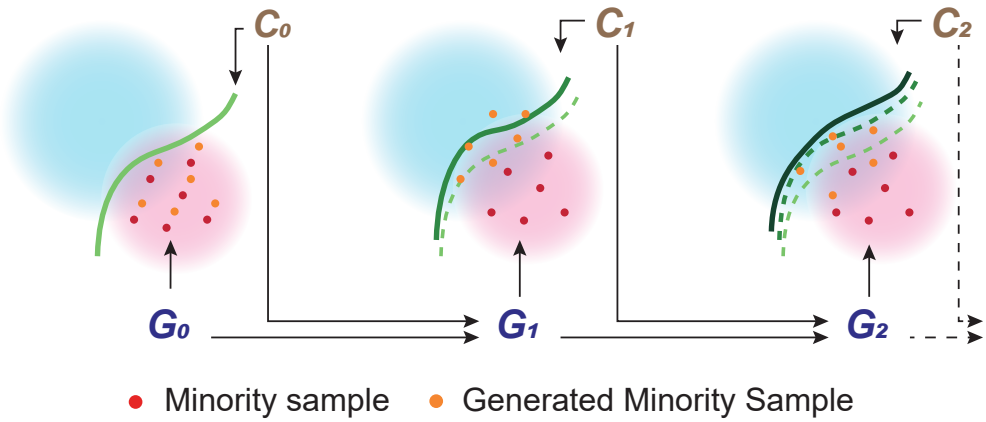



Figure 3.1: The cooperative interactions between the generator and classifier generate the minority samples near the decision boundary to expand the decision region of the minority class.

Various methods have been proposed to overcome the imbalanced data problem [14]. Among existing methods, data-level balancing methods have been widely used to balance training samples [2, 15, 16, 17]. Loss-based methods, in which loss gives larger weighting on minority samples than majority samples, have also been widely used [20, 18, 22]. In recent years, generative adversarial networks (GAN) [19] have been used to generate high-dimensional synthetic samples in the minority class [20, 21, 18, 22, 23]. Most of the existing GAN-based methods addressed in our paper do not consider the effect on a classifier in training a generator and a discriminator of GAN, thus limiting improvement opportunities for the generated samples. To handle this issue, the concept of TripleGAN [34] has been adopted to address imbalanced data classification [35]. However, since TripleGAN was originally proposed for semi-supervised learning, it has an adversarial relationship between a classifier and a discriminator, which limits performance improvement.

The samples near the decision boundary play an important role in training classifiers. For this reason, various research has the attempts to utilize the concept of decision boundary, such as knowledge distillation via decision boundary transfer [36], classifier training robust to adversarial attacks [37],

and out-of-distribution detection problems [38]. To the best of our knowledge, however, there are no attempt to address an imbalanced classification problem by generating samples with GAN to expand a decision boundary of the minority region. The novelty of our study is the decision boundary regularization, which promotes the convergence of the alternating optimization in training our three-player structure for mitigating the imbalance issue

In this paper, instead of the adversarial relationship between a classifier and a discriminator, we propose a novel cooperative relationship between a generator and a classifier. Our key concepts and contributions are as follows:

- A three-player structure (a classifier, a discriminator, and a generator) is proposed in a cooperative relationship between the generator and the classifier to address imbalanced data learning.
- A novel regularization term is embedded to expand the decision boundary of the minority class in a cooperative interaction of the generator and the classifier, as shown in Figure 3.1.
- We develop an alternating optimization strategy, along with a regularization decay scheme, in which the generator and the classifier are trained alternately to learn a desirable distribution.
- The proposed method is validated experimentally using in-depth self-analysis as well as by comparing with the existing methods.

## 3.2 Proposed Method

To formulate our concept for expanding the minority decision region to have a desirable distribution, we design a three-player structure for imbalanced data learning (Section 3.2.1) and develop an alternating training scheme with a cooperative training loop between the generator and the classifier (Section 3.2.2).

### 3.2.1 Three-Player Structure for Imbalanced Data Learning

#### Motivation

As mentioned in the section 2.2.3, TripleGAN [34] is designed to generate pseudo-labels for unlabeled samples for facilitating semi-supervised learning. The discriminator ($D$) in TripleGAN discriminates between a true and a false label which is generated by the classifier ($C$). TripleGAN has an adversarial relationship $\mathcal{U}(C, D)$ between $D$ and $C$, as seen in Figure 3.2 (a). However, in the case of an imbalanced data problem, the adversarial relationship is not useful and may lead to an unstable convergence. In the proposed model, a cooperative relationship is developed between G and C to ensure that both G and C are benefitted by joint training. For developing the cooperative relationship, additional utility terms $\mathcal{U}(G, C)$ and $\mathcal{R}(G, C)$ have been proposed as depicted in Figure 3.2 (b). The proposed three-player structure is designed to expand minority region by generating minority samples towards the borderline between the majority and the minority in the early stage of training, and finally to provide densely distributed samples within an expended minority region. In the following section, we describe the details of the proposed utility function and discuss its impact on imbalanced data learning.

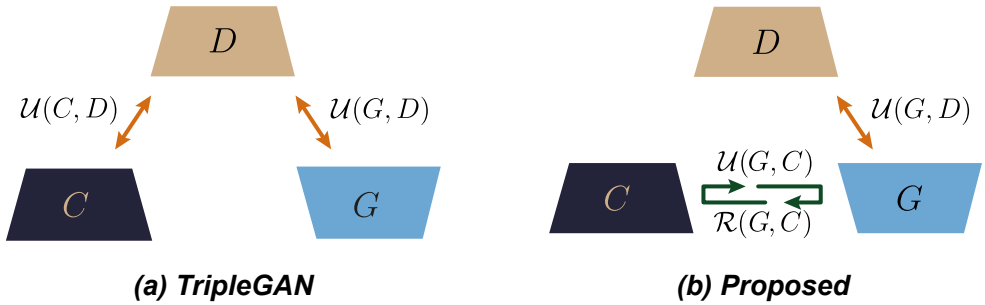



Figure 3.2: Difference between TripleGAN and Proposed.

**Utility Function**

To describe our utility function for our three-player structure shown in Figure 3.3, we define notations. $\mathbf{x}$ denotes the input data and $y$ denotes the output label. Then, $\mathbf{x} = G(\mathbf{z}, y)$ denotes a generated sample from the randomly generated $\mathbf{z}$ and $y$ values. It is assumed that the observed training samples are sampled from unknown $p(\mathbf{x}, y)$ and that samples from both $p(\mathbf{z})$ and $p(y)$ can be easily obtained by using simple known distributions (normal or uniform, etc.) during training. The classified label is denoted by $y = C(\mathbf{x})$ and the output of $D$ is denoted by $D(\mathbf{x}, y)$ for given $\mathbf{x}$ and $y$. Additionally, the joint distributions $p_g(\mathbf{x}, y)$ and $p_c(\mathbf{x}, y)$ are defined as

$$p_g(\mathbf{x}, y) := p(y)\, p_g(\mathbf{x}|y) = p(y)\, p(G(\mathbf{z}, y)|y)\,, \tag{3.1}$$

$$p_c(\mathbf{x}, y) := p(\mathbf{x})\, p_c(y|\mathbf{x}) = p(\mathbf{x})\, p(C(\mathbf{x})|\mathbf{x})\,. \tag{3.2}$$

Here, $p_g(\mathbf{x}|y) = p(G(\mathbf{z}, y)|y)$ in (3.1) indicates the distribution of synthetic samples generated by $G$ for a given label $y$. $p_c(y|\mathbf{x}) = p(C(\mathbf{x})|\mathbf{x})$ in (3.2) indicates the distribution of labels, determined by $C$, for the given samples (generated or observed).

Our goal is to design a utility function $\mathcal{U}(C, D, G)$ for imbalanced data

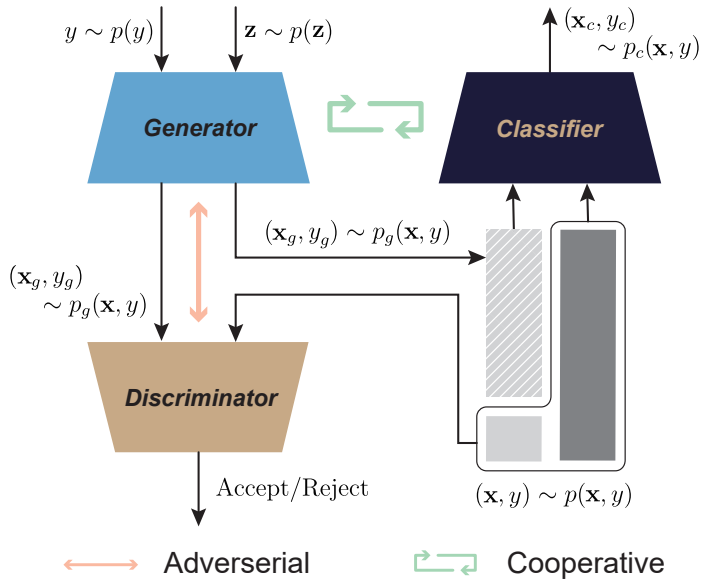



Figure 3.3: The proposed GAN architecture.

learning in three-player game given by

$$\min_{C,G} \max_{D} \mathcal{U}(C, D, G). \tag{3.3}$$

In this paper, the utility function for imbalanced data learning is proposed as

$$\begin{aligned} \mathcal{U}(C, D, G) = \mathcal{U}_g(D, G) + \mathcal{U}_{c_1}(C) + \\ (1 - \lambda)\mathcal{U}_{c_2}(G, C) + \lambda\mathcal{R}(G, C), \end{aligned} \tag{3.4}$$

where the last two terms are distinctive aspects against TripleGAN and they take key roles for cooperative training of $G$ and $C$ in our method. The third term is for jointly training of $G$ and $C$, whereas the fourth term is for minority region expansion. These two terms are linked by a hyper-parameter $\lambda$ for trade-off scheduling between the two terms. Each term is defined formally in the following.

The term $\mathcal{U}_g(D, G)$ is well known utility function of cGAN, which is defined

as

$$\mathcal{U}_g(D,G) = \mathbb{E}_{p(\mathbf{x},y)}\big[\log D(\mathbf{x},y)\big] + \mathbb{E}_{p_g(G(\mathbf{z},y),y)}\big[\log(1-D(G(\mathbf{z},y),y))\big]. \tag{3.5}$$

The term $\mathcal{U}_{c_1}(C)$ is for training $C$ with only the observed (real) data, whereas $\mathcal{U}_{c_2}(G,C)$ is for joint training of $G$ and $C$, which are defined as

$$\mathcal{U}_{c_1}(C) = \mathbb{E}_{p(\mathbf{x},y)}\big[-\log p_c\left(y|\mathbf{x}\right)\big], \tag{3.6}$$

$$\mathcal{U}_{c_2}(G,C) = \mathbb{E}_{p_g(G(\mathbf{z},y),y)}\big[-\log p_c\left(y|G\left(\mathbf{z},y\right)\right)\big]. \tag{3.7}$$

In particular, $\mathcal{U}_{c_2}(G,C)$ makes $C$ be trained to well classify the samples generated by $G$, whereas $G$ be trained to generate extra samples helpful for $C$.

Lastly, $\mathcal{R}(G,C)$ is introduced for expansion of the minority region. To define $\mathcal{R}(G,C)$, the classification scores for the minority and majority classes are denoted by $C_{\mathrm{mi}}\left(G(\mathbf{x})\right)$ and $C_{\mathrm{ma}}\left(G(\mathbf{x})\right)$, respectively, and a generated minority sample is denoted by $\mathbf{x}_g^{\mathrm{mi}}$. Using these terms, $\mathcal{R}(G,C)$ is defined as

$$\mathcal{R}(G,C) = \mathbb{E}_{p_g(\mathbf{x},y)}\big[s_g^{\mathrm{mi}}\big], \tag{3.8}$$

where

$$s_g^{\mathrm{mi}} = \begin{cases} \big[C_{\mathrm{mi}}(\mathbf{x}_g^{\mathrm{mi}}) - C_{\mathrm{ma}}(\mathbf{x}_g^{\mathrm{mi}})\big]^2, & \text{if } C_{\mathrm{mi}}(\mathbf{x}_g^{\mathrm{mi}}) > C_{\mathrm{ma}}(\mathbf{x}_g^{\mathrm{mi}}) \\ 0, & \text{otherwise.} \end{cases}$$

The role of $\mathcal{R}(G,C)$ is presented in the following section.

**Effect to Imbalanced Data Learning**

In $\mathcal{R}(G,C)$, if $\mathbf{x}_g^{\mathrm{mi}}$ is placed in the minority region of the current state of $C$ (upper condition in (3.8)), $\mathbf{x}_g^{\mathrm{mi}}$ moves towards the majority region as $\mathcal{R}(G,C)$ is minimized. While, if $\mathbf{x}_g^{\mathrm{mi}}$ is placed in the majority region (the lower condition

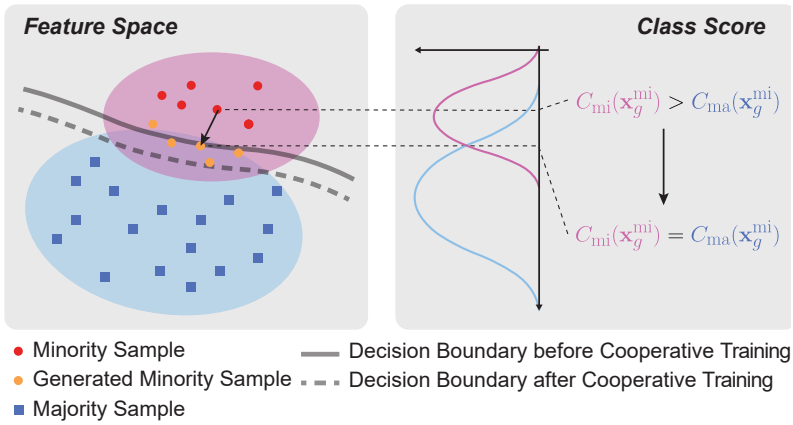



Figure 3.4: Through cooperative training, $G$ is trained to generate minority samples (yellow) crossing the decision boundary between the minority class $C_{\text{mi}}$ and the majority class $C_{\text{ma}}$. As indicated by the dashed line, the samples generated by the tuned $G$ contribute to the expansion of the minority class region for the next training of $C$.

in (3.8)), this sample is already beneficial to expand the region, thus, it does not need to move. Hence, as shown in Figure 3.4, minimizing $\mathcal{R}(G, C)$ plays a role in generating minority samples to expand the minority region in the direction of the majority region in training $C$.

However, if we constantly expand the minority region, the classification performance of the majority class would degrade. To prevent degradation, we introduce a hyper-parameter $\lambda$ for a trade-off scheduling between $\mathcal{U}_{c_2}(G, C)$ and $\mathcal{R}(G, C)$. By reducing $\lambda$ gradually to zero during the alternate training of $C$ and $G$, the role of $\mathcal{R}(C, G)$ vanishes. This implies that $G$ is trained for the expansion of the minority class decision region in the early training stage only. As $\lambda$ decays, the term of $\mathcal{U}_{c_2}(G, C)$ contributes to achieving a densely distributed decision region. More details about the decaying scheme are described in the training section. **Theorem 1** shows the proposed utility function has an equilibrium when $\lambda$ decays to zero.

**Theorem 1.** *The equilibrium of* $\mathcal{U}(C, D, G)$ *with* $\lambda = 0$ *is achieved if and only if*

$$p(\mathbf{x}, y) = p_g(\mathbf{x}, y) = p_c(\mathbf{x}, y) = p_c(G(\mathbf{z}|y), y). \tag{3.9}$$

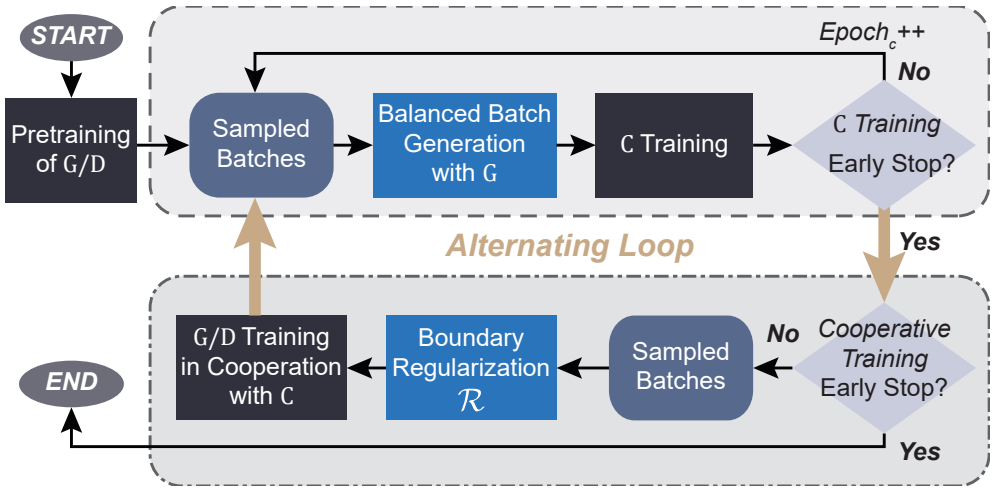



Figure 3.5: Alternating training scheme of proposed method.

Note that $p_c(\mathbf{x}, y) = p_c(G(\mathbf{z}|y), y)$ means that the training of $C$ relies on the distribution of samples generated by $G$ at the equilibrium. Hence, how well $G$ learns the true distribution dominates the performance of $C$. The proof is given in Appendix 1.

### 3.2.2 Training Scheme

#### Overall Scheme

To promote cooperation between $G$ and $C$, along with $\mathcal{U}_{c_2}(C, G)$ and $\mathcal{R}(G, C)$, in optimization process, we adopted an alternating optimization strategy. The overall scheme of the proposed method is outlined in Figure 3.5. Before starting alternating optimization, we pretrained $G/D$ with the observed imbalanced data for the initial generator. As the first step of the alternating loop, $C$ is trained with a balanced batch generated by fixed $G/D$. Thereafter, $G/D$ is trained in cooperation with $C$, along with the decision boundary regularization $\mathcal{R}(G, C)$. These two optimizations are repeated iteratively in an alternating loop. Each optimization is described in the upcoming sections. The alternating loop induces $G$ to generate minority samples that help $C$ expand the minority region during the initial training phase. As $\lambda$ decays with increasing iterations,

the joint term $\mathcal{U}_c(G, C)$ plays a major role in achieving a desirable distribution within each decision region determined by the trained $C$. To escape each loop, we adopted the validation-based early stopping rule [39].

The training parameters of $G$, $D$, and $C$ are denoted by $\theta_g$, $\theta_d$, and $\theta_c$, respectively. Then, the empirical utility function is denoted by parameterized functions as

$$\begin{aligned}\widetilde{\mathcal{U}}(\theta_d, \theta_g, \theta_c) = \widetilde{\mathcal{U}}_g(\theta_d, \theta_g) + \widetilde{\mathcal{U}}_{c_1}(\theta_c) \\ + (1-\lambda)\widetilde{\mathcal{U}}_{c_2}(\theta_c, \theta_g) + \lambda\mathcal{R}(\theta_c, \theta_g).\end{aligned} \tag{3.10}$$

The training of $\theta_c$ with a balanced batch generated by $G$ and the training of $\theta_g/\theta_d$ in cooperation with $C$, are described in the following sections.

**Training of $C$ with Balanced Batch by $G$**

In this stage, only $C$ is trained using the empirical utility function, $\widetilde{\mathcal{U}}(\theta_d, \theta_g, \theta_c))$, in (3.10). That is, only the parameter vector $\theta_c$ of $C$ is updated after fixing $\theta_g$ and $\theta_d$. As $\theta_g$ is fixed, $\theta_c$ is updated by descending the empirical utility function in (3.10) along its stochastic gradient with respect to $\theta_c$. The samples of minority class for balancing are generated by the trained $G$ in a batch-wise manner, whereas the existing GAN-based balancing methods adopt a one-shot balancing policy. In a one-shot balancing policy, the fixed number of minority samples are generated before training $C$ as a preprocessing step. In batch-wise balancing, however, new samples are generated for each batch. Batch-wise balancing is advantageous because it can fully utilize $G$ by generating an unlimited number of samples, as new samples are generated repeatedly in a batch-wise manner until $C$ converges. Another advantage is memory efficiency. Unlike one-shot balancing, batch-wise balancing requires only a small amount of memory for as much as one batch size.

**Training of $G/D$ in Cooperation with $C$ along with $\mathcal{R}$**

This training stage is designed to train $G/D$ to pursue a balanced distribution by expanding the minority decision region and generating sufficient samples within the decision region. To prevent over-expansion of the minority region, we designed a decaying rule of $\lambda$ in the utility function (3.10). Specifically, $\lambda$ is exponentially reduced by multiplying hyper-parameter $\gamma \in (0, 1)$ every iteration loop (*i.e.*, $\lambda = \gamma^i$ for $i$-th iteration). The value of $\gamma$ is empirically selected in experiments. In each alternating loop, by fixing $\theta_c$, $\theta_g/\theta_d$ are updated by descending/ascending $\widetilde{\mathcal{U}}(\theta_d, \theta_c, \theta_g)$ in (3.10) along their stochastic gradient with respect to $\theta_g/\theta_d$. Note that $\theta_g/\theta_d$ also can be trained for several epochs in each loop, but one epoch was empirically sufficient.

## 3.3 Experimental Results

### 3.3.1 Evaluation Setup

#### Datasets to Validate Our Method

We adopted one synthetic and two real imbalanced datasets. Two factors were considered in synthetically constructing an imbalanced dataset that cannot be easily classified. The first factor is the degree of similarity between classes. Learning will be easy if both classes are distinct from each other, even if a considerably small number of samples is provided for the minority class. Therefore, constructing classes with high similarity is desirable for evaluating the performance of imbalanced data learning. The second factor is the imbalance ratio (IR) between classes. Previous studies constructed a synthetic dataset with a low IR, not higher than 2.5 (100:40) [21]; however, these low IR data are insufficient to verify the methods for an extremely imbalanced case. It is therefore desirable to set a sufficiently large value for the IR.

Considering these two factors, we constructed two training datasets from the CIFAR10 [40] data. Based on the first factor, we selected two highly similar classes from the original CIFAR10 dataset: car (majority class) vs. truck (minority class), Cat (majority class) vs. Dog (minority class), Horse (majority class) vs. Deer (minority class), and Airplane (majority class) vs. Ship (minority class). Based on the second factor, we set IR to 20 (100:5). We used all samples in the majority class dataset and randomly selected 5% of the samples from the minority class dataset. A validation dataset was constructed with 20% of samples in the selected training dataset. For the test dataset, we used the original test dataset released by CIFAR10. Using this data, we deeply self-analyze our method in various aspects detailed in Section 3.3.2. Furthermore, to verify the effectiveness of the proposed method on the multi-class classification problem, with reference to [41], we made an imbalanced multi-class dataset by extraction 5% of the samples for half classes (0, 2, 4, 6,

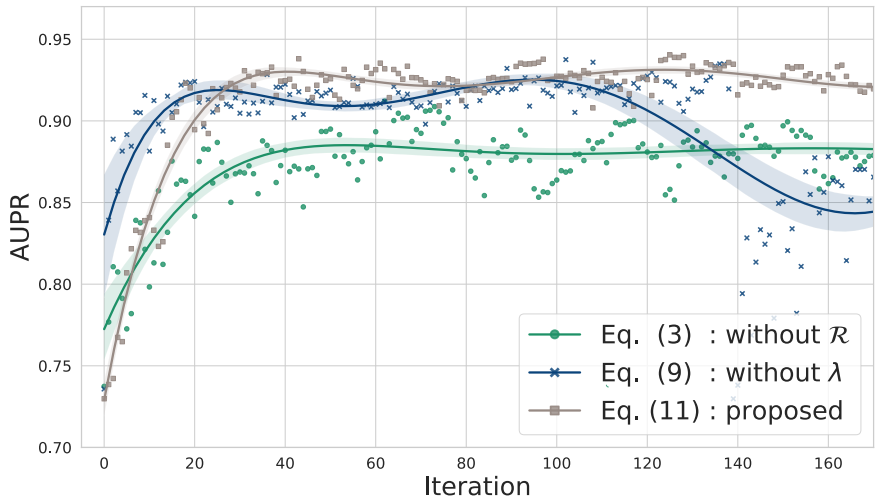



Figure 3.6: Effect of $\lambda$ along with $\mathcal{R}(G, C)$. The proposed method (with $\lambda$ decay) shows more stable and higher performance than the other two cases.

8) in CIFAR10.

To verify the generality on real imbalanced data, we chose two datasets. The first real dataset is a dementia diagnosis data set presented on [42]. Dementia dataset is composed of 92 dimensions of neuro-psychological assessment profiles. There are eight times as many control subjects as dementia subjects (IR = 8). The second real dataset is CelebA [43]. This dataset is composed of 200,000 portraits with 40 classes of multi-labels, and some attributes, such as baldness or hat-wearing, are extremely imbalanced. By validating our method with CelebA, we could confirm its generality for multi-label classification problems as well as for binary classification problems. Two real imbalanced datasets are used for comparative analysis.

### 3.3.2 Self-analysis

**Effect of $\lambda$ along with $\mathcal{R}(G, C)$**

To analyze the influence of $\lambda$ and its decay scheme along with $\mathcal{R}(G, C)$ in (3.4), we evaluated the convergence of the optimization process for each of the

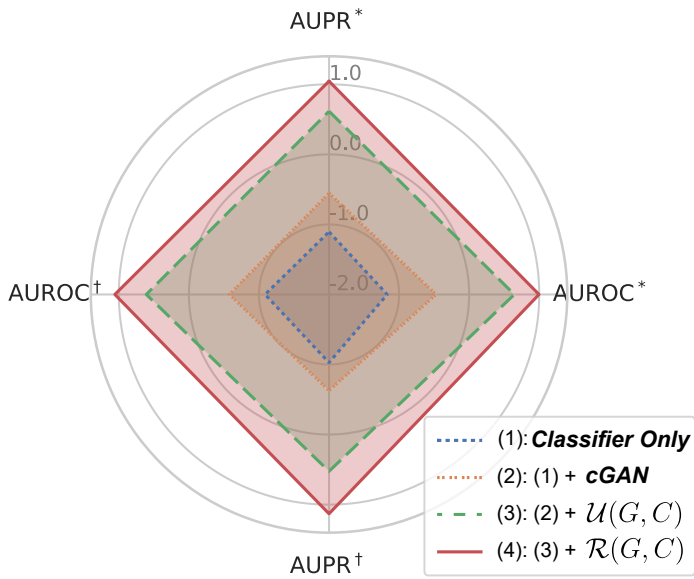



Figure 3.7: Radar chart for ablation comparison of classifier performance on CIFAR10. Scores are from the validation* and test† sets. For better visualization, each score is normalized with mean and variance.

three settings: without $\mathcal{R}(G, C)$, without $\lambda$ decay scheme, and with $\lambda$ decay scheme. Figure 3.6 shows the results of the three cases using CIFAR10. In the case without $\mathcal{R}(G, C)$ (green line), using the utility function in (3.3), the performance was not much improved due to the premature convergence explained in the methods section. In the case without a $\lambda$ decay scheme (blue line), performance degraded after approximately 100 iterations, due to over-expansion of the minority region. In contrast, the case with $\lambda$ decay (proposed, brown line), using the utility function in (3.4), the high and stable performance was achieved as expected. The degree of decay for $\lambda = \gamma^i$ is determined by the value of $\gamma$, which is observed to be dependent on the dataset. We determined $\gamma$ empirically as 0.9, 0.1, and 0.5 for CIFAR10, Dementia, and CelebA, respectively.

### Ablation Study

The ablation study was conducted with CIFAR10 by sequentially adding each ablation component because each component could not be implemented without the previous components. The role of the components is validated through an ablation study on CIFAR10 through ablation of one baseline and three variants as listed in the table.

| | | | |
|---|---|---|---|
| Baseline | w/o $\mathbf{x}_g^{\text{mi}}$ | w/o $\mathcal{U}_{c_2}(G, C)$ | w/o $\mathcal{R}(G, C)$ |
| Variant 1 | w/ $\mathbf{x}_g^{\text{mi}}$ | w/o $\mathcal{U}_{c_2}(G, C)$ | w/o $\mathcal{R}(G, C)$ |
| Variant 2 | w/ $\mathbf{x}_g^{\text{mi}}$ | w/ $\mathcal{U}_{c_2}(G, C)$ | w/o $\mathcal{R}(G, C)$ |
| Variant 3 | w/ $\mathbf{x}_g^{\text{mi}}$ | w/ $\mathcal{U}_{c_2}(G, C)$ | w/ $\mathcal{R}(G, C)$ |

Note[1] w/o: without, w/ : with, $\mathbf{x}_g^{\text{mi}}$: generated minority samples.
Note[2] Variant 3 is the proposed one. Baseline uses only $C$.

Figure 3.7 depicts the results of the ablation study. First, on Variant 1 (orange line), performance improves slightly compared to learning using only $C$ (blue line). As this variant corresponds to existing cGAN, the amount of improvement is not significant. On Variant 2 (green line), a significant improvement of performance is achieved in addition to the first ablation (green line). This implies the terms for joint training of $G$ and $C$ along with alternating training contributes to both $G$ and $C$ so that $C$ helps $G$ generate samples beneficial to $C$, consequently improving $C$'s performance. Finally, when the $\mathcal{R}(G, C)$ term was added as Variant 3, it significantly improved since $G$ generated samples to interactively expand the minority region (red line).

### Validity of Samples Generated Throughout Cooperative Training

Figure 3.8 shows a map of the samples generated by the proposed GAN in the feature space. The blue and red contours represent the majority and minority class distributions, respectively, for the given training data. The dark red dots represent the 64 samples generated by $G$. Features in the intermediate layer of $C$ were extracted for all samples and were visualized in two-dimensional space

using the parametric t-distributed stochastic neighbor embedding scheme [44]. For fair visualization, we used a fixed $\mathbf{z}$ to generate samples at each iteration.

In Figure 3.8, the leftmost panel shows the samples generated by cGAN learning, which was only trained in the initial phase, without cooperative training. Most of the samples are mapped in a small region within the training data distribution. The remaining panels show a map of the samples generated through repeated cooperative training. Although the samples were generated using the same $\mathbf{z}$ values, they are mapped in different positions of the feature space in every iteration. Especially in the first cooperative training, as the value of $\lambda$ is 1, most of the generated minority samples cross the decision boundary between two classes. We can see that as the $\lambda$ value decays, the tendency of generating samples cross the decision boundary decreases.

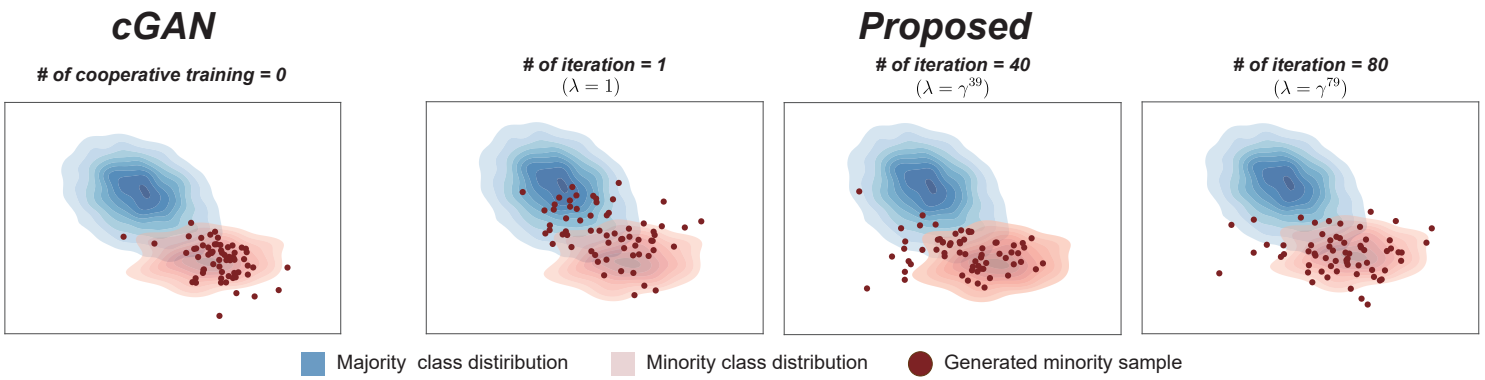



Figure 3.8: Distribution of each class and generated minority samples in feature space. Without cooperative training, generated samples are located within the training data distribution. However, with cooperative training, generated samples tend to be located on the borderline. As $\lambda$ decays, generated samples return to the distribution with broader coverage.

As most data-level sampling methods provide samples only in the inner region of the training data distribution, they risk over-fitting [45]. In contrast, we can observe that several samples generated by our method are positioned over the decision boundary between two classes. This result implies that the proposed method can expand the minority region to improve the generalization performance of $C$ on the minority class. After the regularization term vanishes by reducing $\lambda$ to almost zero, the generated samples cover a wide region of the minority class, as shown in the fourth map of Figure 3.8.

Figure 3.9 shows the locations of the generated minority samples in feature space. Top right images in (a) and (b) are the generated sample images. The numbers left to the generated images are the indexes which correspond to the numbers in feature space. Figure 3.9-(a) represents the generated sample locations after the first cooperative training. As discussed in section 4.5, due to $\lambda = 1$, the generated minority samples are located around the borderline of two classes. Figure 3.9-(b) represents the generated sample locations after 80th cooperative training. As $\lambda$ converges to 0, the generated samples are located within the original distribution rather than the borderline. Even though the images with the same index in (a) ad (b) are generated with the same value of $z$, appearances of the two images with the same index are different from each other. Many of the generated images in Figure 3.9-(a) look like a car (low and round). This figure illustrates that $G$ trained in the initial cooperative training phase can generate the ambiguous minority samples that look like majority samples. These ambiguous minority samples are beneficial to the expansion of the minority region. However, as $\lambda$ converges to zero, the generated images become similar to truck image (high and box-style) as shown in Figure 3.9-(b).

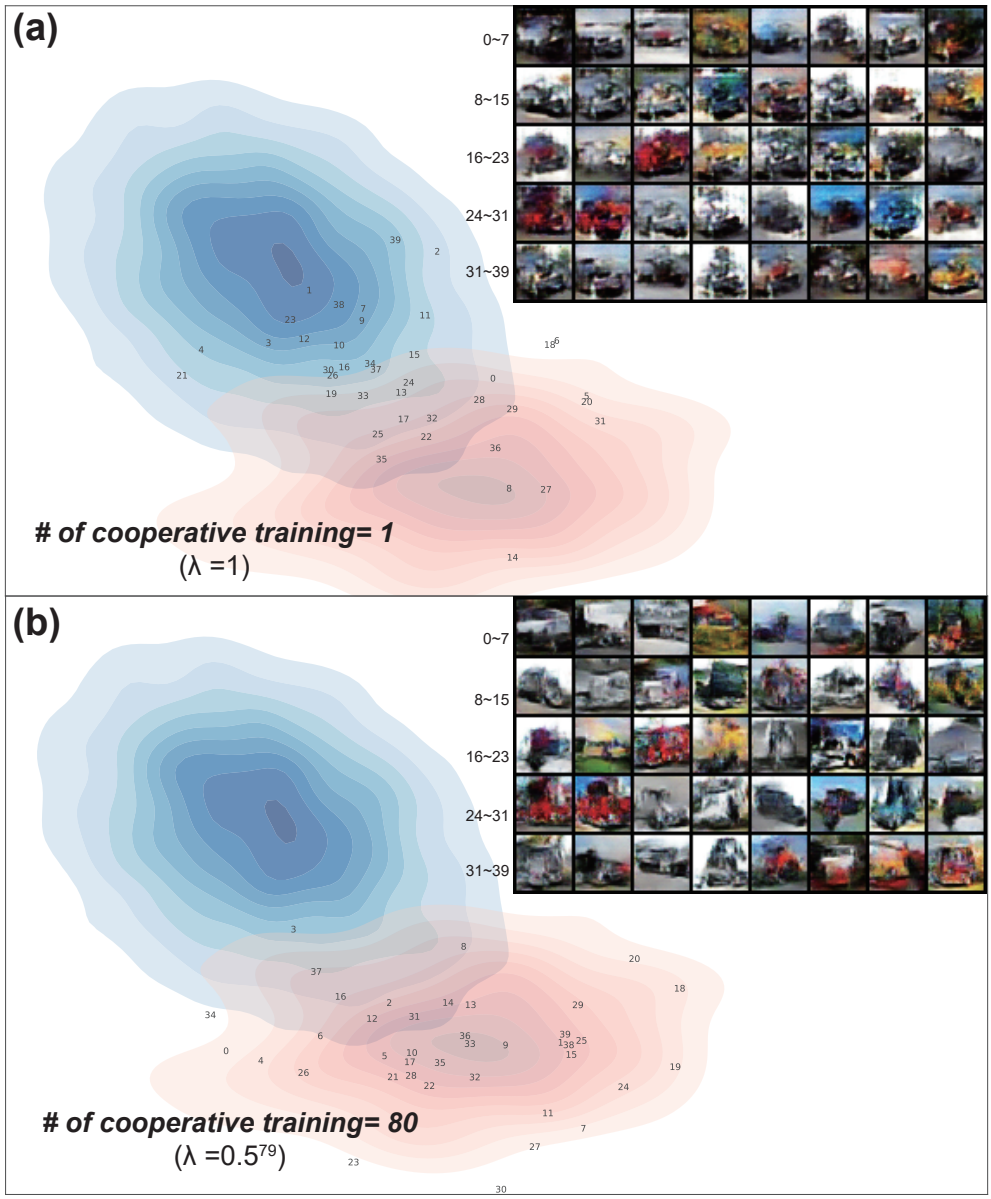



Figure 3.9: Feature space mappings and images of generated minority samples (truck) against majority samples (car).

### 3.3.3 Comparative Analysis

To verify the validity of the proposed method, we compared the classification performance to several existing techniques using three datasets.

#### Compared Methods

For the conventional data-level methods, we adopted six methods described in section 2.2.1. For implementation, we used the imbalanced-learn tool [46]. The compared loss-based methods are CRL [27], MPL [28], and focal loss [29]. GAN-based techniques were compared to three other methods. The first method is based on cGAN, which is used in most GAN-based approaches. The structure of cGAN is the same as that used in our work. The second method is BAGAN [21]. The authors of BAGAN released the source code, and the structure and hyper-parameters specified in their paper were used. The third GAN-based method is TripleGAN [34], HexaGAN [35] uses the concept of TripleGAN for imbalanced data problem.

#### Hyper-parameters and Experimental Settings

For a fair comparison, hyper-parameters of the classifier for each dataset are searched for the classifier only (baseline) case. Then, the same set of classifier's hyper-parameters were used for the others. Besides, the unique hyper-parameters of each method, such as $\gamma$ of focal loss [29], were searched within a specific range following their guidelines and selected with the values which showed the best validation performance. In the case of GAN-based techniques, the same structure of G and D were used, except for BAGAN having its own structure.

Table 3.1: Test set performance of CIFAR10 (car vs. truck), Dementia, and CIFAR10 (all labels) dataset.

| Methods | | CIFAR10 (car vs. truck)† | | | Dementia (control vs. patient) | | | CIFAR10 (all labels - multi-class)‡ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUPR | AUROC | F1 | AUPR | AUROC | F1 | Mean Accuracy | Micro AUPR | Micro AUROC | Micro F1 |
| Classifier only | | 0.8275 | 0.8223 | 0.3498 | 0.9260 | 0.9741 | 0.8449 | 0.5824 | 0.6543 | 0.9033 | 0.5569 |
| Conventional Data-level | SMOTE | 0.8361 | 0.8287 | 0.4764 | 0.9183 | 0.9712 | 0.4515 | 0.5582 | 0.6315 | 0.8927 | 0.5208 |
| | B-SMOTE | 0.7966 | 0.7861 | 0.4638 | 0.9083 | 0.9679 | 0.4910 | 0.5467 | 0.6187 | 0.8906 | 0.5121 |
| | ADASYN | 0.8795 | 0.8711 | 0.5685 | 0.9045 | 0.9661 | 0.3637 | 0.5914 | 0.6215 | 0.9003 | 0.5666 |
| | CC | 0.6551 | 0.6591 | 0.5806 | 0.8830 | 0.9643 | 0.8050 | 0.3121 | 0.3071 | 0.7575 | 0.2778 |
| | CNN | 0.7399 | 0.7417 | 0.5520 | 0.9099 | 0.9702 | 0.8063 | 0.5052 | 0.5518 | 0.8709 | 0.4724 |
| | SMOTEENN | 0.8405 | 0.8421 | 0.6731 | 0.9049 | 0.9694 | 0.4847 | 0.4500 | 0.4533 | 0.8466 | 0.4482 |
| Loss-based | CRL | 0.8743 | 0.8714 | 0.4445 | 0.9149 | 0.9732 | 0.7711 | 0.5746 | 0.5758 | 0.8500 | 0.5522 |
| | Focal | 0.8595 | 0.8417 | 0.5647 | 0.8821 | 0.9661 | 0.8197 | 0.5770 | 0.5603 | 0.8870 | 0.5506 |
| | MPL | 0.8321 | 0.8253 | 0.6379 | 0.9188 | 0.9733 | 0.8471 | 0.5593 | 0.6307 | 0.8951 | 0.5282 |
| GAN-based | cGAN | 0.8479 | 0.8491 | 0.5869 | 0.9283 | 0.9748 | 0.9200 | 0.5388 | 0.6136 | 0.8873 | 0.5032 |
| | BAGAN | 0.7641 | 0.7740 | 0.4071 | 0.9112 | 0.9691 | 0.9042 | 0.5793 | 0.6486 | 0.9019 | 0.5516 |
| | TripleGAN | 0.8412 | 0.8381 | 0.4902 | 0.9305 | 0.9743 | 0.8740 | 0.6031 | 0.6734 | 0.9142 | 0.5790 |
| | Proposed | **0.9375** | **0.9350** | **0.7602** | **0.9391** | **0.9773** | **0.9215** | **0.6106** | **0.6889** | **0.9146** | **0.5866** |

† The additional results on other two class combinations beside car vs. truck in CIFAR 10 are given in Appendix 2.

‡ Majority class: 1, 3, 5, 6, 7, Minority class (5% of training samples): 0, 2, 4, 6, 8

Table 3.2: Test set performance of multi-label CelebA

| Methods | | CelebA | | |
|---|---|---|---|---|
| | | Mean AUPR | Mean AUROC | Mean F1 |
| Classifier only | | 0.7137 | 0.9162 | 0.7861 |
| Loss-based | CRL | 0.6982 | 0.9207 | 0.6938 |
| | Focal | 0.6857 | 0.9044 | 0.7881 |
| | MPL | 0.6154 | 0.8630 | 0.7428 |
| GAN-based | cGAN | 0.7138 | 0.9177 | 0.7944 |
| | TripleGAN | 0.7033 | 0.9123 | 0.7853 |
| | Proposed | **0.7293** | **0.9226** | **0.7979** |

Conventional data-level methods and BAGAN do not support multi-label.

### Comparison Results

The comparative results are listed in Table 3.1 and 3.2. The results for other class combinations, other than car vs. truck for CIFAR10, are shown in Appendix 2. Our method outperformed all the compared methods on all the datasets consistently. For binary CIFAR10 (car vs. truck, etc.), most of the existing methods tend to show improvement against the classifier only. But some methods such as under-sampling methods only show improvements in F1 score and not for AUPR and AUROC. The proposed outperforms existing methods on not only binary-class problems but also CIFAR10 multi-class problems. The baseline exhibits a relatively higher performance in the case of dementia due to the low value of IR. It is observed that the proposed method is accurate and shows consistent improvement in all metrics. As CelebA also has various IR values (1 to 43) for each attribute, some methods are degraded, but our method shows improvement in all metrics consistently.

In particular, like the proposed method, B-SMOTE also considers data boundary found using simple $k$ nearest neighbors in the input space. However, in general, the separability in feature space is crucial for the classification problem rather than input space. The proposed method, differently from B-SMOTE, utilizes the decision boundary in the feature space induced by the trained classifier. Hence the proposed method effectively performs minority

sample generation in the feature space. In addition, TripleGAN (HexaGAN) has a three-player structure like the proposed model. However, the classification performance was lower than that of ours, because the adversarial relationship between $C$ and $D$.

## 3.4 Conclusions

To overcome the difficulty of imbalanced data learning, we proposed a novel methodology based on a three-player game and decision boundary regularization. First, we designed a three-player structure to improve imbalanced data learning performance and analyzed the equilibrium point of the proposed utility function. Second, we introduced a decision boundary regularization to expand the minority region determined by the trained classifier with samples generated by the generator in our three-player structure. Third, we proposed an alternating training scheme to effectively train the three-player structure, in cooperation with the decision boundary regularization. The experiment illustrated that the proposed method outperforms the existing methods by yielding abundant samples to expand the minority decision region, which is beneficial in addressing imbalanced data learning problems.

Although the proposed method showed promising results, certain issues remain. In this study, with a relatively simple form of cGAN, the proposed method achieved a considerable performance improvement in imbalanced data classification. As further work, the use of more precise generators and discriminators is expected to yield higher performance for the imbalanced data learning. Additionally, for the $\lambda$ decay schedule, an exponential decay rule was used where the decaying degree was empirically determined depending on the datasets. For further improvement, a more elaborate design or adaptive scheme for $\lambda$ decay should be adopted, which would consider the imbalance ratio and complexity of the target dataset.

# Appendix 1

The proof of the following **Lemma 1** is equivalent to the proof[1] of the original GAN [19], here we briefly summarize the original proof by rewriting it. For the details, refer to the reference in footnote. Here we add **Theorem 1** for the proof of the three-player game proposed in this paper.

**Lemma 1.** *For any fixed $G$ in $\mathcal{U}_g(D,G)$, the optimal discriminator $D$ is given by*

$$D^*(\mathbf{x},y) = \frac{p(\mathbf{x},y)}{p(\mathbf{x},y)+p_g(\mathbf{x},y)}. \tag{3.11}$$

**Proof.** Given $G$, $\mathcal{U}_g(D,G)$ can be rewritten as

$$\begin{aligned}\mathcal{U}_g(D,G) = &\int\int p(\mathbf{x},y)\log D(\mathbf{x},y)dyd\mathbf{x} \\ &+\int\int p_g(\mathbf{x},y)\log(1-D(\mathbf{x},y))dyd\mathbf{x}.\end{aligned} \tag{3.12}$$

This function achieves the maximum at $\frac{p(\mathbf{x},y)}{p(\mathbf{x},y)+p_g(\mathbf{x},y)}$. □

**Theorem 1.** *For given $D^*$, the equilibrium of $\mathcal{U}(C,D,G)$ is achieved if and only if*

$$p(\mathbf{x},y) = p_g(\mathbf{x},y) = p_c(\mathbf{x},y) = p_c(G(\mathbf{z}|y),y). \tag{3.13}$$

**Proof.** Given $D^*$, we can reformulate the minimax game with value function

[1] https://srome.github.io/An-Annotated-Proof-of-Generative-Adversarial-Networks-with-Implementation-Notes/

$\mathcal{U}_g(D,G)$ as

$$\mathcal{U}_g(D,G) = \int\int p(\mathbf{x},y)\log\frac{p(\mathbf{x},y)}{p(\mathbf{x},y)+p_g(\mathbf{x},y)}dyd\mathbf{x} + \int\int p_g(\mathbf{x},y)\log\frac{p_g(\mathbf{x},y)}{p(\mathbf{x},y)+p_g(\mathbf{x},y)}dyd\mathbf{x}. \tag{3.14}$$

Following the proof in GAN, $\mathcal{U}_g(D,G)$ can be rewritten as

$$\mathcal{U}_g(D,G) = -\log 4 + 2JSD(p(\mathbf{x},y)||p_g(\mathbf{x},y)), \tag{3.15}$$

where $JSD$ is the Jensen-Shannon divergence. In addition, according to the definition of Kullback–Leibler(KL) divergence, $\mathcal{U}_c(C,G)$ can be rewritten as

$$\begin{aligned}\mathcal{U}_c(C,G) &= \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}\left[-\log p_c(y|\mathbf{x})\right] &(3.16)\\ &+ \mathbb{E}_{(\mathbf{x},y)\sim p_g(\mathbf{x},y)}\left[-\log p_c(y|G(\mathbf{z}|y))\right]\\ &= D_{KL}(p(\mathbf{x},y)||p_c(\mathbf{x},y)) + H_p(y|\mathbf{x})\\ &+ D_{KL}(p_g(\mathbf{x},y)||p_c(G(\mathbf{z}|y),y)) + H_{p_g}(y|G(\mathbf{z}|y)). &(3.17)\end{aligned}$$

From (3.15) and (3.17), $\mathcal{U}(C,D,G)$ becomes

$$\begin{aligned}\mathcal{U}(C,D,G) =& \mathcal{U}_g(D,G) + \mathcal{U}_c(C,G) &(3.18)\\ =& 2JSD(p(\mathbf{x},y)||p_g(\mathbf{x},y)) + D_{KL}(p(\mathbf{x},y)||p_c(\mathbf{x},y))\\ &+ D_{KL}(p_g(\mathbf{x},y)||p_c(G(\mathbf{z}|y),y))\\ &+ \left(H_p(y|\mathbf{x}) + H_{p_g}(y|G(\mathbf{z}|y)) - \log 4\right). &(3.19)\end{aligned}$$

Since $JSD(\cdot)$ and $D_{KL}(\cdot)$ are non-negative, their minimum values become zero if and only if $p(\mathbf{x},y) = p_g(\mathbf{x},y)$, $p(\mathbf{x},y) = p_c(\mathbf{x},y)$, and $p_g(\mathbf{x},y) = p_c(G(\mathbf{z}|y),y)$. Hence, the equilibrium of $\mathcal{U}(C,D,G)$ become $p(\mathbf{x},y) = p_g(\mathbf{x},y) = p_c(\mathbf{x},y) = p_c(G(\mathbf{z}|y),y)$. □

# Appendix 2

In this appendix, we provide the results for other binary combinations of CIFAR10. As we discussed in section 4.1.1, we chose two very similar classes to make classification difficult. As shown in Table 3.3, the proposed method consistently outperforms the existing methods on other combinations of CIFAR10.

Table 3.3: Test set performance of other CIFAR10 binary combinations.

| Methods | | Cat vs. Dog | | | Horse vs. Deer | | | Airplane vs. Ship | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AUPR | AUROC | F1 | AUPR | AUROC | F1 | AUPR | AUROC | F1 |
| Classifier only | | 0.6214 | 0.6123 | 0.4690 | 0.8018 | 0.7864 | 0.5228 | 0.8487 | 0.8141 | 0.6029 |
| Conventional Data-level | SMOTE | 0.6038 | 0.6064 | 0.5496 | 0.8114 | 0.8040 | 0.4826 | 0.8427 | 0.8215 | 0.5500 |
| | B-SMOTE | 0.6010 | 0.6028 | 0.3419 | 0.8057 | 0.8050 | 0.4551 | 0.8615 | 0.8527 | 0.5741 |
| | ADASYN | 0.5877 | 0.6078 | 0.3950 | 0.8526 | 0.8414 | 0.5586 | 0.8577 | 0.8409 | 0.6484 |
| | CC | 0.6000 | 0.6087 | 0.5345 | 0.6376 | 0.6700 | 0.5815 | 0.6462 | 0.6159 | 0.4145 |
| | CNN | 0.5544 | 0.5595 | 0.3665 | 0.7454 | 0.7462 | 0.5072 | 0.8255 | 0.8083 | 0.5085 |
| | SMOTEENN | 0.6134 | 0.6122 | 0.5197 | 0.7943 | 0.8013 | **0.6263** | 0.8393 | 0.8354 | **0.7573** |
| Loss-based | CRL | 0.6847 | 0.6856 | 0.3344 | 0.8331 | 0.8371 | 0.4364 | 0.8447 | 0.8356 | 0.3454 |
| | Focal | 0.6116 | 0.6292 | 0.3820 | 0.8161 | 0.8105 | 0.5234 | 0.7744 | 0.7809 | 0.6606 |
| | MPL | 0.6217 | 0.6099 | 0.3506 | 0.7820 | 0.7738 | 0.3863 | 0.7848 | 0.7548 | 0.5318 |
| GAN-based | cGAN | 0.6707 | 0.6633 | 0.3333 | 0.7279 | 0.7188 | 0.3333 | 0.8068 | 0.7966 | 0.3833 |
| | BAGAN | 0.8356 | 0.8146 | 0.6293 | 0.8159 | 0.8181 | 0.5269 | 0.8356 | 0.8146 | 0.6293 |
| | TripleGAN | 0.6484 | 0.6380 | 0.3333 | 0.8462 | 0.8480 | 0.5645 | 0.8555 | 0.8465 | 0.5271 |
| | Proposed | **0.8497** | **0.8327** | **0.6986** | **0.9005** | **0.8953** | 0.6041 | **0.8996** | **0.8965** | 0.6320 |

# Chapter 4

# Application I: Dementia Diagnosis Data Learning

## 4.1 Introduction

Neuropsychological assessments are essential for early diagnosing dementia and monitoring progression of dementia in both clinical and research settings, in advance of high-cost neuroimaging-based diagnosis such as magnetic resonance imaging (MRI) and positron emission tomography (PET). However, the abundant information of neuropsychological batteries other than their conventional total and/or subscale scores are not optimally employed in diagnosing and/or subclassifying dementia. [47, 48, 49, 50]. In our previous works, we showed that a simple cognitive test such as a categorical verbal fluency test would provide an accurate diagnostic reference of dementia if we employed various response patterns in the test instead of its simple total score [51, 52]. In this regard, neuropsychological batteries that consist of multiple cognitive tests for evaluating multiple cognitive domains may improve the diagnostic accuracy of dementia considerably if we employ the response patterns of multiple cognitive tests together instead of conventional total and/or subscale scores.

In this paper, to develop a practical data mining framework overcoming

the issues raised in the previous works, we propose a deep learning based low-cost and high-accuracy diagnostic framework of dementia with the response profiles of the Korean Longitudinal Study on Cognitive Aging and Dementia Neuropsychological Battery (KLOSCAD-N). The framework includes design procedures on missing data imputation, input variable selection, and cascaded classifier design for cost effective classification. First, in contrast to the previous works discarding the missing data samples which lead to information loss, we introduce a missing data imputation procedure to increase the accuracy and robustness in data analysis. Second, to maximize the diagnostic performance, a deep neural networks (DNNs) architecture are designed and validated in comparison with the other well-known classifiers. Third, to prevent a degradation of classification performance arising from the useless or redundant variables, we suggest a procedure to check the existence of useless or redundant variables and prune them. Fourth, we design a two-stage classifier to reduce time and cost for diagnosis using KLOSCAD-N and MMSE.

## 4.2 Background

Recently, data mining has shown remarkable performance in various fields including the medical fields [53]. Data mining is an interdisciplinary field of statistics, machine learning, visualization, database systems, and so on [54]. It focuses on discovering new meaningful information from a large dataset and provides us the information as understandable structure [54]. Especially, deep learning has recently emerged owing to big data and high-performance computing power. The deep learning is capable of exploiting the unknown structure from data to discover good representation. Thanks to this representation learning, the deep learning has overcome previous limitations of conventional approaches. Furthermore, the deep learning made great contributions to major advances in diverse fields including bioinformatics and medicine [1, 55, 56, 57, 58, 59, 60]. As we discussed ahead, although a large number of neuropsychological assessment data have been accumulated, hidden patterns in the data are not fully analyzed yet. To analyze the neuropsychological assessment data, the data mining using deep learning techniques can be utilized as a suitable approach. Mani et al. [61] first applied the data mining approach to neuropsychological assessment data, but simple classifiers were used to show the possibility of data mining application to neuropsychological data. Leighty [62] and Maroco et al. [63] provided the useful comparison on applications of multiple machine learning classifiers to neuropsychological assessment data, but these research studies did not consider variable redundancy, which may cause the performance degradation arising from the curse of dimensionality. Lemos [64] applied variable selection algorithms to overcome the curse of dimensionality, but the approach just removed the data with missing values, which may lead to loss of information.

## 4.3 Methods

Figure 4.1 depicts the overall scheme of the proposed diagnostic framework which includes five steps: (1) acquisition of KLOSCAD-N response profiles, (2) imputation of missing variables, (3) design of DNNs and validation by comparing with other classifiers, (4) input variable selection based on mutual information, and (5) design of two-stage classification scheme via the combination of MMSE and KLOSCAD-N. This study was approved by the institutional review board of Seoul National University of Bundang Hospital. The details of each step are provided in the following.

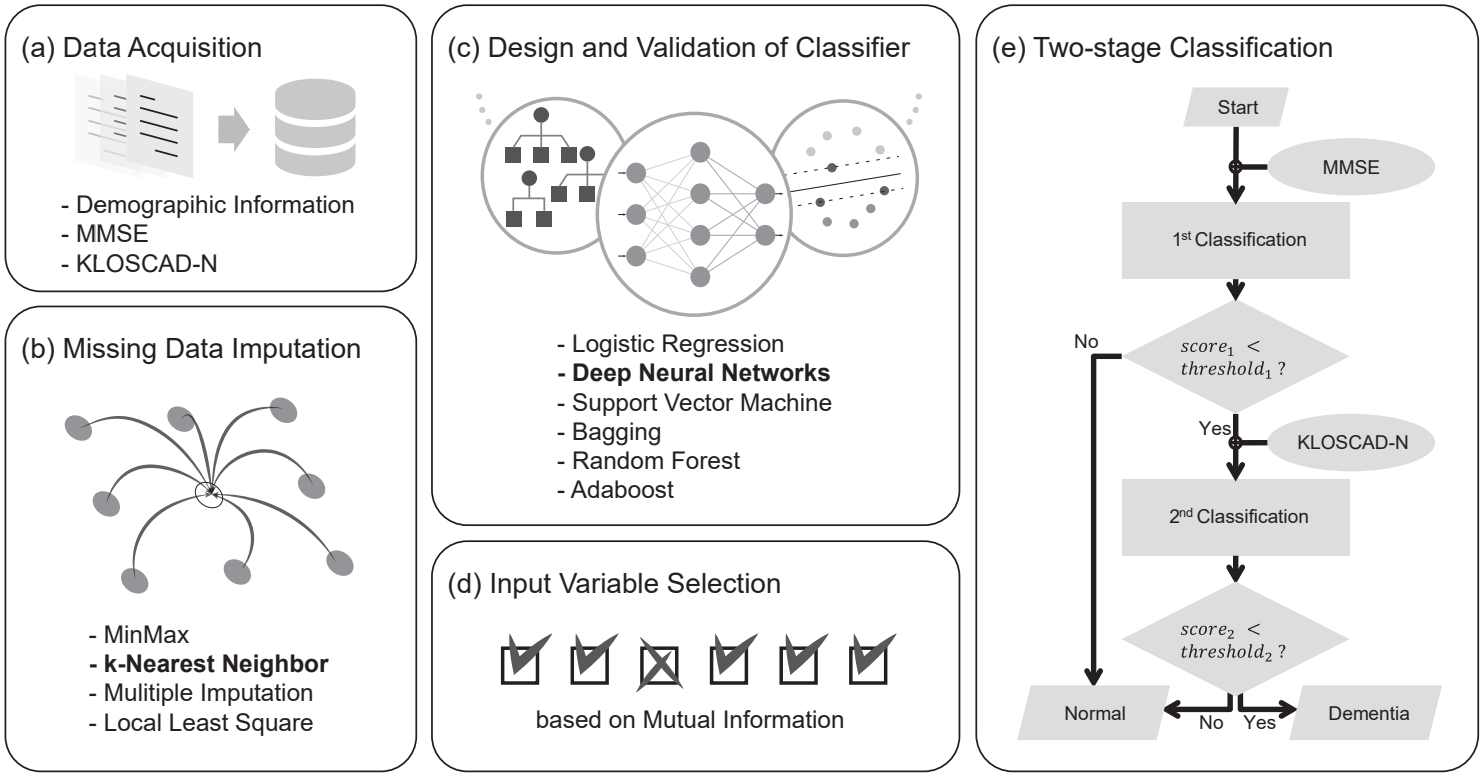



Figure 4.1: Overall scheme of the diagnostic framework. The proposed diagnostic framework includes five steps.

Table 4.1: Characteristics of the subjects.

| | Controls | Dementia | | Statistics | |
|---|---|---|---|---|---|
| | | CDR=0.5 | CDR=1 | For $X^2$ | post hoc$^{\ddagger}$ |
| Number | 2666 | 189 | 246 | | |
| Age (years) | $69.54 \pm 6.52^a$ | $75.01 \pm 7.23^b$ | $76.61 \pm 7.43^b$ | $174.927^{***}$ | $a < b$ |
| Sex (female, %) | 53.2 | 56.6 | 65.4 | $20.138^{**}$ | |
| Education (years) | $9.57 \pm 5.33^a$ | $8.40 \pm 5.75^b$ | $6.61 \pm 5.75^c$ | $30.520^{**}$ | $a > b > c$ |

$^{***}p < .001$, $^{**}p < .001$, $^{\ddagger}$Games-Howell post hoc comparisons

### 4.3.1 Subjects

We analyzed the KLOSCAD-N response profiles of 2,666 cognitively normal elderly (CNE) individuals and 435 dementia patients. The CNE individuals were the participants of the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD), which is a community-based longitudinal study of cognitive aging and dementia of community-dwelling Korean elderly cohort [65]. The dementia patients were either participant of the KLOCSCAD or visitors to the 14 dementia clinics that participated in the KLOSCAD. All subjects were 60 years or older. We excluded subjects with major axis I psychiatric disorders, such as major depressive disorder, and those who had serious medical or neurological disorders that could affect cognitive functions. The demographic and clinical characteristics of the subjects are summarized in Table 4.1. The 20% of subjects were randomly chosen as a test dataset for evaluating the proposed framework. The test dataset was not used in any of training procedure. Using the remaining 80% of subjects as a train dataset, we carried out five-fold cross-validation for training and model selection.

### 4.3.2 Diagnostic Assessments

Research neuropsychiatrists evaluated each subject using a standardized clinical interview, physical and neurological examinations, and laboratory tests according to the protocol of the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease Assessment Packet (CERAD-K) [66] and the

Mini International Neuropsychiatric Interview (MINI) version 5.0 [67]. When dementia was suspected, brain computerized tomography (CT) or magnetic resonance imaging (MRI) was also performed. The subjects diagnosed as having dementia according to the criteria of the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) (American Psychiatric Association 1994) were enrolled in the dementia group. The global severity of dementia was determined according to the Clinical Dementia Rating (CDR) [68].

### 4.3.3 Neuropsychological Assessments

Trained research neuropsychologists who were blind to the diagnosis of the subjects administered the KLOSCAD-N to each subject. The KLOSCAD-N consists of the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease Assessment Neuropsychological Battery (CERAD-N) [69, 66], Digit Span Test (DST) [70], Frontal Assessment Battery (FAB) [71], and Executive Clock Drawing (CLOX) [72]. The CERAD-N consists of nine neuropsychological tests: Categorical Verbal Fluency Test (CVFT), 15-item Boston Naming Test (BNT15), MMSE, Word List Memory Test (WLMT), Constructional Praxis Test (CPT), Word List Recall Test (WLRT), Word List Recognition Test (WLRCT), Constructional Recall Test (CRT), and Trail Making Test A and B (TMT-A and TMT-B). Conventionally, test scores of the nine neuropsychological tests were used to ascertain the presence of cognitive impairment objectively in diagnosing dementia and monitor the progress of cognitive impairment objectively with advancing dementia.

### 4.3.4 Missing Data Imputation

Inputs with missing values is unable to apply most of supervised machine learning models including deep learning. On the other hand, since the missing values often appear in neuropsychological tests, it is necessary to make up the

missing values in order to apply the model to the subjects having the missing values. Among the 3,101 samples of KLOSCAD-N response profiles, 75 have at least one missing value. Samples with one or two missing values are most frequent. CLOX1 and CLOX2 scores have the most frequent missing values. We have implemented four imputation methods: minimum-maximum (MinMax) imputation, k-nearest-neighbor (kNN) imputation [73], multiple imputations (MI) (Schafer 1999), and local least squares (LLS) imputation [74].

First, the MinMax imputation method is based on the assumption that the missing is caused by the subject's deficiency. The missing values are imputed according to the correlation between variables and labels. If the correlation is positive (or negative), the missing value is imputed with the maximum (or minimum) value of the variable. Second, the kNN imputation method attributes the missing values using the information of other subjects with a similar pattern in that sense of the nearest neighbor. After finding k number of neighbors, the imputation value is computed by averaging the values of those neighbors. In this study, Euclidean distance is used, and k is set to 5 empirically via experiments. Third, the MI method provided by the SPSS software is the most popular method in statistics, which has been developed to solve a single imputation's underestimating problem. The missing values are replaced by averaging a number of complete datasets which are estimated by the Monte Carlo technique. Each estimated complete dataset is imputed by linear regression. Lastly, the LLS imputation method shows the best performance for the missing value estimation on microarray data [75]. After finding the top k number of relevant genes (variables) using Pearson correlation, the target gene and its missing value are obtained by a linear combination of those relevant genes through solving a least squares problem.

Each method is evaluated in two ways: direct evaluation via error computation and indirect evaluation via classification performance. The direct evaluation is to compute an error between the original value and the imputed

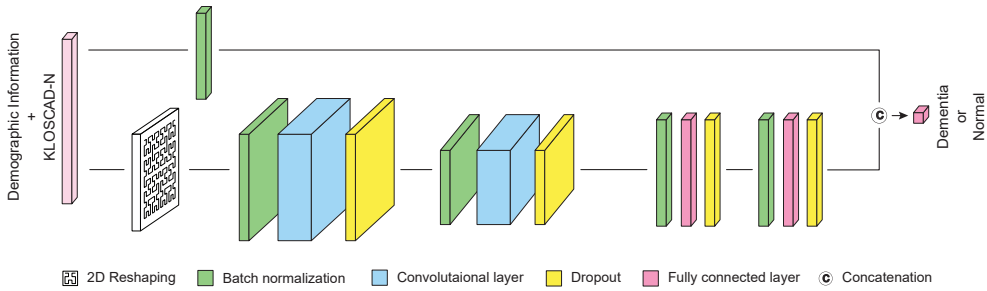



Figure 4.2: Architecture of proposed deep neural networks for KLOSCAD-N assessment and demographic information.

value. After we randomly generate artificial missing data from the complete data by considering the missing ratio in each variable, four kinds of imputation values for the artificial missing data are obtained through the four methods, respectively. The error between the original value and the estimated values is computed by matrix Euclidean norm. The indirect evaluation is to check a classification performance on imputed samples using the classifier trained with the complete data. By utilizing the four kinds of imputed samples generated by the four methods, respectively, we check which method shows the best classification performance by various classifiers.

### 4.3.5 Constructing Deep Learning Classifiers

Artificial neural network (ANN) is a computation model inspired by the biological brain. The hidden layer of ANN takes a role of feature extraction from input or lower hidden layer information. The responses in the hidden layer represent features extracted via a linear transformation of inputs and a nonlinear activation function. The DNN is a kind of ANN with deep hidden layers between the input and output layers. The deep layers composite the features from lower layers hierarchically, and learn complex data by associative memorizing through connection weights [76].

To construct a promising diagnosis framework, we design the DNNs for MMSE and KLOSCAD-N respectively. Since MMSE is composed of only five

dimensions (four demographic variables and one MMSE total-score), the fully-connected network (FCN) is enough to cover this simple classification problem. For KLOSCAD-N, we construct a two-dimensional convolutional neural network (2D-CNN) to achieve the best performance. As shown in Figure 4.2, we cascade a fully-connected layer following the convolutional layers. Also skip connection [77] is utilized to explicitly feed low level features to the output layers. In addition, we reshape the input into 2D image-like form with the Hilbert space-filing curve [78] which has been successfully used for DNA sequence classification with CNN [79]. Hilbert curve, which is shown in Figure 4.2, give a mapping 1D to 2D space that fairly well preserves locality. Since our data is a sequence of assessments followed by demographic information, continuity and clustering property of Hilbert curve would be appropriate for our data characteristics. To prevent an over-fitting, dropout [80], batch normalization [81] and early stop training technique is applied.

In this study, the ratio of the negative label samples to the positive label samples is approximately $9:1$ because the positive samples indicating the subjects of dementia are relatively rare compared to the negative samples indicating normal subjects. To solve this problem, the cost-sensitive loss is defined as (4.1) by multiplying a weight with the positive target.

$$l_c(y_i, \hat{y}_i) = -w_c y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i), \tag{4.1}$$

where $y_i$ is target value, $\hat{y}_i$ is predicted value, and $w_c$ is set for balancing as: $w_c = (\#\text{ of positive})/(\#\text{ of negative})$. Cost-sensitivity loss has the advantage to learn of minority data without additional computation. However, if there is no limit on the amount of computation, the other methodologies including the methodology proposed in Chapter 3 can be tried for the enhancement of performance. The trial will be conducted in the experiments.

To achieve the optimal architecture, we empirically evaluate the model with all combination of hyper-parameters as follows: the number of convolution

layer: $[1, 4]$, the number of filters: $32, 64, 128$, kernel size: $[2, 4]$, the number of fully connected layer: $[1, 4]$, and the number of hidden unit: $32, 64, 128$.

In addition, empirical evaluations are conducted for other architectures of DNNs such as 1D-CNN, fully-connected networks (FCN). Also, we compare with the transfer learning by adopting a pre-trained model (NasNet [82]) since NasNet is capable to handle low dimensional inputs unlike other networks for imagnet. Also, we compare our classifier with six well-known classifiers: XGBoost [83], Adaboost [84], Random Forest [85], Bagging [86], SVM [87], and Logistic Regression [88]. Hyper-parameters are empirically established through greedy search. Each algorithm is implemented by calling the java object of libSVM [89] and Weka [90] in MATLAB. To evaluate the generalization of each classifier, a five-fold cross validation on train dataset is applied. The area under curve (AUC) is used as the main evaluation metric.

### 4.3.6 Input Variable Selection

Since useless or redundant variables cause a degradation of classification performance due to a curse of dimensionality, it is necessary to check the existence of useless or redundant variables among KLOSCAD-N. Furthermore, by eliminating the redundant variables, the assessment time and monetary costs can be reduced. If there is a hierarchical property between variables, it is difficult to independently remove each variable. In this study, we thus do not consider subtotal variables that belong to the upper part of the hierarchical structure but use only the scores of the lowest-level variables. The relationships (or hierarchical properties) among the selected variables are then analyzed through the 2D-CNN.

For this purpose, we adopt the feature selection toolbox (FEAST) [91] which provides a computation toolbox of mutual information and other information theoretic functions. FEAST calculates the ranking of all variables by their contribution of information. In our work, we utilize eight functions

in FEAST: MIM, MRMR, CMIM, JMI, DISR, CIFE, ICAP, and CONDRED (see [91], the paper of FEAST toolbox, for details of each function). The ranking information of the eight functions is combined to determine the final ranking of each variable in an ensemble manner. For each variable, the eight ranking scores are averaged. The averaged ranking score is used to determine the ranking order of each variable.

Let $S_i, i = 1, ..., m$ be the variable set containing $i$ number of variables in ranking order. For example, $S_1$ only includes the highest ranked variable, and $S_5$ includes the variables from the first rank to the fifth rank. Then the classification performance is evaluated for each set $S_i$, and the set with the maximum performance is denoted by $S_{max}$. DeLong's test [92] is a statistical nonparametric approach to check whether two area under curve (AUC) values are having significant different. If the p-value from the test is less than 0.05, this indicates that the two sets show significant differences in AUC performance. Conversely, if the p-value is greater than 0.05, it can be judged that there is no significant loss of AUC performance between the two sets. Since the goal is to select the set with the lowest number of variables without loss of performance, we finally choose the set with the smallest number of variables from $S_i$ with p-value over 0.05.

### 4.3.7 Two-stage Classification

MMSE is the most popular screening test for dementia [65, 66, 93, 94]. MMSE is advantageous at low cost, but it is known to be less accurate than high-cost batteries such as KLOSCAD-N. Therefore, we propose a novel framework that combines the advantages of MMSE and KLOSCAD-N. In the first stage, MMSE is applied as a coarse screening test, and in the second stage, the KLOSCAD-N is administered for a fine diagnosis. If the candidate for KLOSCAD-N can be reduced through the first stage (MMSE) in advance without loss of diagnostic performance, a low-cost and high-performance di-

agnostic framework could be established.

The brief block diagram of the two-stage classification framework is shown in Figure 4.1-(d). The suggested framework has been established using the DNNs which showed the best performance among the other classifier on each test in the classifier comparison step. The MMSE total-score and demographic information are utilized to decide the further execution of the second stage, KLOSCAD-N, or not. By changing the threshold on the first-stage decision score to pass the subjects to the second-stage, we compute the cost and accuracy of the two-stage classification framework with test dataset. The cost is defined as

$$cost = n_{all} \times c_M + n_2 \times c_K, \tag{4.2}$$

where $c_M$ and $c_K$ is the cost per single subject of MMSE and KLOSCAD-N respectively, $n_{all}$ is the number of all subjects, and $n_2$ is the number of subjects who need the second-stage. Based on Korean insurance fees, the cost of each assessment per subject is approximately 10 USD and 180 USD for MMSE and KLOSCAD-N, respectively. We determine the best threshold on the decision score which shows the lowest cost while the performance does not show loss of classification performance.

Table 4.2: Classification performances on the imputed dataset indicated by the area under the receiver operator curve (AUC).

| | Proposed DNNs | XGBoost | Logistic Regression | Random Forest | Adaboost | Bagging | Support Vector Machine |
|---|---|---|---|---|---|---|---|
| MinMax | 0.9489 | 0.9506 | 0.9083 | 0.9405 | 0.9149 | 0.9334 | 0.8898 |
| kNN | **0.9603** | **0.9541** | **0.9356** | **0.9466** | **0.9444** | **0.9559** | 0.9321 |
| MI | 0.9586 | 0.9524 | 0.9312 | 0.9211 | 0.9184 | 0.9418 | 0.9347 |
| LLS | 0.9594 | 0.9471 | 0.9295 | 0.9343 | 0.9109 | 0.9339 | **0.9383** |

MinMax: minimum-maximum imputation, kNN: k nearest neighbor imputation, MI: multiple imputation, LLS: local least square imputation

## 4.4 Results

### 4.4.1 Missing Data Imputation

As suggested in the method section, the four imputation methods were evaluated via two ways, and the best imputation method was chosen. The first evaluation result (Euclidian norm) which gives the error between the original value and the imputed value was 1438.5621 for MinMax, 196.2499 for kNN, 255.7012 for MI, and 245.9988 for LLS. kNN had the smallest Euclidean error, whereas MinMax had the largest error. In consequence, kNN was evaluated to reconstruct the missing variable with the most similar value to the original one. Table 4.2 shows the result of the second evaluation approach, where the validity of imputed data had been evaluated by the classification performance tested via six classifiers trained with the complete data. Every classifier, except SVM, showed the best performance on kNN-based imputed data, whereas SVM showed the best performance on LLS. According to the result, kNN imputation method is chosen as the best one for the completion of missing values in KLOSCAD-N.

Table 4.3: Classification performances of various deep neural network architectures on Mini Mental Status Exam (MMSE) and Korean Longitudinal Study on Cognitive Aging and Dementia Neuropsychological Battery (KLOSCAD-N) indicated by the area under the receiver operator curve (AUC) via five-cross validation on train dataset.

| | | 2D-CNN | 2D-CNN Naïve | 2D-CNN w/o SC | 1D-CNN | 1D-CNN w/o SC | FCN | FCN w/o SC | NasNet |
|---|---|---|---|---|---|---|---|---|---|
| MMSE | mean | - | - | - | - | - | **0.9702** | 0.9583 | - |
| | std | - | - | - | - | - | **0.0144** | 0.0139 | - |
| KLOSCAD-N | mean | **0.9863** | 0.9850 | 0.9782 | 0.9848 | 0.9805 | 0.9830 | 0.9771 | 0.9813 |
| | std | **0.0048** | 0.0058 | 0.0057 | 0.0053 | 0.0042 | 0.0060 | 0.0070 | 0.0046 |

*Since MMSE is composed with only five dimensions (four demographic variables and one MMSE total-score, the other architectures are not applicable except FCN.

### 4.4.2 Classifier Validation

As we mentioned in the method section, hyper-parameters for every candidate model were searched via greedy search. The best FCN for MMSE is composed of one layer with 128 number of hidden units. The best 2D-CNN model for KLOSCAD-N is composed with two convolutional layers which contains 128 and 32 number of filters respectively with kernel size of 2, and two fully connected layers with 64 hidden units. Skip connection leads to a performance improvement over all structures. For 2D-CNN, our input reshaping method with Hilbert curve achieves higher performance than naïve reshaping method that simply stacks a sliced 1D input to form of 2D matrix (see the second column in Table 4.3).

Transfer learning with weights pretrained from imagenet (NasNet) has shown AUC value of 0.9813, which is smaller than those of the other networks trained with random initialization. This implies the pretrained information from imagenet datasets is not helpful to solve our problem. Table 4.3 shows the classification performance of various deep learning architectures from five-fold cross validation. For MMSE, the designed FCN in our work has AUC value of 0.9702. For KLOSCAD-N, the proposed architecture for 2D-CNN shows the best performance (AUC value of 0.9863) among all the candidate architectures.

Table 4.4: Comparative analysis with other conventional classifiers indicated by the area under the receiver operator curve (AUC) via five-cross validation on train dataset.

| | | Proposed DNNs | XGBoost | AdaBoost | Random Forest | Bagging | Support Vector Machine | Logistic Regression |
|---|---|---|---|---|---|---|---|---|
| MMSE | mean | **0.9702** | 0.9605 | 0.9573 | 0.9581 | 0.9631 | 0.9627 | 0.9642 |
| | std | **0.0144** | 0.0144 | 0.0171 | 0.0192 | 0.0169 | 0.0196 | 0.0171 |
| KLOSCAD-N | mean | **0.9863** | 0.9850 | 0.9774 | 0.9762 | 0.9724 | 0.9744 | 0.9807 |
| | std | **0.0048** | 0.0065 | 0.0107 | 0.0079 | 0.0069 | 0.0093 | 0.0080 |

Table 4.4 shows the classification performance of other type of classifiers. For both MMSE and KLOSCAD-N, the proposed DNNs show the best performance. It is known that the DNNs show inherently a good generalization capability, even its large number of parameters when trained with the sufficient number of train data samples. As a result, our dataset is enough to achieve reasonable performance for the both assessments using the designed DNNs.

Table 4.6 shows the comparative efficiency of the proposed two-stage classification in view of various metrics including the cost. As shown in the fourth and fifth columns, the existing works for KLOSCAD-N [51] and MMSE [52] do not show good performance relatively because they rely on the simple total score of KLOSCAD-N or MMSE. As shown in the first and third columns DNNs improves the accuracy with 2.90% for MMSE and 6.61% for KLOSCAD-N compared to the existing methods because it can utilize the hidden patterns of input variables (demographic information, subscale scores, and so on). As shown in the second column, the proposed two-stage classification framework shows the best efficiency through all evaluation metrics with a reasonable cost (Details are discussed in the following section on two stage classification).

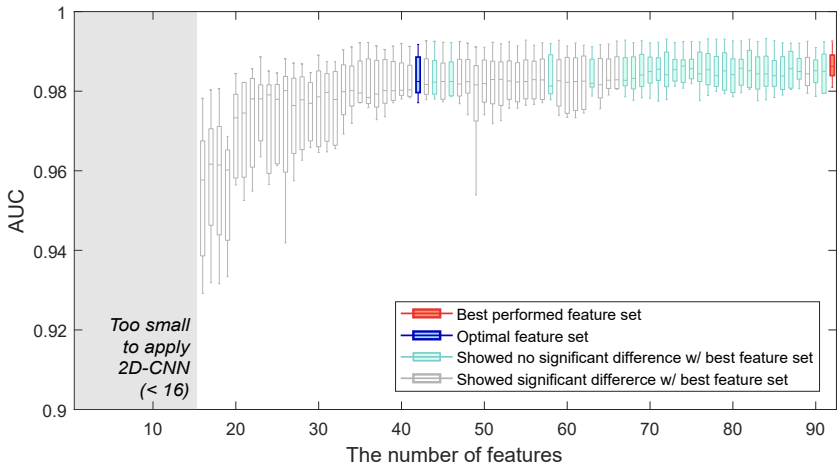



Figure 4.3: Trends of the area under the receiver operator curve (AUC) as a function of the number of variables included in order from the highest ranging variable.

### 4.4.3 Input Variable Selection

The final rankings of 92 input variables were yielded through the ensemble of eight methods for feature selection provided in FEAST. The performances on the input variable sets, $S_i, i = 1, ..., 92$, are shown in Figure 4.3. As shown in Figure 4.3, the performance increases as the variables are added one by one in order from the highest-ranking variable, but the degree of increase lessens after 30 variables and becomes saturated after 43 variables. The best performance was achieved with 92 variables which are depicted as red boxplot in Figure 4.3. Among $S_i$, we removed the variable set (gray boxplot) that showed a significant difference ($p < 0.05$ on DeLong's test) with the best-performed variable set $S_{max}$ (red boxplot). Among the remaining candidate variable set (blue boxplot and red boxplot), we chose the final variable set which contains the least number of variables. As a result, we could reduce the number of variables 92 to 43. The final variable set and variable ranking information is described in Table 4.5.

Table 4.5: Top 40 variables selected for classifying dementia from normal controls

| Ranking | Variable description |
|---|---|
| 1 | Time to complete the Trail Making Test A |
| 2 | Retention index of Constructional Recall Test[1] |
| 3 | Age |
| 4 | Response bias index of the Word List Recognition Test[2] |
| 5 | Recency index of the Word List Memory Test[3] |
| 6 | Executive Clock Drawing Test (CLOX) 1 score |
| 7 | Consistency index of the Word List Memory Test[4] |
| 8 | Correct responses at the second quarter (15-30 seconds) in the Verbal Fluency Test |
| 9 | The number of repetitive recalls in trial 3 of the Word List Memory Test |
| 10 | Geriatric Depression Scale score |
| 11 | Cube recall score of the Constructional Recall Test |
| 12 | Clustering index of Verbal Fluency Test |
| 13 | Correct responses in the middle-frequency objects of the 15-item Boston Naming Test without cues |
| 14 | The number of correct recall in trial 2 of the Word List Memory Test |
| 15 | Digit Span Test Forward score |
| 16 | Years of education |
| 17 | Perceptual error index in the low-frequency objects of the 15-item Boston Naming Test |
| 18 | Ineffective switch index of the Verbal Fluency Test |
| 19 | Retention index of the Word List Recall Test[5] |
| 20 | Consistency index of the Word List Recall Test[6] |
| 21 | Primacy index of the Word List Memory Test[7] |
| 22 | Word List Recall Test score |
| 23 | Switch index of the Verbal Fluency Test[8] |
| 24 | The number of correct recall in trial 1 of the Word List Memory Test |
| 25 | Forward span of the Digit Span Test |
| 26 | Word List Recognition Test total score |
| 27 | Correct responses in the low-frequency objects of the 15-item Boston Naming Test with phonemic cues |
| 28 | Learning curve of the Word List Memory Test[6] |
| 29 | Digit Span Test Backward score |
| 30 | Correct responses at the last quarter (45-60 seconds) in the Verbal Fluency Test |
| 31 | Constructional Recognition Test score |
| 32 | Go-No-Go score of the Frontal Assessment Battery |
| 33 | The umber of correct recall in trial 3 of the Word List Memory Test |
| 34 | Correct responses in the high-frequency objects of the 15-item Boston Naming Test without cues |
| 35 | Correct responses at the first quarter (0-15 seconds) in the Verbal Fluency Test |
| 36 | 'Do not know' responses in the low-frequency objects of the 15-item Boston Naming Test |
| 37 | The number of intrusion errors in the Word List Recall Test |
| 38 | Intersecting rectangles recall score of the Constructional Recall Test |
| 39 | Recency index in trial 1 of the Word List Memory Test |
| 40 | Correct responses at the third quarter (30-45 seconds) in the Verbal Fluency Test |
| 41 | Backward span of the Digit Span Test |
| 42 | Diamond recall score of the Constructional Recall Test |
| 43 | Cube score of the Constructional Praxis Test |

[1](Constructional recall test score/constructional praxis test)×100

[2](False positive score−false negative score)/(false positive score+false negative score)

[3](The number of recalled words among the last 3 words of the Word List Memory Test/Word List Memory Test score)×100

[4]The sum of the numbers of words consistently recalled in between trial 1, trial 2 and trial 3 of the Word List Memory Test

[5](Word List Recall Test total score/trial 3 score of Word List Memory Test)×100

[6](The number of words consistently recalled in the Word List Recall Test among the recalled words in trial 3 of the Word List Memory Test / the numbers of words recalled in trial 3 of the Word List Memory Test)×100

[7](The number of recalled words among the first 3 words of the Word List Memory Test/Word List Memory Test score)×100

[8]The number of switches between clusters during Verbal Fluency Test

Table 4.6: Comparative results of two-stage classification on test dataset

| | KLOSCAD-N w/ DNNs | Proposed Two-stage Classification | MMSE w/ DNNs | KLOSCAD-N w/o DNNs | MMSE w/o DNNs |
|---|---|---|---|---|---|
| Accuracy (%) | 92.74 | 92.90 | 87.74 | 86.13 | 84.84 |
| AUC | 0.9790 | -* | 0.9383 | 0.9349 | 0.9143 |
| F1 Score | 0.7805 | 0.7800 | 0.6667 | 0.6356 | 0.6179 |
| Sensitivity | 0.9287 | 0.9343 | 0.8780 | 0.8621 | 0.8736 |
| Specificity | 0.9195 | 0.8966 | 0.8736 | 0.8612 | 0.8443 |
| Likelihood Ratio Plus | 11.5425 | 9.0319 | 6.9446 | 6.2092 | 5.6097 |
| Likelihood Ratio Minus | 0.0775 | 0.0732 | 0.1396 | 0.1602 | 0.1498 |
| Positive Predictive Value | 0.5673 | 0.5064 | 0.4410 | 0.4136 | 0.3892 |
| Negative Predictive Value | 0.9913 | 0.9917 | 0.9844 | 0.9821 | 0.9833 |
| Pre Test Odd | 0.1136 | 0.1136 | 0.1136 | 0.1136 | 0.1136 |
| Post Test Odd | 1.3111 | 1.0259 | 0.7888 | 0.7053 | 0.6372 |
| Post Test Probability | 0.5673 | 0.5064 | 0.4410 | 0.4136 | 0.3892 |
| Cost† | $111,600 | $40,030 | $6,200 | $111,600 | $6,200 |

*Since each stage provides their own probability, single AUC value cannot be calculated.
†Total cost for test dataset including 620 subjects

### 4.4.4 Two-stage Classifications

Accordingly, at two-stage classification, performance and cost were evaluated by changing the threshold of the first stage classification on MMSE to pass subjects to the second stage (KLOSCAD-N). The results are shown in Figure 4.4. Figure 4.4-(a) shows a value of sensitivity and specificity as a function of threshold on the first classification. It is noted that the two curves meet at the threshold of 0.075, and the point is referred to as equal error rate (EER). Figure 4.4-(b) shows the trends of performance and cost in the threshold range $[0, 0.075]$. As shown in Figure 4.4-(b), the higher threshold (fewer subjects take KLOSCAD-N) leads to the less performance and cost. On certain the threshold, f1 scores are smaller than that of when the threshold is zero. In conclusion, at threshold equal to 0.0362, the proposed framework saves as much as cost without loss of performance. The second column in Table 4.6 is the final performance of the proposed two-stage classification. As a result of the proposed combination of MMSE and KLOSCAD-N, the cost is reduced by 64.13% without loss of accuracy compared to the case that every subject takes KLOSCAD-N (the first column in Table 4.6).

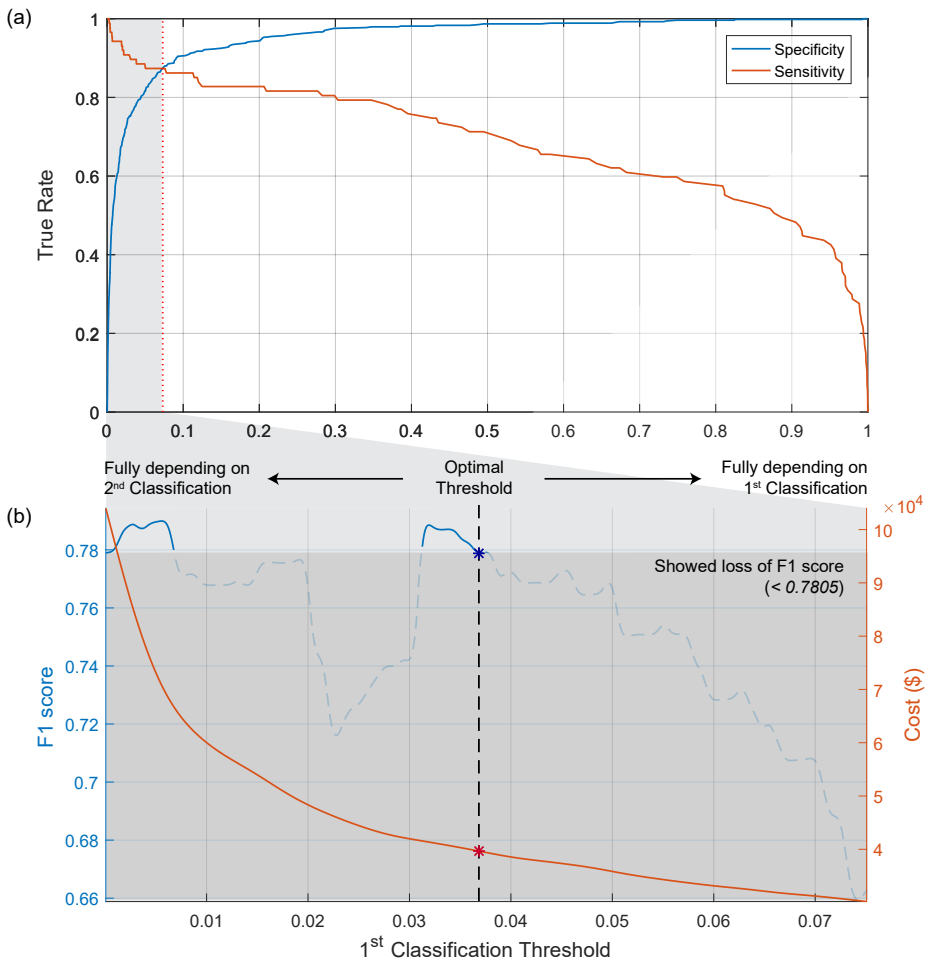



Figure 4.4: Two-stage classification performance trends as function of a sweeping threshold of deep neural networks (DNNs) with MMSE for the second-stage diagnosis with Korean Longitudinal Study on Cognitive Aging and Dementia Neuropsychological Battery. (a) Equal error rate (EER) curve on DNNs for MMSE. (b) Empirically estimated performance and cost on test dataset. When first-stage classification threshold value is 0.0362, cost is minimized without any loss on performance (f1 score).

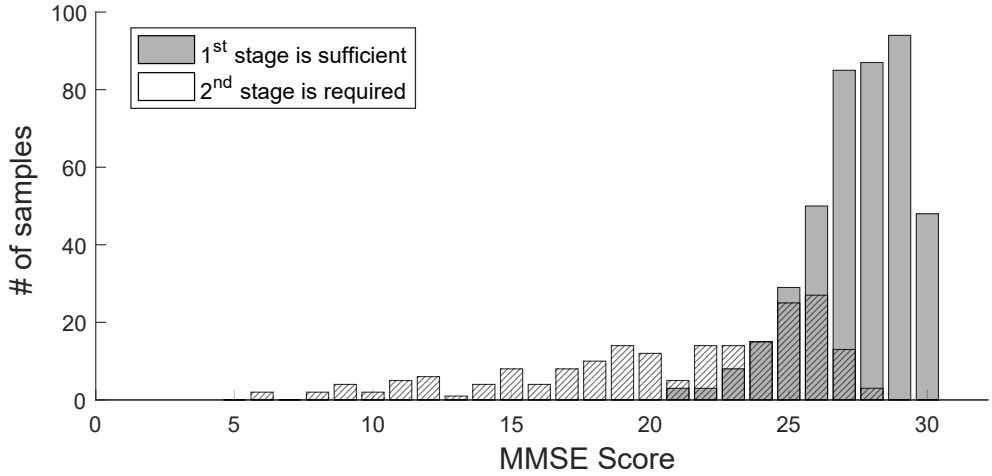



Figure 4.5: The distribution of the MMSE scores of the test set subjects requiring only first-stage and those requiring two-stages. The two distributions are roughly divided around 25 points, but can not be clearly distinguished only by the MMSE score.

Figure 4.5 is the histogram distribution of the MMSE scores of the test dataset subjects. Subjects that require only first-stage are represented by hatched bars and are represented by shaded bars that require a second-stage. Two groups are roughly divided by point 26, but there are still overlapping parts. The existence of overlapping means that the MMSE score alone cannot make a clear diagnosis. In other words, in order to judge whether or not to take the second-stage more clearly, it is necessary to use the designed DNNs.

Table 4.7: Comparison of various imbalanced learning techniques on dementia diagnose

| | 5-fold Cross Validation | | | | |
|---|---|---|---|---|---|
| Model | AUPR | AUROC | F1 | Sensitivity | Specificity |
| Classifier Only | 0.9413 | 0.9840 | 0.8677 | 0.8448 | **0.9831** |
| Cost-sensitive Loss | 0.9433 | 0.9843 | 0.8354 | 0.9368 | 0.9498 |
| SMOTE | 0.9320 | 0.9803 | 0.4713 | 0.9885 | 0.6250 |
| Borderline SMOTE | 0.9122 | 0.9741 | 0.5090 | 0.9828 | 0.6812 |
| ADASYN | 0.9159 | 0.9756 | 0.3699 | **0.9943** | 0.4253 |
| Cluster Centroids | 0.9059 | 0.9712 | 0.8389 | 0.8393 | 0.9733 |
| Condensed Nearest Neighbor | 0.9350 | 0.9819 | 0.8366 | 0.9224 | 0.9536 |
| SMOTE ENN | 0.9314 | 0.9809 | 0.4976 | 0.9885 | 0.6610 |
| Proposed in Chapter 3 | **0.9482** | **0.9846** | **0.8780** | 0.8766 | 0.9803 |
| | Test Set | | | | |
| Model | AUPR | AUROC | F1 | Sensitivity | Specificity |
| Classifier Only | 0.9260 | 0.9741 | 0.8449 | 0.8138 | 0.9816 |
| Cost-sensitive Loss | 0.9235 | 0.9734 | 0.7936 | 0.9126 | 0.9366 |
| SMOTE | 0.9183 | 0.9712 | 0.4515 | 0.9770 | 0.6011 |
| Borderline SMOTE | 0.9083 | 0.9679 | 0.4910 | 0.9724 | 0.6642 |
| ADASYN | 0.9045 | 0.9661 | 0.3637 | **0.9885** | 0.4199 |
| Cluster Centroids | 0.8830 | 0.9643 | 0.8050 | 0.7885 | 0.9722 |
| Condensed Nearest Neighbor | 0.9099 | 0.9702 | 0.8063 | 0.8851 | 0.9490 |
| SMOTE ENN | 0.9049 | 0.9694 | 0.4847 | 0.9770 | 0.6514 |
| Proposed in Chapter 3 | **0.9330** | **0.9761** | **0.8591** | 0.8414 | 0.9809 |

### 4.4.5 Imbalanced Data Classifications

Table 4.7 shows the results of applying various imbalanced learning techniques to dementia diagnosis. When applying the classifier only, it is possible to confirm the imbalanced learning tendency is biased since the sensitivity value is much lower than the specificity value. As a result of applying the cost-sensitive loss, the sensitivity value is increased with the slight decrease of specificity value, which means that learning is more balanced than classifier without cost-sensitive loss. Data-level techniques increase the sensitivity but reduce the specificity largely. As a result, the AUPR, AUROC, and F1 value was adversely affected. This implies that the generated data by the conventional data-level techniques are not enough to represent the actual data distribution. However, the methodology proposed in Chapter 3 achieves the highest records in AURP, AUROC, and F1 which represent the performance in consideration of both classes, whereas the sensitivity is improved without loss of specificity.

## 4.5 Discussion

Comprehensive neuropsychological assessments, in spite of their variety and abundance of information, have not been optimally employed for diagnosing and/or subclassifying dementia by their conventional total and/or subscale scores. In the current study, we developed a low-cost high-accuracy diagnostic framework for diagnosing dementia using a comprehensive neuropsychological battery that includes MMSE. The proposed framework proceeds through four steps: missing data imputation, classifier validation, input variable selection, and two-stage classifications.

Although neuropsychological batteries can provide useful diagnostic information (such as reaction patterns and inter-correlations among them), only overall performance (such as total scores or subscale scores) has been quantified so far in both clinical and research settings. Even if we simultaneously used data from multiple cognitive tests, we could not have improved the diagnostic accuracy for dementia if we had used only the overall performance of each test. For example, Seo et al. [48] proposed the total score of CERAD-N (CERAD-TS), which was a simple sum of multiple cognitive test scores included in the CERAD-N. However, the diagnostic accuracy of the CERAD-TS for dementia was only approximately 3% higher than that of MMSE in a given population.

In our previous work, we showed that the reaction patterns of cognitive tests may provide better performance in diagnostic dementia than simple total scores of the tests [51, 52]. For example, patients with Alzheimer's showed impaired knowledge-based semantic associations compared with the cognitively normal elderlies who had the same overall performance in the categorical verbal fluency test as the Alzheimer's disease patients [51]. In addition, we showed that we could improve the diagnostic accuracy for dementia of categorical verbal fluency tests by approximately 10% if we used reaction patterns in the test instead of the total score of the test [52].

Therefore, we may improve the diagnostic accuracy for dementia if we can

use the hidden patterns of responses in the multiple cognitive tests included in neuropsychological batteries simultaneously. Data mining approaches have shown remarkable performance in discovering new meaningful information from large datasets and summarizing the information in understandable structure [54]. As we discussed earlier, although a large amount of neuropsychological assessment data has been accumulated, hidden patterns in the data have not been fully analyzed yet. The proposed framework achieved better improvements in diagnostic performance than the CERAD-TS [48] as shown in the fourth column in Table 4.6. The improvement compared with CERAD-TS was +6.61% for accuracy, 0.044 for AUC, and +0.14 for f1 score.

There were some studies to improve screening accuracy for dementia with MMSE by supplementing other brief cognitive test scores [95] or informant questionnaires [96]. However, it has never been studied whether and how much the supplementation of comprehensive neuropsychological batteries can improve diagnostic accuracy for dementia. To the best of our knowledge, our methodology is the first approach that cascades the screening test (MMSE) and the neuropsychological battery (CLOSCAD-N) for diagnosing dementia.

The proposed framework is effective in three aspects. First, by the proposed two-stage classification approach, 71,570 USD (64.13%) of the cost for 620 subjects was evaluated to be saved without loss of classification performance. Second, through the variable selection step, it was confirmed that only a small number of KLOSCAD-N variables with 2D-CNN achieved higher performance than the full number of variables. This implies that it is possible to develop more compact assessments with saving time and monetary cost. Third, the proposed framework will be implemented and distributed as a form of software. Non-expert will also be able to obtain additional information about the diagnosis of dementia in addition to the total score by entering the results of the neuropsychological tests into the software. It is expected that the social cost for the overall diagnosis of dementia can be reduced by increasing

the usefulness of clinical neuropsychological tests and the possibility of early diagnosis of dementia.

Regarding the limitation of our framework, the diagnosis only focuses on a binary classification problem (normal versus dementia). As for future works, the proposed framework can be extended to a multi-class classification problem such as dementia progress classification (normal versus mild cognition impairment verses dementia) or dementia type classification (Alzheimer's disease versus vascular dementia versus dementia with Lewy bodies, and so on). However, neuropsychological assessments alone may not be enough to diagnose specific dementia types. In fact, to diagnose the specific dementia types, neuroimaging techniques (MRI and PET) and genetic analysis are performed. Cascading these advanced tests as the next stage of the proposed two-stage classification will further enhance the advantages that we have gained in this study. Another limitation of this study is that the proposed framework cannot explain the hidden patterns learned by DNNs because of the black-box property of deep learning. However, the field of explainable artificial intelligence is being actively studied for visualizing these hidden patterns in nowadays [97]. For the future work, it will be possible to specify meaningful patterns to clinicians through explainable artificial intelligence methodology.

# Chapter 5

# Application II: Drowsiness EEG Data Learning

## 5.1 Introduction

A sufficient amount and good-quality sleep are directly related to cognitive function. Surprisingly, however, over 30% of adults are chronically sleep-deprived, sleeping less than seven hours [98, 99, 100, 101, 102]. They commonly suffer from major sleep disorders such as insomnia or sleep apnea, which usually results in extreme daytime drowsiness[101, 103, 104]. The extreme daytime drowsiness is associated with lowered attention, which causing considerable socioeconomic burden to the community[105, 106, 107]. It hinders work productivity[108], lowers academic achievements[109, 110], and increases the risks of traffic or workplace accidents[111, 112]. Therefore, a virtuous cycle system that can monitor drowsiness or attention and provide proper feedback is of vital importance both for improving work efficiency and for the safety of our society.

We can categorize previous drowsiness detection approaches into two types, *i.e.*, task performance and biosignal-based methods. One representative example of the task performance-based method is vehicle motion (VM) monitor-

ing [113, 114, 115, 116]. However, its usage is restricted to a driving condition only. Another representative method is psychomotor vigilance task (PVT)[117, 118, 119, 120], where subjects are asked to respond to certain stimuli (visual or audio) as fast as possible. However, the limitation of PVT is that the subject must stop their ongoing task. Electrooculography (EOG) monitoring [116, 121, 122, 123, 124, 125] is one of the biosignal-based methods. It tracks eye movements or blinking patterns (*e.g.* frequency and speed) to detect drowsiness using a video camera or an infrared device. The limitation of the EOG-based approaches is that they only capture secondary eye responses caused by the homeostatic or circadian sleep drive in the central nervous system. Inevitably they have a low temporal resolution, and are influenced by environmental factors such as wind, temperature, and humidity.

Current algorithms for drowsiness detection are mostly based on electroencephalography (EEG) monitoring. These studies measure electrical activities of the brain associated with drowsiness, extract meaningful features from EEG, and use a classifier to distinguish various states of one's alertness and drowsiness[126, 127, 128]. Some previous works have adopted machine learning-based classifiers and have shown promising classification results[129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140]. However, they mainly focused on classifying long-term states such as conditions before or after driving, and they fail to capture more *instantaneous drowsiness* (*i.e.* lapse) which may cause severe accidents.

In this work, to address the limitations of previous works, we propose a novel framework to detect *instantaneous drowsiness* using a short time segment ($\sim$ 2 seconds) of EEG. The main contributions of this paper, as an extension of our previous work [141], are as follows:

- We define a novel phenotype labeling method for *instantaneous drowsiness* by combining the advantages of PVT (as a standard reference) and EOG monitoring (as a task-independent measurement).

- We propose a novel framework outperforming the previous approaches by adopting multitaper power spectral density (MPSD) for EEG feature extraction and extreme gradient boosting (XGBoost) as a classifier.
- We identify key frequency bands and channels for drowsiness detection.
- We demonstrate that using only seven channels (Fp1, Fp1, T3, T4, O1, O2, and ECG) can provide comparable performance to using the original twenty with less than 2% accuracy degradation.
- We verify the applicability of the proposed framework for a mobile environment by using a wireless EEG with dry-sensors as well as a wired EEG with wet-sensors.

## 5.2 Background

### 5.2.1 Task Performance-based Drowsiness Detection Methods

The VM monitoring module [113, 114, 115, 116], and the PVT [117, 118, 119] determine the drowsiness level by measuring task performance. The VM monitoring modules receives input from driving directions and from changes in lane-keeping. It is vulnerable to weather, road condition, and vehicle type. The PVT is a well-established measure of alertness or sustained attention. For the standard ten-minutes PVT, the subjects are required to click a button with the dominant hand's thumb as soon as possible when the visual signals are presented in random intervals (2–10 seconds) [117, 120]. Response time is a validated indicator of the alertness level. The PVT can measure changes in alertness caused by sleep disorders and deprivation [117, 118, 119, 120]. Sleep deprivation leads to a fluctuation in sustained attention because of the interaction of involuntary sleep-initiating and counteracting wake-maintaining systems, thus resulting in lapses (errors of omission to respond for given stimuli). However, to complete the given task, subjects need to interrupt the ongoing task, thus hindering the work-continuity. In this study, the PC-PVT platform (Biotechnology High Performance Computing Software Applications Institute, `http://bhsai.org/software/pcpvt`, MD, USA) is adopted [142] (Section 5.3.3). Both PC-PVT and standard PVT have the same functionality. PC-PVT uses a personal computer, whereas PVT uses a specific hardware. The PC-PVT data are used to label drowsiness for EEG segments. The labeled data are then utilized to train the supervised classification-based drowsiness detection model.

### 5.2.2 EOG-based Drowsiness Detection Methods

EOG-based drowsiness detection uses a video camera to analyze eye movement markers such as speed, frequency, blinking, and winding. Daytime drowsiness

is closely associated with several ocular parameters (*e.g,* slow movements, increased closure time, and increased blinking frequency) [116, 121, 123, 124, 125]. R100 (Phasya, Belgium) is a recently developed EOG-based method [143]. The R100 exploits glasses equipped with a high-speed camera to sense eye and eyelid movements. It can continuously monitor the level of drowsiness with minimal or no disruption of the ongoing task, and provide a task-independent measure of drowsiness. In this study, we use the R100 device to acquire a long length of drowsiness information. The obtained information is used to define the drowsiness state label discussed in Section 5.3.3.

### 5.2.3 EEG-based Drowsiness Detection Methods

EEG changes are closely related to alertness fluctuations. Regarding drowsiness, Putilov *et al.* [144] reported that when the subjects' eyes are open, the power of the alpha and theta bands increase, and with eye closure only the theta band power increases. Alloway *et al.* [145] reported that the alpha band power increases when the eyes are open but decreases when they close.

In the previous studies, the drowsiness label was variously defined. The methods for labeling can be categorized into three types. The first category is a questionnaire-based method such as the Epworth [146] and Karolinska sleepiness scale [144]. A simple questionnaire such as "how sleepy are you now?" is given to the subject. The output is hardly objective. The second approach is task performance evaluation that uses VM detection modules [147, 135, 129, 148, 130] or response time to specific stimuli [134, 148, 149]. The third type is EOG-based method. Gang *et al.* [150] defined the label with EOG information but only used the parameter of eyelid closure ratio.

To the best of our knowledge, our drowsiness labeling method is the first approach that utilizes both R100 and PVT. R100 provides a task-independent and long-term measurement of drowsiness, and PVT sets the references for extreme sleepiness.

### 5.2.4 Feature Extraction for EEG Signals

Many types of feature extraction techniques have been proposed to retrieve useful EEG information and apply it to drowsiness detection studies. The most widely used feature is band power (BP). Especially theta ($4-8$ Hz) and alpha ($8-16$ Hz) bands have played a key role [134, 135, 151, 152]. To achieve more detailed frequency information than BP, some studies [147, 138, 149] have used power spectral density (PSD) estimation or wavelet decomposition. However, the PSD-based approaches suffer from high bias and variance in estimation. A single-taper PSD (SPSD) method was developed to address high bias issues [153]. Nonetheless, the SPSD technique was not able to fix the high variance in estimation. To solve this issue, a multitaper PSD (MPSD) method has been developed [154]. MPSD allows precise estimation by aggregating multiple independent SPSDs, thus outperforming SPSD for estimating the sleep stages [155]. To the best of our knowledge, this is the first study that applies MPSD for detecting drowsiness with EEG.

### 5.2.5 Machine Learning Methods for Drowsiness Detection

Many studies based on machine learning techniques have been proposed for detecting drowsiness. Linear regression was first applied for drowsiness detection [134]. Since then, several studies have used artificial neural networks (ANNs) [138, 149, 131, 148] and support vector machines (SVMs) [130, 151, 148]. In our study, we use the extreme gradient boosting (XGBoost) method [83], which offers the following benefits: ease of use, scalability, accuracy, and computational efficiency. In addition, it has good records on recent machine learning competitions [156]. Furthermore, we can estimate the importance of each input feature by using the branch gain in XGBoost. This information can suggest which EEG frequency bands and channels are relevant to the drowsiness level, and guide the target of the electrode application and the frequency band analysis, thereby enhancing performance and reducing resource inputs.

## 5.3 Methods

As depicted in Figure 5.1, the proposed framework includes four steps: data acquisition, feature extraction, drowsiness labeling, and drowsiness detection. For the first step, data acquisition, we acquire the EEG signals from the recruited subjects. Furthermore, we perform an R100 and a PC-PVT (see section 5.3.1). As the second step, we extract the MPSD [154] feature from the preprocessed EEG segments (see section 5.3.2). In the next step, the instantaneous drowsiness label is defined for each EEG segment using the R100 ensemble outputs, and is validated by comparing lapse information from PC-PVT (see section 5.3.3). For the last step, XGBoost [83] is trained with the features and labels which are respectively acquired from previous steps. The trained XGBoost decides whether a given EEG segment is acquired during a drowsiness condition or not (see Section 5.3.4).

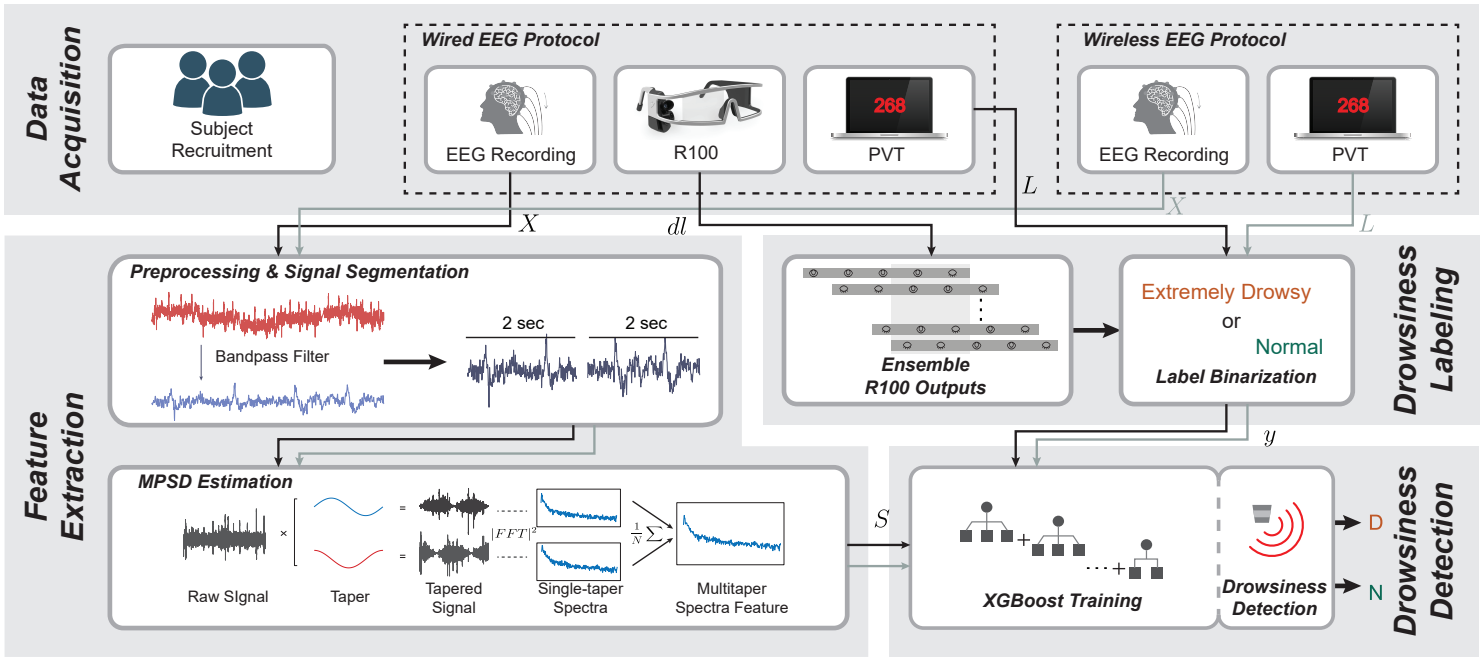



Figure 5.1: The Framework is composed of four main steps: data acquisition, feature extraction, drowsiness labeling, and drowsiness detection. For wireless EEG protocol, the drowsiness level was evaluated only with PVT (please see Section 5.3.5 for the details).

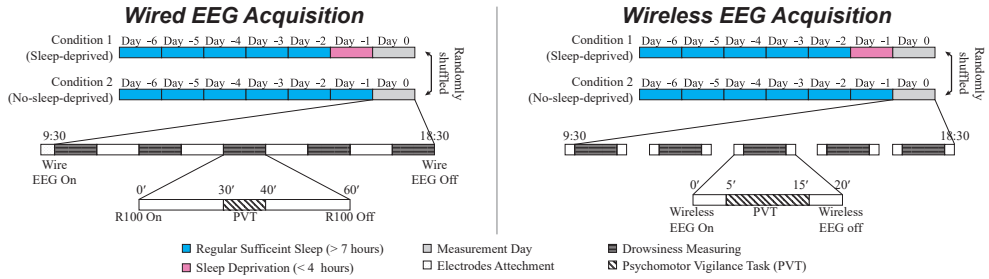



Figure 5.2: Illustration of two EEG acquisition protocol (wired and wireless). There are two random ordered conditions (sleep-deprived and no-sleep deprived) for each protocol. While the wired EEG device is worn for an entire, the wireless EEG device was only worn during the PVT task.

### 5.3.1 Data Acquisitions

#### Subjects

We recruited eight healthy subjects (4 men and 4 women; age, 26.8±3.4 years old; body mass index, 20.9±2.1 kg/m$^2$) who were nonsmokers and did not have any neurologic diseases and sleep disorders such as obstructive sleep apnea, insomnia, primary hypersomnia, or restless legs syndrome. None took any medication affecting sleep or alertness, and they reported no involvement in shift work for the previous one-year period and no travel to an area with a different time zone during the month before enrollment. All participants had intermediate chronotype, and slept regularly, sleeping between six and eight hours per day.

The institutional review committee of Seoul national university Bundang hospital and the Seoul national university approved the wired and wireless EEG protocols, respectively. All participants submitted written informed consent.

#### Experimental Protocol

A schematic presentation is shown in Figure 5.2. We conducted two protocols (wired EEG and wireless EEG), and each protocol consisted of two condi-

Table 5.1: Sleep time (minutes) for each condition

| No-sleep-deprived | | Sleep-deprived | |
|---|---|---|---|
| Night -6 – -2 | Night -1 | Night -6 – -2 | Night -1 |
| 441.25±38.50 | 431.25±31.82* | 440.52±44.43 | 216.25±30.21* |

*Sleep time of the night before measurement on each condition showed significant difference ($p < 0.05$).

tions (sleep-deprived and no-sleep-deprived). Two conditions were randomly ordered, and the inter-condition interval was two weeks. During the study period, subjects were instructed to have regular sleep for more than seven hours per night; however, in the sleep-deprived condition, they were instructed to sleep less than four hours on the night before the EEG acquisition. Each subject's compliance to the given sleep-wake schedule was monitored through daily sleep logs. Sleep duration was successfully controlled as shown in Table 5.1, 216.3±30.2 minutes on a sleep-restricted day. All the subjects abstained from alcohol and caffeine for at least the 48 hours before and during the day of the EEG measurement.

All EEG measurements were performed from 9:30 AM to 6:30 PM in an isolated space. Subjects were allowed to have regular meals and water, but abstained from drinking caffeine and alcohol, smoking, and napping. The first meal was before 9:30 AM and the lunch was between 12:30 PM and 1:30 PM. The subjects' behaviors were continuously monitored by real-time video.

### EEG and Drowsiness Measurement

A standard wet-electrode EEG (Beehive Horizon, Grass Technologies, Natus, USA) and a cap-type dry-electrode EEG (Ybrain Inc., Republic of Korea) device were adopted for the wired and wireless EEG protocols, respectively. A total of 19 EEG electrodes were placed according to the standard 10-20 system, and an additional single ECG channel (modified type II lead) was recorded. The recording list is as below:

$$\begin{aligned} \text{set } \mathcal{C} = \{&\text{Fp1}, \text{Fp2}, \text{F7}, \text{F3}, \text{Fz}, \text{F4}, \text{F8}, \text{T3}, \text{T4}, \\ &\text{T5}, \text{T6}, \text{C3}, \text{Cz}, \text{C4}, \text{P3}, \text{Pz}, \text{P4}, \text{O1}, \text{O2}, \text{ECG}\}. \end{aligned} \tag{5.1}$$

A referential montage was adopted in which all the electrodes were referenced to the left mastoid electrode (A1).

For the wired EEG protocol, the EEG data were continuously recorded from 9:30 AM to 6:30 PM, and five drowsiness measurements were taken at two-hour interval (Figure 5.2). Each measurement consisted of two recordings with continuous background EEG monitoring: a 60-minute R100 recording and a ten-minute PVT in the middle of the former. For the wireless protocol, subjects wore the EEG cap every two hours for 20 minutes (Figure 5.2). As wireless EEG cap interfered the stable application of R100 and the EEG artifacts increased due to the contacts between the temporal EEG electrodes and the temple bows of the R100 eye glasses, only PVT information was used to phenotype drowsiness. For both protocols, all the recording programs were installed on a single laptop to synchronize the time axis for EEG, R100, and PVT data.

### 5.3.2 Feature Extraction

#### EEG Preprocessing

The $j$th channel EEG signal $\mathbf{x}^j$ is computed by the subtraction from the electric potential $\mathbf{i}^j$ to the $\mathbf{i}^{A1}$ signal from the left mastoid (channel A1) as below:

$$\mathbf{x}^j = \mathbf{i}^j - \mathbf{i}^{A1}. \tag{5.2}$$

Other montage (average referential and longitudinal bipolar montage) did not show significant performance difference with the left mastoid reference montage. A 1 Hz low-pass filter and a 50 Hz high-pass filter is set. The sampling rate is 200 Hz.

To prevent drowsiness-related accidents, the length of the EEG signal required for detecting instantaneous drowsiness should be short. In our approach, we split the EEG signal without overlap into a specific length (1–16 seconds) segment. When the size of the window was set to $w$, the EEG segment could be expressed as follows:

$$\mathbf{x}_i^j = [x^j_{(i-1)*w+1}, x^j_{(i-1)*w+2}, ...x^j_{i*w}]. \tag{5.3}$$

#### Multitaper Spectral Feature Extraction

In this study, multitaper power spectral density (MPSD) [154] based feature extraction is performed on each EEG segment. For the $\mathbf{x}_i^j$ in (5.3), single power spectral density (SPSD) $s_i^j(f)$ [153] at the frequency $f$ is calculated as follows:

$$s_i^j(f) = \Delta t \left| \sum_{k=1}^{w} w_k^{(l)} x^j_{(i-1)*w+k} e^{2\pi k f \Delta t} \right|^2, \tag{5.4}$$

where $w$ is a taper function and $\Delta t$ indicates the sampling duration. MPSD is computed by averaging $L$ number of SPSDs generated with orthogonal tapers

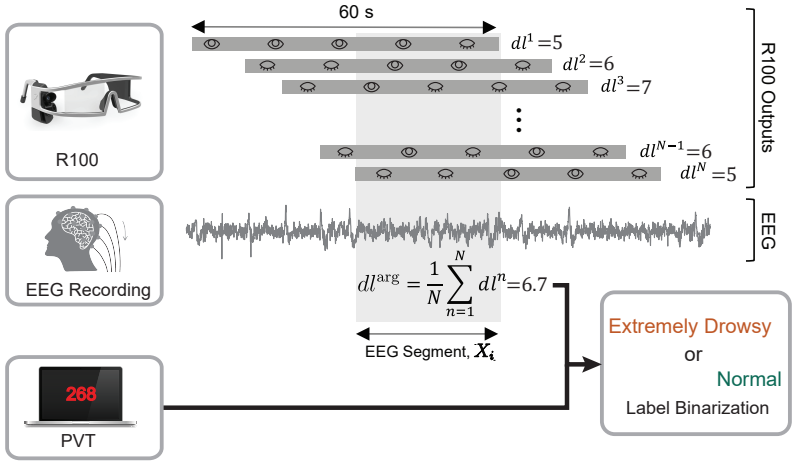



Figure 5.3: Since multiple R100 outputs correspond to given an EEG segment $X_i$, one representative drowsiness level $dl^{\arg}$ is aggregated by averaging.

as follows:

$$\bar{s}_i^j(f) = \frac{1}{L}\sum_{l=1}^{L} s_i^{j(l)}(f). \tag{5.5}$$

The MPSD of the $\mathbf{x}_i^j$ for $m$ frequencies is given as follows:

$$\mathbf{s}_i^j = [\bar{s}_i^j(f_1), \bar{s}_i^j(f_2), ..., \bar{s}_i^j(f_m)]. \tag{5.6}$$

The spectral information of $n$ number of channels for the EEG segment $X_i$ is defined as $\mathbf{S}_i = [\mathbf{s}_i^1, \mathbf{s}_i^2, ..., \mathbf{s}_i^n] \in \mathrm{R}^{m*n}$.

### 5.3.3 Drowsiness Labeling

Since our approach is based on supervised learning, it is necessary to define a label indicating drowsiness for given EEG segments. To focus on detecting instantaneous drowsiness, we define a label of whether the subject showed lapse or not during PVT measurement. However, EEG with PVT information (50 minutes per day) is not enough to train the model. Thus, to define the label $y_i$, we aggregate the R100 information (five hours per day) in section 5.3.3. Then, we find a proper threshold to binary $y_i$ and validate how well $y_i$

is related to the occurrence of the lapse in section 5.3.3.

### R100-based Drowsiness Labeling

The EEG segment (1–14 seconds) is relatively short compared to the time required by the R100 (60 seconds); thus, the multiple outputs $dl^n$ of the R100 correspond to a single EEG segment. Therefore, in our work, to increase robustness, we aggregate the R100 outputs into one representative value, as shown in Figure 5.3. We adopt an averaging method to define the label $dl_i^{\text{arg}}$ corresponding to the $i$th EEG segment as follows:

$$dl_i^{\text{arg}} = \frac{1}{N} \sum_{n=1}^{N} dl_i^n. \tag{5.7}$$

To validate the averaging method, the majority of $dl$s and the first $dl^1$ output are compared empirically in the experiment.

To binarize the $dl_i^{\text{arg}}$, the following formula is defined:

$$y_i = \begin{cases} 1 & \text{if } dl_i^{\text{arg}} \geq \text{threshold} \\ 0 & \text{if } dl_i^{\text{arg}} < \text{threshold.} \end{cases} \tag{5.8}$$

$y_i = 1$ represents the instantaneous state and $y_i = 0$ represents the normal state. The following section discusses how to decide the threshold using PVT information.

### Threshold searching for Label Binarization

A lapse is a temporary episode of drowsiness which lasts for a second where a subject fails to respond in a certain time ($t$) [157]. Let $\text{RT}_i$ be the PVT response time in $X_i$, then occurrence of the lapse $L_i$ can be defined as

$$L_i = \begin{cases} 1 & \text{if } \text{RT}_i \geq t \\ 0 & \text{if } \text{RT}_i < t \end{cases}. \tag{5.9}$$

To determine a precise threshold of (5.8), we use the method proposed by Francois' [143]. The threshold in (5.8) is chosen which to minimize the distance between $y_i$ in (5.8) and $L_i$ in (5.9). By assuming $L_i$ as target and $y_i$ as a predicted target, we calculate the sensitivity and specificity based on their difference. Thus, the receiver operating characteristic (ROC) curve can be plot by modifying the threshold value. The optimal threshold is determined when the average of the sensitivity and specificity is maximized. Once the threshold is determined, we are able to determine $y_i$ over the entire duration when the subject wears R100.

### 5.3.4 Drowsiness Detection

A binary classifier using the XGBoost is trained for each subject with the feature vector $S_i$ and the corresponding label $y_i$ respectively as the input and the output. XGBoost is an ensemble method that aggregates outputs from $K$ number of classification and regression trees (CART) as follows:

$$\hat{y}_i = \sum_{t}^{T} f_t(S_i), \tag{5.10}$$

where $f_k$ represents the $k$th tree model. The objective function is given by

$$Obj = \sum_{i}^{M} l(y_i, \hat{y}_i) + \sum_{k}^{K} \Omega(f_k), \tag{5.11}$$

where $l(\cdot)$ and $\Omega(\cdot)$ are the loss function and the complexity of trees respectively with $M$ training samples. $l(\cdot)$ is the cross-entropy, which is defined as

$$l(y_i, \hat{y}_i) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i). \tag{5.12}$$

In our problem, the ratio of the negative class (normal state) samples against the positive class (drowsiness) samples is approximately $8 : 2$. Since these imbalanced data can cause biased classification [5], by multiplying a

weight, we define a cost-sensitive loss $l(y_i, \hat{y}_i)$ as (5.13)

$$l_c(y_i, \hat{y}_i) = -w_c y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i), \tag{5.13}$$

where $w_c = (\#$ of positive$)/(\#$ of negative$)$.

XGBoost is based on an additive training technique that progressively adds trees to increase the precision of the prediction as

$$\hat{y}_{(t)} = \hat{y}_{(t-1)} + f_t(S), \tag{5.14}$$

where $t = 1, ..., T$, and $\hat{y}_0 = 0$. Since XGBoost is vulnerable to overfitting, the area under curve (AUC) values of the validation sets are used as a criterion for early stopping. The training stops when there is no further improvement after the addition of more than 300 trees. In our work, we used five-fold cross-validation to validate our framework.

By letting $j$ as an index of a single leaf among the $K$ number of leaves in a tree, the object function (5.11) converges as

$$Obj^* = -\frac{1}{2} \sum_{k=1}^{K} \frac{G_k^2}{H_k + \lambda} + \gamma K, \tag{5.15}$$

where $\lambda$ and $\gamma$ are parameters that control the trade-off relationship between the complexity and accuracy. $G$ and $H$ is defined as

$$G_k = \sum_{i \in I_k} \partial_{\hat{y}_i} l(y_i, \hat{y}_i), \qquad H_k = \sum_{i \in I_k} \partial^2_{\hat{y}_i} l(y_i, \hat{y}_i), \tag{5.16}$$

where $I_k$ denotes a set of samples reaching the $k$th leaf in each tree. The $Obj^*$ in (5.15) value measures how well a tree is structured. The optimization of the tree is conducted by expanding the leaves based on information gain ($Gain$)

(5.17) which is defined by

$$\text{Gain} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_L^2 + G_R^2}{H_L + H_R + \lambda} - \gamma. \tag{5.17}$$

Each term respectively denotes the left children score, and the right children score, as well as the score without split. By considering these three cases, a tree is structured to maximize the total gain value. Following a pruning technique, if the gain of each leaf is smaller than $\gamma$, tree addition is terminated.

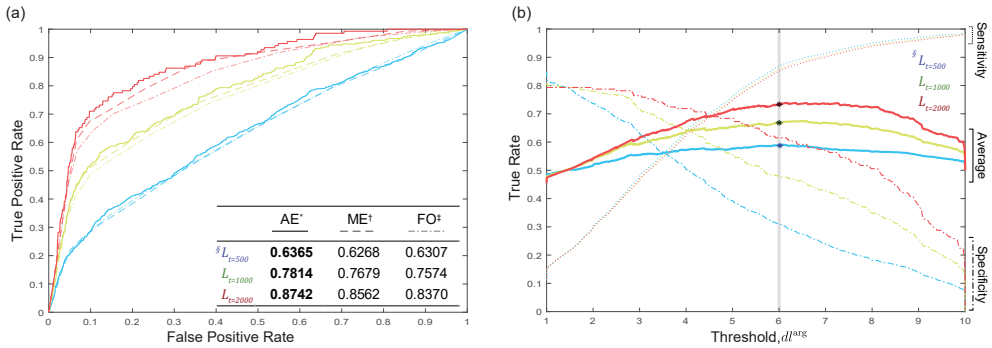
Since each branch of the tree corresponds to one of the features, we can compute the total gain for a specific feature by summing the gain of each branch. The sum represents the importance of the corresponding feature. In our study, since spectral information of each node is used as a feature, we are able to easily compute the importance of each frequency and each channel.

### 5.3.5 Applicability in a Wireless EEG environment

Wireless EEG was additionally acquired to confirm that the framework is applicable in a wireless EEG environment. However, since drowsiness information from the R100 was not acquired in the case of wireless EEG, as below, the extreme drowsiness for wireless protocol, $y_i^{\text{wl}}$, was defined directly with the lapse during the PVT tasks.

$$y_i^{\text{wl}} = \begin{cases} 1 & \text{if } \text{RT}_i \geq t \\ 0 & \text{if } \text{RT}_i < t \end{cases}. \tag{5.18}$$

$t$ is adaptively considered for each subject as $t = \mu + 2 * \sigma$ where $\mu$ and $\sigma$ are respectively the mean and the standard deviation of RT. Except for drowsiness labeling, the rest of the process is the same as with the wired EEG.




*Average Ensemble (Proposed), †Majority Ensemble, ‡First Output, §$L$ in (5.9) for certain $t$

Figure 5.4: Evaluation results on three methods of drowsiness level definition. (a) ROC curves and AUC values for the three methods on three values of $t$. (b) Sensitivity and specificity trends as function of a sweeping AE drowsiness level on three values of $t$.

## 5.4 Results

### 5.4.1 Evaluation of Drowsiness Label

ROC curves of the drowsiness label defined by the three methods (average ensemble, majority ensemble, and first output) in Section 5.3.3 are shown in Figure 5.4 (a). The proposed averaging method shows a high AUC value regardless of $t$. Therefore, we can conclude that the labeling defined by the averaging aggregation method of R100 is most highly related to lapse during PVT.

As discussed in Section 5.3.3, the proper threshold value should be determined. Figure 5.4 (b) shows the sensitivity and specificity values between $L_i$ in (5.9) and $y_i$ in (5.7) for every $t$. The mean values of sensitivity and specificity are depicted as bold lines. The mean values were maximized on $dl_{ens} = 6$. In other words, if $dl_{ens} \geq 6$ (=threshold), we assumed that the subjects were suffering from instantaneous drowsiness.

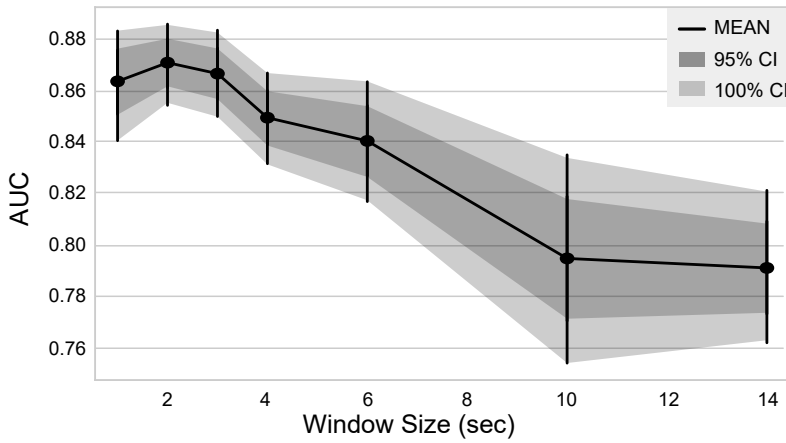



Figure 5.5: The AUC distributions according to window size, $w$ in (5.3)

### 5.4.2 Compatibility as Instantaneous Drowsiness Detection

To instantaneously detect the drowsiness, the length of the EEG signal should be short. We analyzed the AUC values on various window sizes, $w$ in (5.3). The highest accuracy was reached with the two-second window size, which is sufficiently short to detect instantaneously. Longer EEG segments resulted in lower accuracy. We think that the chance of a mixed condition of normal and drowsiness states in the long EEG segment might produce a negative effect.

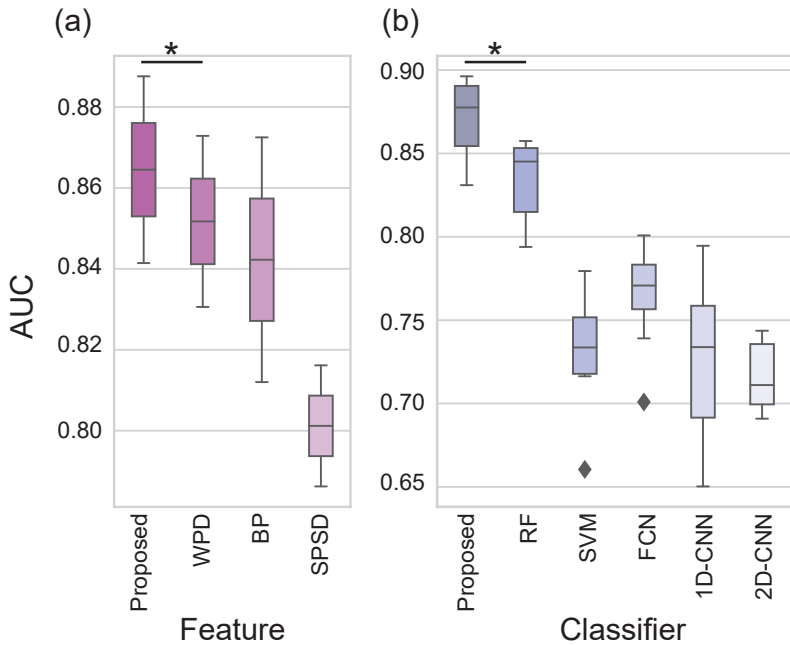



Figure 5.6: AUC distributions on (a) features and (b) classifiers.

### 5.4.3 Comparative Analysis

Comparisons between the techniques used in proposed frameworks and other conventional techniques (three methods of feature extraction and five classifiers) were conducted. Figure 5.6 depicts the distribution of AUC values. By testing Delong's method [92], the statistical significance among the methods was analyzed.

#### Comparison against Other Feature Extraction Methods

Three widely used feature extraction methods were implemented and compared against MPSD. The first method was band power (BP) [151], which uses the power values on specific bands, such as the delta ($0.5-4$ Hz), theta ($4-8$ Hz), alpha ($8-13$ Hz), beta ($13-30$ Hz), and gamma ($30-50$ Hz) bands. The second was Welch's method based on SPSD [153]. The third method was

the wavelet packet decomposition (WPD) [158]. WPD is known to overcome the previous methods' limitation, *i.e.,* the lack of precision with nonstationary signals. For every feature extraction technique, the XGBoost were used. As shown in Figure 5.6 (a), the method that we used in our work, MPSD, shows the best performance, whereas SPSD shows the worst. The multi-taper technique significantly boosts the performance of extracting the spectral information accurately.

**Comparison against Other Classifiers**

Five widely used machine learning classifiers were implemented and compared. The first classifier was random forest (RF) [86], which is a bagging-based ensemble technique. The second classifier was SVM [87]. The remaining types of classifiers were deep-learning (DL) models [159]. DL models could be implemented into various architectures. We implemented three different architectures: fully connected networks (FCNs), 1D convolutional networks (1D-CNNs), and 2D convolutional networks (2D-CNNs). Due to class imbalance, the training of the SVM and the DL models showed difficulties on convergence. For this reason, we balanced the training data by generating synthetic data using the synthetic minority over-sampling technique (SMOTE) [2] for the drowsiness class. All the hyper-parameter sets were selected using a greedy search.

As shown in Figure 5.6 (b), XGBoost which we used in our work, shows the best results. RF-based detection shows relatively good performance against the others. Even though SMOTE was additionally applied as prepossessing to overcome imbalanced data issue, SVM shows relatively low performance with a biased classification on a major class (normal state). Even though we applied various techniques (*e.g.,* such as skip connection, drop-out, and batch normalization), DL models also do not show good performance.

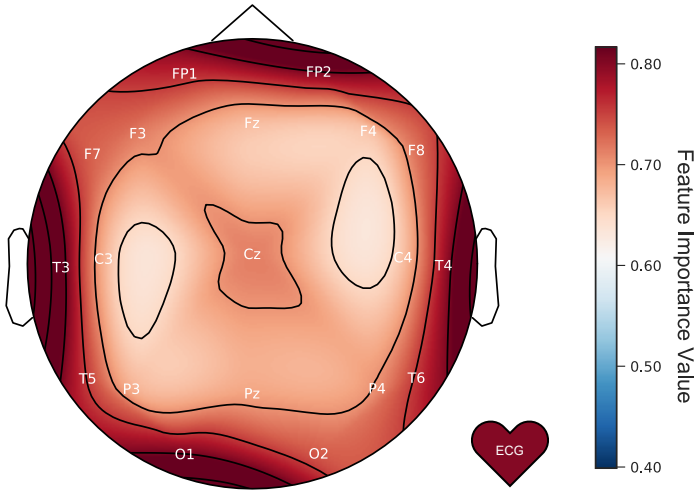



Figure 5.7: Topographic mapping of feature importance value

### 5.4.4 Feature Importance

As described in Section 5.3.4, the importance of the input features was computed using the gain value. The importance values of a specific frequency are shown in Figure 5.8. Peaks are shown in the theta and the alpha bands, which means that these bands play a key role in decisions. This finding shows consensus with the existing studies (e.g., the alpha attenuation test [145] and the Karolinska drowsiness test [144]). Furthermore, we found that the range $(45-50$ Hz) in the gamma band is also important. Figure 5.7 shows a topological mapping of importance. Seven channels (Fp1, Fp2, T3, T4, O1, O2, and ECG) are the key channels among 10-20 system channels.

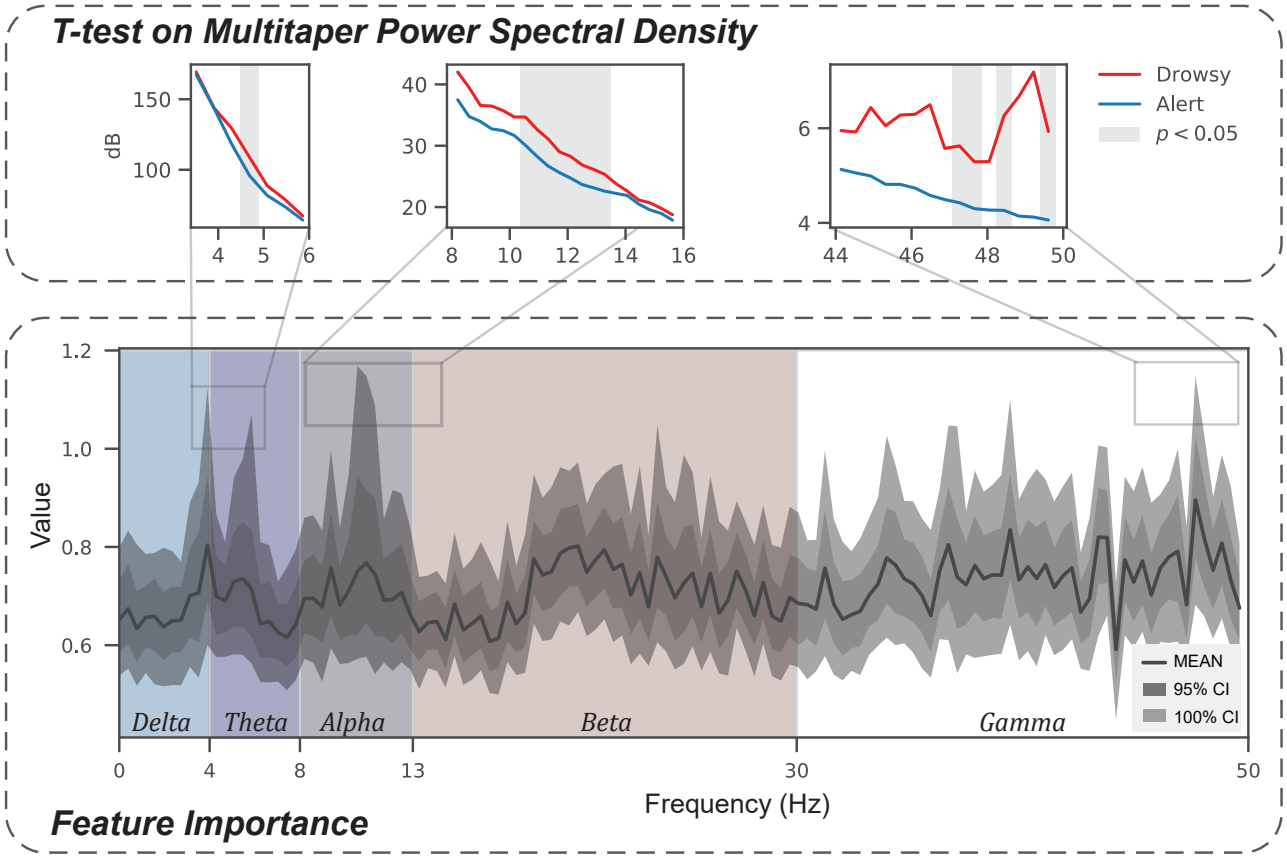



Figure 5.8: Feature importance as a function of frequency. On certain frequencies that show peaks on feature importance, t-test showed significant difference.

### 5.4.5 Channel Reduction

If the framework is based on fewer channels, a more lightweight and cost-effective system can be implemented. Detection performance according to various channel combinations is shown as Figure 5.9. When all channels are used, the best performance is obtained. When using seven selected channels (Fp1, Fp2, T3, T4, O1, O2, and ECG) based on the importance, a small performance drop of -1.80% is observed. As the number of channels decreased, the performance of the system continued to degrade. The performance increased when the ECG channels were added for all the cases. Furthermore, using only the channel of a specific region, the performance was in the order of the ECG-temporal-frontal-occipital region. There was no significant difference in performance between the left hemisphere (Fp1, T3, O1, and ECG) and the right hemisphere (Fp2, T4, O2, and ECG).

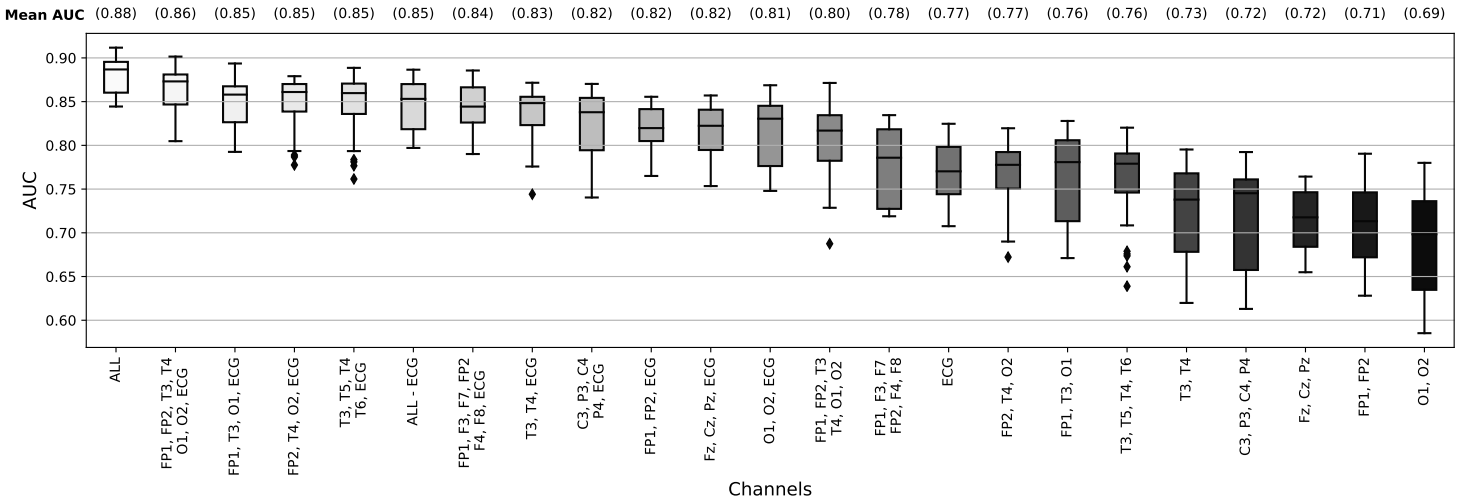



Figure 5.9: Performance degradation according to channel reduction

### 5.4.6 Results on Wired and Wireless EEG

As mentioned in Section 5.3.5, in addition to the wired EEG device, we have acquired additional data with the wireless EEG system, which uses dry sensors. The performance of each subject using the wired EEG and wireless EEG is described in Table 5.2. The performance for each subject represents the mean and standard deviation of the five-fold cross-validation. For the wired EEG, most of the metrics show values above 0.78, which means that the proposed framework can be applied to the actual situation. Additionally, all the metrics have small standard deviations (less than 0.05 among folds and subjects). For the wireless EEG, the AUC values for all subjects showed an average performance degradation of -6.55%. In addition, the standard deviation of performance was higher in the wireless EEG than with the wired EEG. In other words, it can be interpreted that detection is less stable with the wireless EEG than with the wired EEG. This performance degradation could be caused by the instability of the dry sensor or the lack of EEG length used during learning. However, since all indicators show a value of over 0.70, we expect the wireless EEG to be useful for drowsiness detection in a mobile environment.

Table 5.2: Subject-specific performance of wire EEG and wireless EEG

| | Wire EEG + R100 + MPSD + XGBoost | | | | Wireless EEG + PVT + MPSD + XGBoost | | | |
|---|---|---|---|---|---|---|---|---|
| Subject | Accuracy | AUROC | Sensitivity | Specificity | Accuracy | AUROC | Sensitivity | Specificity |
| 1 | 0.8033 ± 0.0056 | 0.8913 ± 0.0038 | 0.8030 ± 0.0053 | 0.8034 ± 0.0057 | 0.8474 ± 0.0374 | 0.9154 ± 0.0259 | 0.8133 ± 0.0183 | 0.8486 ± 0.0383 |
| 2 | 0.7665 ± 0.0139 | 0.8489 ± 0.0042 | 0.7678 ± 0.0167 | 0.7664 ± 0.0137 | 0.6425 ± 0.0709 | 0.6766 ± 0.0582 | 0.6067 ± 0.0683 | 0.6436 ± 0.0711 |
| 3 | 0.7661 ± 0.0091 | 0.8461 ± 0.0066 | 0.7660 ± 0.0087 | 0.7661 ± 0.0092 | 0.8200 ± 0.1459 | 0.8054 ± 0.1941 | 0.6000 ± 0.4183 | 0.8247 ± 0.1487 |
| 4 | 0.7980 ± 0.0138 | 0.8889 ± 0.012 | 0.7984 ± 0.0141 | 0.7980 ± 0.0138 | 0.9132 ± 0.0969 | 0.9555 ± 0.0401 | 0.7333 ± 0.4346 | 0.9167 ± 0.0998 |
| 5 | 0.7428 ± 0.0146 | 0.8211 ± 0.0147 | 0.7417 ± 0.0141 | 0.7429 ± 0.0146 | 0.6536 ± 0.1215 | 0.6760 ± 0.1604 | 0.6000 ± 0.1369 | 0.6548 ± 0.1216 |
| 6 | 0.8269 ± 0.0080 | 0.9056 ± 0.0062 | 0.8267 ± 0.0085 | 0.8269 ± 0.0079 | 0.7730 ± 0.0513 | 0.8238 ± 0.0564 | 0.7179 ± 0.0542 | 0.7754 ± 0.0534 |
| 7 | 0.7786 ± 0.0115 | 0.8609 ± 0.0077 | 0.7788 ± 0.0119 | 0.7785 ± 0.0115 | 0.7706 ± 0.0736 | 0.8158 ± 0.0756 | 0.7000 ± 0.0935 | 0.7727 ± 0.0769 |
| 8 | 0.7986 ± 0.0149 | 0.8813 ± 0.0106 | 0.7985 ± 0.0152 | 0.7986 ± 0.0149 | 0.7574 ± 0.0905 | 0.8170 ± 0.0835 | 0.6933 ± 0.1738 | 0.7594 ± 0.0940 |
| Average | 0.7851 ± 0.0266 | 0.8680 ± 0.0284 | 0.7851 ± 0.0267 | 0.7851 ± 0.0266 | 0.7722 ± 0.0917 | 0.8107 ± 0.0988 | 0.6831 ± 0.0763 | 0.7745 ± 0.0925 |

Table 5.3: Comparison of various imbalanced learning techniques on drowsiness detection

| Model | Accuracy | AUPR | AUROC | F1 | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Classifier Only | **88.72 %** | 0.4084 | 0.7722 | 0.2875 | 0.2160 | **0.9770** |
| Cost-sensitive Loss | 77.81 % | 0.4048 | 0.7742 | 0.4023 | 0.5886 | 0.8075 |
| SMOTE | 63.70 % | 0.3962 | 0.7642 | 0.3427 | 0.7470 | 0.6178 |
| Borderline SMOTE | 67.82 % | 0.3914 | 0.7613 | 0.3536 | 0.6984 | 0.6657 |
| ADASYN | 60.08 % | 0.3819 | 0.7600 | 0.3283 | 0.7644 | 0.5722 |
| Cluster Centroids | 55.70 % | 0.2892 | 0.6785 | 0.2788 | 0.6562 | 0.5580 |
| SMOTE ENN | 55.37 % | 0.3723 | 0.7774 | 0.3156 | **0.8396** | 0.5077 |
| Proposed in Chapter 3 | 87.66 % | **0.4128** | **0.7823** | **0.4235** | 0.3975 | 0.9368 |

### 5.4.7 Imbalanced Data Classifications

Table 5.3 shows the results of applying imbalanced learning techniques to drowsiness detection. In this chapter, Xgboost has been used for a classifier. However, if a large amount of data is accumulated later, deep learning classifier must be an ultimate choice. Therefore, the imbalanced learning analysis has been done based on deep learning classifier. For equivalent comparisons, we have used an equivalent deep learning classifier architecture for all techniques. When applying the classifier only, it is possible to confirm the imbalanced learning tendency is biased since the sensitivity value is much lower than the specificity value. As a result of applying the cost-sensitive loss, the specificity value decreases slightly, but sensitivity value is increased, which means that learning is more balanced than classifier without cost-sensitive loss. As shown in the table, data-level techniques increase the sensitivity but reduce the specificity largely. As a result, the AUPR, AUROC, and F1 value was adversely affected. This implies that the generated data by the conventional data-level techniques are not enough to represent the actual data distribution. However, the methodology proposed in Chapter 3 achieves the highest records in AUPR, AUROC, and F1 which represent the performance in consideration of both classes, whereas the sensitivity is improved without loss of specificity.

## 5.5 Discussion

We have proposed a novel framework for detecting instantaneous drowsiness, which can be applied regardless of the subject's circumstance. In other words, our framework is able to detect drowsiness in any situation. Once the model is trained, it does not require any extra tasks, such as PVT, to evaluate the drowsiness. Moreover, since our framework uses a short EEG segment of two seconds, it is able to quickly detect and notify instantaneous drowsiness states such as a lapse.

The proposed framework includes feature extraction using MPSD and a classifier using XGBoost. The MPSD successfully contains meaningful spectral information within the EEG data. As shown in the experiments, the XGBoost successfully detects drowsiness by using spectral information. Our framework shows the most outstanding performance among various feature extraction techniques and machine learning methods, as shown in Section 5.4.1.

Our framework also has the advantage of being able to obtain the frequency bands and channels that play the most important roles in detecting drowsiness. Through the experiments, we reconfirmed that the alpha and the theta bands are critical to detect drowsiness, which is consistent with the previous works [144, 145]. Furthermore, we have demonstrated that the gamma frequency of $45 - 50Hz$ also contributes to detection of drowsiness, which supports the recent studies on the gamma-based sleepiness detection. Abnormal patterns of gamma waves have been reported to be discovered from patients with neurological disorders (e.g.,/Alzheimer's disease, Parkinson's disease, schizophrenia, and epilepsy) [146].

As a result of our experiment, the Fp1, Fp2, T3, T4, O1, O2, and ECG channels have been shown to play important roles for detecting drowsiness. According to the topological findings, drowsiness detection is possible with the placement of electrodes in the highly relevant regions (frontal, temporal, occipital, and ECG). If only a small number of channels is sufficient for drowsi-

ness detection, we can expect cost and weight reduction of the EEG device. Therefore, we evaluated the performance of various channel combinations. As the number of channels decreases, so does the performance. However, when using the seven channels with the highest importance, performance degradation is less than 2%. In other words, even if only seven channels are used, the performance of drowsiness detection is not dramatically decreased.

To examine the applicability of our framework to drowsiness detection in a mobile device environment, this study acquired additional data through a wireless EEG device using dry sensors. However, the adopted wireless EEG system and R100 were not compatible to be acquired simultaneously due to interference. Since wireless EEG was acquired exclusively during PVT, only approximately 50 minutes of EEG signals per day were used for learning. Detection performance using wireless EEG showed a performance degradation of less than 7% AUC compared to wired EEG, and a value of 0.70 or more in most performance indicators. Two main factors contribute to performance degradation. The first factor is that a relatively small amount of EEG was used for learning, compared to the wired EEG. In the case of the wired EEG, we used five hours per day, whereas only 50 minutes a day of EEG data were available for wireless EEG. Since a small amount of learning data causes performance degradation, it is necessary to accumulate sufficient learning data for wireless brain waves. The second factor is the instability of the dry sensors in the wireless EEG. Although the dry sensors have the advantage of being easy to wear, they have lower signal quality and higher motion-sensitivity than the wet sensors. Nevertheless, performance degradation is not drastic; therefore, if these factors are improved, good and stable detection performance can be expected in the mobile device environment using wireless EEG. In future work, we need to develop and implement an algorithm to handle the EEG artifacts and their effects on the performance to detect drowsiness, especially for dry wireless EEG system.

During the data acquisition, subjects were given various constraints such as long measurements, limited movement, and controlled sleep time. Such constraints make recruiting difficult, which resulted in a small data: eight subjects for the case of our study. Since the amount of data was insufficient, the performances of the deep learning methods such as FCN and CNN were not outstanding. According to recent researches, deep learning has shown a state-of-the-art performance in various fields [1, 55, 56, 57, 58, 59, 60]. However, deep learning can automatically extract meaningful features and achieve great performance only with a sufficient amount of data. As more data is collected later, we hope not only to achieve better performance, but also to uncover new useful features through deep learning. Another limitation of the proposed framework due to the lack of data is that only the intra-subject approach was considered. One model was trained for one subject, which means that if there is a new subject, the new model will need to be trained. For our model to generalize over any subjects, more data should be accumulated.

# Chapter 6

# Application III: ECG-based Authentication Data Learning

## 6.1 Introduction

The rapid growth in physiological sensors, low-power integrated circuits, and wireless communication has enabled body area networks (BANs) [160, 161, 162]. BANs are now utilized for various purposes such as monitoring traffic, crops, infrastructure, and health [163, 164, 165, 166]. The progression of BANs is vital, but security remains a challenge yet to be resolved [167]. Biometrics, which is a key solution for security issues, deals with identity-recognition problems by utilizing the biological peculiarity–e.g., voice, retina, fingerprint, blood vessel, face, and cardiovascular cycle. Among them, cardiovascular signals were first proposed by Biel *et al.* [168] as biometric signals measured by electrocardiogram (ECGs). Since the ECG is intrinsically connected to biological functions, circumvention is significantly more difficult than other biometrics such as retina, fingerprint, hand, or face. Measurability and the measuring cost of the ECG is also relatively more reasonable than other biometrics [169].

Starting with the pioneering work of Biel *et. al.* [168], ECG-based biometrics has been steadily studied. The existing methods suggest various ap-

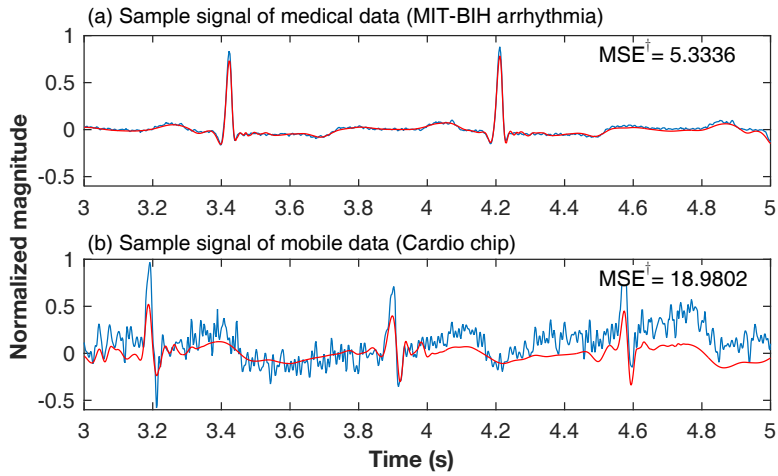



Figure 6.1: Comparison of (a) conventional medical data (MIT-BIH [178]) and (b) data from mobile sensor (Cardio chip). Signals of blue line are raw signal of each data, where as those of red line are denoised signal. †MSE refers to mean squared error between raw signal and denoised signal.

proaches to improve authentication performance.

An ECG-authentication algorithm typically comprises three steps: preprocessing, feature extraction, and classification [169]. For preprocessing, most algorithms have used bandpass filters [170, 171] and novel approaches such as wavelet decomposition [172] have also been developed. The feature extraction algorithms can be categorized into two major types: fiducial-based (time domain) [168, 173, 174] and non-fiducial-based methods (frequency domain) [175, 176]. For the classification, various classifiers have been used, such as k-nearest neighbors (kNN) [171], neural networks [177], and support vector machine (SVM) [170].

Most of the methods verified their approach by using open medical data called MIT-BIH and PTB [178], which are available at Physionet (`http://www.physionet.org/`). The databases offer various types of medical ECGs influenced by heart diseases and conditions, such as sleep stage, apnea, arrhythmia, noise stress, and atrial fibrillation [179, 178]. Various levels of annotations are also available: subject level (demographic information, and di-


Figure 6.2: Pictures of ECG sensing module named CardioChip (composed of two electrodes, Bluetooth network module, and SoC.)

agnostic class), beat level (atrial fibrillation and atrial flutter), and time-point level (peak and end markers for P, QRS, T). Since they are acquired through medical devices, the signals are precise and clean, as shown in Figure 6.1(a). However, the conventional ECG sensors used in medical devices could not be used for practical applications to mobile biometric authentications due to aspects of high cost and measurability. Recently, real-time health-caring devices are attracting considerable interest, and cardiac-disorder diagnosis techniques are being developed on mobile devices with ECG sensors, such as smart watches, earphones, and smart clothing [180]. The use of low-cost ECG signals from mobile sensors is expected to increase the usability of authentication in actual environments.

In this paper, we propose a practical biometric authentication method using ECG signals acquired via mobile sensors. To acquire ECG signals, we use a one-chip-solution module–i.e., the CardioChip made by Neurosky (`http://neurosky.com/`), which is shown in Figure 6.2. The ECG sensing module is composed of two electrodes, the Bluetooth network module, and an SoC named BMD101. The chip BMD101 provides 0.5-100 Hz signal bandwidth, 512 samples per second (sps) sampling rate, 61 dB SNR, and 3 mm $\times$ 3 mm

size.

In addition, we validate the actual utility of the present method for biometric authentication via ECG signals acquired by a low-cost mobile sensor. As shown in Figure 6.1, the ECG signals from this mobile chip are very noisy, with a high MSE of 18.98, unlike the ECG signals from the high-cost medical device, which has an MSE of 5.33.

To remove the various noises embedded in the ECG signals acquired by the mobile sensors, we have designed a band-pass filter by cascading a low-pass filter and a high-pass filter. As the unit of authentication, we present a segmentation from ECG signals into a heartbeat unit, and a feature extraction procedure together with empirically tuned values of parameters. Finally, to achieve a satisfactory performance, a classifier-based authentication scheme is proposed, where training data is constructed to avoid the unbalanced problem in the one-against-all authentication. Nine classifiers are evaluated to choose the best classifier for our authentication purpose.

The rest of the paper is organized as follows: Section 6.2 provides preliminary information about ECGs, mobile devices, classifiers, and evaluation metrics. Section 6.3 details the methodology specified in four steps. Section 6.4 shows the experimental results. Section 6.5 discusses limitations, the difficulty of our approach, and future works.

Table 6.1: Mobile health care devices

| |  |  |  |  |  |
|---|---|---|---|---|---|
| **Category** | Smart watch | Smart watch | Smart watch | Earphone | Sensor |
| **Product** | Galaxy gear S2 | Apple watch | Charge HR | HRM FR74 | iPhone 6/6s Case |
| **Brand** | Samsung elec. | Apple | Fitbit | LG elec. | AliveCor |
| **Release** | Oct. 2015 | Apr. 2015 | July. 2015 | April. 2014 | Jan. 2013 |
| **Price** | $ 300 | $ 400 | $150 | $ 180 | $100 |

## 6.2 Background

### 6.2.1 Electrocardiogram

Electrocardiogram (ECG) is the physical interpretation of the depolarization electrical activity created by the heart muscles [181]. Propagation of the depolarization as a wave is transmitted to the entire body as well as the heart [181]. This wave produces a current that is individually unique depending on the anatomic structure of the individual heart and body, and the current can be easily detected using skin electrodes. ECG is composed of three main electrical entities that are depicted in Figure 6.3: P wave, QRS complex, and T wave. P wave is produced by muscle contraction of the atria. QRS complex marks the ending of atria contraction and the beginning of ventricular contraction. T wave marks the ending of ventricular contraction. The shapes of these entities are determined primarily by human anatomical features so that fiducial information (interval and amplitude) of these entities can be an individually unique feature [182]. In that sense, ECG satisfies the requirements for biometric characteristics [183]: universality (individual possession of the characteristic), measurability (convenience to obtain the characteristic), uniqueness (no identity on two individuals with the characteristics), and permanence (no change in the characteristic over time).

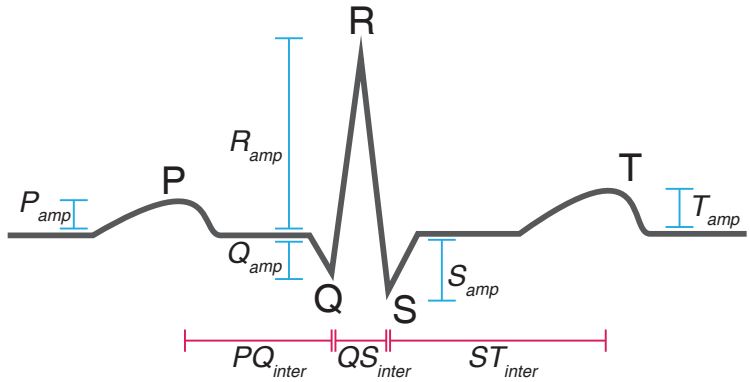



Figure 6.3: General ECG signal shape and extracted fiducial feature per beat; $PQ_{inter}$, $QS_{inter}$, and $ST_{inter}$ refers to intervals among peaks. $P_{amp}$, $Q_{amp}$, $R_{amp}$, $S_{amp}$, and $T_{amp}$ refers to amplitudes of peaks.

### 6.2.2 Mobile Devices for Cardiogram Monitor

As mobile technologies have become a powerful form of media for providing individual-level support to health care, plenty of mobile wearable devices have been released by pioneering major companies [184]. Galaxy Gear Fit (Samsung Electronics), Apple watch (Apple), and Charge HR (Fitbit) are the representative products. Not only watches, but also other types of products such as earphones and clothing, have been developed. Table 6.1 shows the specifications of these representative products, which measure user conditions (exercise quantity, cardiac information, and food-intake patterns) and suggest guidelines for these users. Most key products measure heart rates only, but the devices that provide ECG measurement in addition to the heart rate have also been developed [185]. Even though the current developments for the device with ECG monitoring are only for health diagnosis, this kind of effort will increase the utility of the device with ECG monitoring by easily extending its application to the authentication.

Table 6.2: Classifier list

| Type | Classifier name | Reference |
|---|---|---|
| Classical model | Naïve Bayes | [186] |
| | Logistic regression | [187] |
| | Support Vector Machine | [188] |
| Graphical model | Bayesian Network | [189] |
| | Multilayer Perceptron | [190] |
| | RBF Network | [191] |
| Ensemble model | Bagging | [192] |
| | Random Forest | [86] |
| | Adaboost | [84] |

### 6.2.3 Classification Algorithms

To choose a suitable classifier for our ECG-based authentication, the nine classifiers listed in Table 6.2 were implemented and evaluated. The classifiers can be categorized into three types: classical, graphical, and ensemble model. Naïve Bayes [186] is a simple classifier that uses conditional independence of each variable and Bayesian properties. Logistic regression [187] measures the linear relationship between the dependent variable and independent variables by estimating probabilities using a cumulative logistic distribution. The support vector machine [188] is a robust classifier with good generalization capability that obtains an optimal hyper-plane that maximizes the margin from support vectors, which is the nearest data from the hyperplane. Bayesian networks [189] are probabilistic graphical models that represent a set of random variables and their conditional dependencies via a directed acyclic graph (DAG), which is widely used for hierarchical inference of complex decision systems. Multilayer perceptron (MLP) [191] is a feed-forward artificial neural network with one or more hidden layers, which can work as a universal

approximator. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. The radial basis function (RBF) network [191] is an artificial neural network that uses radial basis functions as activation functions. Bagging [192] trains multiple weak classifiers by bootstrapping, then aggregates the outputs of the weak classifiers to produce a strong classifier. Random forest [86] averages or votes the outputs of multiple decision trees, which is generated by bagging. Adaboost [84] averages the outputs of weak classifiers with weights. Unlike bagging, each weak classifier is trained by the same dataset.

**Algorithm 1** Fiducial feature extractor $\mathscr{F} : \mathbb{R}^l \rightarrow \mathbb{R}^m$

**procedure** DETECTION OF R PEAKS
  $R_{candidate} \leftarrow \arg_i(ECG \geq R_{thr} * \max(ECG))$ ▷ Find all the candidate R peaks
  **if** $|R_{candidate}[i] - R_{candidate}[i+1]| < R_{over}$ **then** ▷ Find all the overlap beats which are too close to each other
    $R_{overlap} \leftarrow \arg\max_i(ECG[R_{candiate}[i]], ECG[R_{candiate}[i+1]])$
  **end if**
  $R_{loc} \leftarrow R_{candidate} \setminus R_{overlap}$ ▷ Remove overlap beats from the candidates
  $R_{amp} \leftarrow ECG(R_{loc})$ ▷ Get the amplitude feature of R peaks
**end procedure**

**procedure** DETECTION OF OTHER PEAKS
  **for** $i$ to length($R_{loc}$) **do** ▷ Get the location feature of the other peaks which correspond to the $i$th R peak
    $P_{loc} \leftarrow \arg\max_i (ECG[R_{loc}[i] - P_{begin}, R_{loc}[i] - P_{end}])$
    $Q_{loc} \leftarrow \arg\min_i (ECG[R_{loc}[i] - Q_{begin}, R_{loc}[i] - Q_{end}])$
    $S_{loc} \leftarrow \arg\min_i (ECG[R_{loc}[i] + S_{begin}, R_{loc}[i] + S_{end}])$
    $T_{loc} \leftarrow \arg\max_i (ECG[R_{loc}[i] + T_{begin}, R_{loc}[i] + T_{end}])$
  **end for**
**end procedure**

**procedure** COMPUTATION OF FIDUCIAL FEATURES
  **for** $i$ to length($R_{loc}$) **do** ▷ Get the interval and amplitude features which correspond to the $i$th R peak
    $PQ_{inter}[i] \leftarrow Q_{loc}[i] - P_{loc}[i]$
    $QS_{inter}[i] \leftarrow S_{loc}[i] - Q_{loc}[i]$
    $ST_{inter}[i] \leftarrow T_{loc}[i] - S_{loc}[i]$
    $P_{amp}[i] \leftarrow ECG(P_{loc}[i])$
    $Q_{amp}[i] \leftarrow ECG(Q_{loc}[i])$
    $R_{amp}[i] \leftarrow ECG(R_{loc}[i])$
    $S_{amp}[i] \leftarrow ECG(S_{loc}[i])$
    $T_{amp}[i] \leftarrow ECG(T_{loc}[i])$
  **end for**
**end procedure**

Parameters:

- $\mathrm{R}_{thr} = 0.4$ (threshold minimum amplitude of R peak, $\mathrm{R}_{over} = 0.1$ s (minimum interval among R peaks)
- $\mathrm{P}_{begin}$ =$\mathrm{RR}_{int}/6$, $\mathrm{P}_{end}$ =$\mathrm{RR}_{int}/10$
- $\mathrm{Q}_{begin}$ =$\mathrm{RR}_{int}/10$, $\mathrm{Q}_{end}$=0
- $\mathrm{S}_{begin} = 0$, $\mathrm{S}_{end}$ =$\mathrm{RR}_{int}/10$
- $\mathrm{T}_{begin}$ =$\mathrm{RR}_{int}/10$, $\mathrm{T}_{end}$ =$\mathrm{RR}_{int}/2$

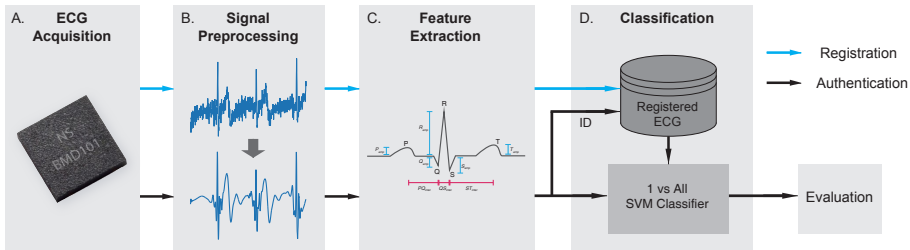



Figure 6.4: An overall scheme of proposed framework: (a) acquiring ECG signals via mobile sensor, (b) cancelling noise for robustness, (c) extracting fiducial features in each heartbeat, (d) authenticating an identity through a one-against-all classification. Blue line refers to registration (train) flow and black line refers to authentication (test) flow.

## 6.3 Method

The overall scheme of the proposed authentication algorithm is depicted in Figure 6.4: (a) acquiring ECG signals via mobile sensor, (b) canceling noise for robustness, (c) extracting fiducial features in each heartbeat, and (d) authenticating an identity through a one-against-all classification. More details of each step are provided in the following.

### 6.3.1 ECG Acquisition

For ECG signal acquisition, we used Neurosky's CardioChip, which provides one lead signal by contacting two electrodes with a pair of thumb fingers. Compared to multi-lead medical devices, the device has a higher level of noise. However, the sensor has strong merit in that it can be easily installed to wearable or mobile devices owing to the strength of its small size and its Bluetooth communication module.

In this study, ECG was recorded in a sitting position at a resting state. Each signal is 60 seconds long, with 512 Hz sampling frequency. A total of 175 subjects (144 male and 31 female) participated in our experiments. We acquired the ECG signal once for 78 subjects, twice for 67 subjects, and three times for the remaining 30 subjects.

### 6.3.2 Noise Cancellation

Noise cancellation should be accompanied, since noise disturbs the robustness of authentication. ECGs commonly contain 60 Hz interference noise, muscle noise ($> 40$ Hz), and electrode contact noise ($< 0.5$ Hz)[193]. Since our sensing device has no glue-on electrode, it has much more electrode contact noise than medical devices. Fortunately, the electrode contact noise has a divisibly lower frequency than those of real ECG signals. In addition, the interference noise and muscle noise have a divisibly high frequency. The 5-15 Hz band is appropriate to extract the fiducial feature for authentication [194]. To focus on this band, we adopted a cascading filter that consists of low- and high-pass filters in turn. The transfer functions of the two filters designed are as follows:

$$H_{\text{low}}(z) = \frac{(1 - z^{-6})^2}{(1 - z^{-1})^2} \tag{6.1}$$

and

$$H_{\text{high}}(z) = \frac{-1 + 32z^{-16} + z^{-32}}{(1 + z^{-1})}. \tag{6.2}$$

### 6.3.3 Fiducial Feature Extraction

A single ECG has a periodic cycle of three electrical entities (see Figure 6.3): a P wave, a QRS complex, and a T wave, generated by atrial depolarization, ventricular repolarization, and ventricular depolarization, respectively [193]. Our algorithm computes an authentication score for each heartbeat, and a procedure is required to segment the whole signal into single heartbeat signals. The segmentation is done by detecting an R-peak in the QRS complex. Based on the detected R-peak location, eight fiducial features ($\text{P}_{amp}$, $\text{Q}_{amp}$, $\text{R}_{amp}$, $\text{S}_{amp}$, $\text{T}_{amp}$, $\text{PQ}_{inter}$, $\text{QS}_{inter}$, and $\text{ST}_{inter}$) are extracted for each heartbeat as shown in Figure 6.3. The detailed procedure to extract the features is described in Algorithm 1. The parameters defining the peak are established by referring

to the normal ECG range suggested in [195]. The values of parameters are described in the bottom part of Algorithm 1.

### 6.3.4 Classification-Based Authentication

In the authentication stage, the feature vector extracted from each heartbeat is an input to the classifier corresponding to a target user. The classifier produces a score indicating the degree of how much the feature matches the target user. The final authentication is done by the ensemble (arithmetic mean) of the scores from the multiple heartbeats in the window. The larger window size improves the performance but requires more time for authentication. That is, there is a trade-off between performance and authentication time. This trade-off will be examined in Section 6.4.2.

For the classifier in our work, we have chosen the SVM that has shown the best performance among the nine classifiers presented in Section 6.2.3, through empirical evaluations. For the evaluations, each classification algorithm was implemented by calling the Java object of libSVM [89] for SVM, and Weka [90] for the other algorithms in MATLAB2015a. To evaluate the generalization ability of each classifier, a 10-fold cross-validation has been applied.

For the validation, an early 30-second signal from each ECG is used to train a classifier, and the next 25-second signal is used to test. A set of multiple test signals is established by moving a window every second for an authentication signal.

Let $h_j^{(i)} \in R^l$ and $x_j^{(i)} \in R^m$ be the $l$-length heartbeat ECG vector and the $m$-dimensional feature vector, respectively, of the $j$-th heartbeat of the $i$-th object, where $i = 1, \cdots, M$ and $j = 1, \cdots, N$. Then

$$x_j^{(i)} = \mathscr{F}(h_j^{(i)}) \tag{6.3}$$

with the feature extracting function $\mathscr{F} : R^l \rightarrow R^m$. To train the one-against-all classifier for the $i$-th object authentication, the positive sample set $\chi_p^{(i)}$ for

$i$-th object is constructed by

$$\chi_p^{(i)} = \{x_j^{(i)} | j = 1, \cdots, N\}, \tag{6.4}$$

whereas its negative sample set $H_n^{(i)}$can be constructed by all samples from the other objects except the $i$-th objects, that is,

$$\chi_n^{(i)} = \{x_j^{(k)} | j = 1, \cdots, N \text{ and } k = 1, \cdots, M \text{ and } k \neq i\}. \tag{6.5}$$

In this case, the number of negative samples is much larger than that of positive samples, which is known to result in degradation of performance. To balance the two sample sets, we use one representative feature vector per each negative object, by averaging all feature vectors of the object. That is,

$$\bar{\chi}_n^{(i)} = \{\bar{x}^{(k)} | \text{ and } k = 1, \cdots, M \text{ and } k \neq i\} \tag{6.6}$$

where

$$\bar{x}^{(k)} = \frac{1}{N} \sum_{j=1}^{N} x_j^{(k)}. \tag{6.7}$$

This approach reduces negative samples with a small amount of computation, which is suitable for a computing environment of the mobile device targeted in our work. However, this approach can lose a large amount of information due to under-sampling. Because of the rapid development of computing power on mobile devices, we expect that algorithms with higher computational complexity can be applied. Thus, there is also a need to examine various imbalance learning techniques in addition to the representative negative method.

Table 6.3: Classifier comparison results

| | SVM | Simple Logistic | Naive Bayes | Random Forest | Adaboost |
|---|---|---|---|---|---|
| **ACC**(%) | **95.990** | 93.160 | 90.330 | 95.680 | 94.300 |
| **EER**(%) | **4.460** | 8.810 | 9.510 | 6.230 | 6.510 |
| **AUC** | **0.996** | 0.951 | 0.981 | 0.996 | 0.995 |
| **SEN** | **0.915** | 0.823 | 0.889 | 0.874 | 0.874 |
| **SPE** | **0.972** | 0.964 | 0.908 | 0.981 | 0.981 |
| **FPR** | **0.028** | 0.036 | 0.092 | 0.019 | 0.019 |
| **FNR** | **0.085** | 0.177 | 0.111 | 0.126 | 0.126 |

| | Bagging | Multi-layer Perceptron | Bayes Net | RBF Network | Lugovaya *et al.* [182] |
|---|---|---|---|---|---|
| **ACC**(%) | 94.400 | 93.790 | 94.940 | 93.900 | 96.00 |
| **EER**(%) | 6.720 | 8.540 | 6.100 | 9.300 | - |
| **AUC** | 0.994 | 0.98 | 0.993 | 0.987 | - |
| **SEN** | 0.858 | 0.83 | 0.884 | 0.844 | - |
| **SPE** | 0.969 | 0.97 | 0.969 | 0.967 | - |
| **FPR** | 0.031 | 0.03 | 0.031 | 0.033 | - |
| **FNR** | 0.142 | 0.17 | 0.116 | 0.156 | - |

## 6.4 Results

Our approach was implemented in MATAB2015a under an Ubuntu 12.04 Linux environment. The experiments were carried out on a machine equipped with four AMD 6172 CPUs, 256GB RAM, Radeon 7970 (2048 cores, 3GB) GPU, and a 2TB hard disk.

### 6.4.1 Single-Beat Authentication Performance

To verify the authentication performance of our approach for single beats, testing and training sets were randomly constructed by a 10-fold cross-validation strategy. Table 6.3 shows the performance of each classifier in terms of six evaluation indices, which have been described in Section 2.3. Each classifier's performance was averaged on all 175 subjects. In each type of classifier, SVM [188], Bayesian network [189], and random forest [86] each showed satisfactory performance with an EER of 4.46%, 6.1%, and 6.23%, respectively. The Radial Basis Function (RBF) kernel-based SVM [188] was the best classifier among all the classifiers used. By parameter tweak, the sigma value of the RBF ker-

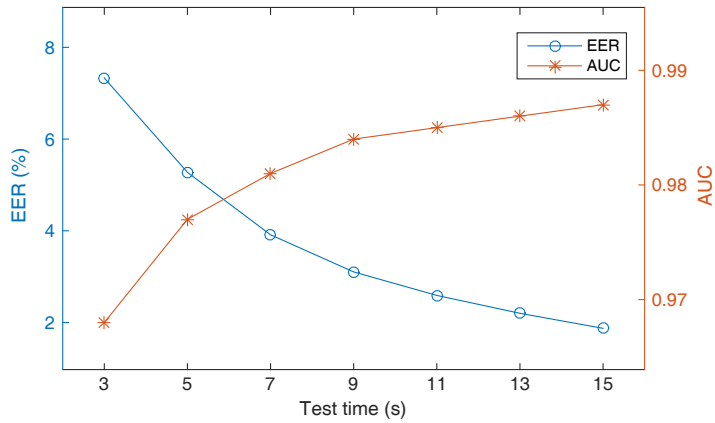



Figure 6.5: Test time validation. Performance trends on EER (blue) and AUC (red) as function of extended test time.

nel in SVM was set to 1.5, which showed the best performance. As shown in Table 6.3, the proposed scheme (SVM), even using a noisy signal, achieves a satisfactory performance comparable to the results of Lugovaya's work, which used an ECG signal from the medical device [182].

### 6.4.2 Actual Authentication Scenario

Since the RBF kernel-based SVM showed the best performance on a single heartbeat, an actual authentication scenario was verified, with an increasing test time from 3 to 15 seconds. As shown in Figure 6.5, the more test time was used, the more accurate a performance was achieved. Compared with the scenario using three seconds of test time, the 15-second scenario shows about 3.91 times higher accuracy in terms of EER. We could achieve a stable authentication performance owing to the ensemble scores of multiple heartbeats.

Table 6.4: Comparison of various imbalanced learning techniques on ECG authentication

| Model | Accuracy | AUPR | AUROC | EER | F1 | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|
| Classifier Only | 96.39 % | 0.8011 | 0.9860 | 4.79 % | 0.2865 | 0.2222 | **0.9997** |
| Cost Sensitive Loss | 93.68 % | 0.7375 | 0.9855 | 4.04 % | 0.6301 | 0.9976 | 0.9338 |
| Representative Negative | 86.36 % | 0.3934 | 0.9254 | 13.05 % | 0.4053 | 0.8788 | 0.8628 |
| Cluster Centroids | 83.52 % | 0.4958 | 0.9492 | 10.41 % | 0.3854 | 0.9647 | 0.8288 |
| SMOTE | 96.45 % | 0.8509 | 0.9927 | 2.44 % | 0.7541 | 0.9990 | 0.9628 |
| Borderline SMOTE | 96.25 % | 0.8432 | 0.9924 | 2.72 % | 0.7440 | 0.9990 | 0.9606 |
| ADASYN | 96.38 % | 0.8485 | 0.9927 | 2.38 % | 0.7511 | **0.9992** | 0.9621 |
| SMOTE ENN | 96.25 % | 0.8432 | 0.9924 | 2.72 % | 0.7440 | 0.9990 | 0.9606 |
| Proposed in Chapter 3 | **99.88 %** | **0.9981** | **0.9999** | **0.04 %** | **0.9871** | 0.9941 | 0.9990 |

### 6.4.3 Imbalanced Data Classifications

Table 6.4 shows the results of applying imbalanced learning techniques to authentication. In this work, SVM has been used for a classifier. However, due to the rapid development of mobile devices, deep learning techniques are expected to be available. Therefore, the analysis has been done based on deep learning classifier. For equivalent comparisons, we have used an equivalent deep learning classifier architecture for all techniques. As a result of applying the classifier only, it is possible to confirm the unbalanced learning tendency with the sensitivity value being extremely low. Based on the F1 score, we can confirm that all the imbalanced learning techniques have improved performance. However, the representative negative and cluster centroid methods, belonging to under-sampling, show a large loss in AUPR, AUROC, and specificity. The cost-sensitive loss shows the same tendency. Unlike Chapter 4 and 5, due to low-dimension of input feature, the over-sampling and hybrid techniques achieve successful performance improvement on every metric. However, the methodology proposed in Chapter 2 still shows the highest records in AURP, AUROC, and F1 which represent the performance in consideration of both classes, whereas the sensitivity is largely improved without significant loss of specificity.

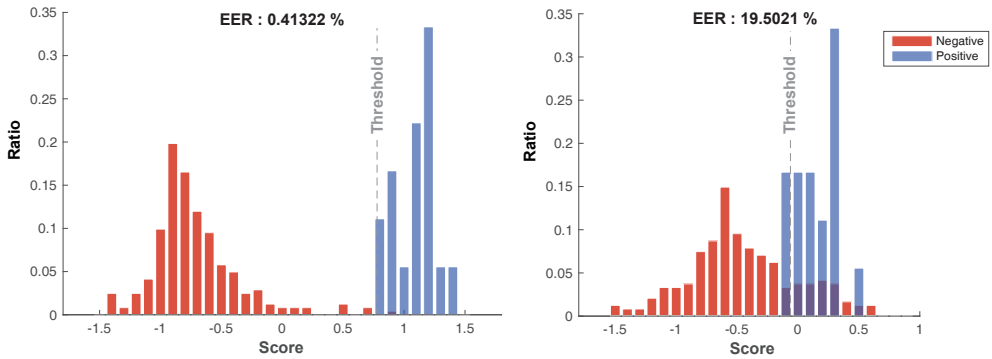



Figure 6.6: Investigation on authentication results: score histogram examples of ECG signals that give the low EER (a) or a high EER (b). The red (blue) bar refers to a score histogram of the negative (positive) class. The gray dot line depicts the decision threshold for authentication.

## 6.5 Discussion

Figure 6.6 shows score histogram examples of ECG signals that give a high accuracy or a low accuracy depending on a subject. Figure 6.6(a) depicts the score histogram of a high authentication performance signal (0.42% EER). The distribution of both positive and negative objects had complete partition on threshold 0.7. In contract, as shown in Figure 6.6(b), the score histogram of a relatively low-performing signal (19.50% EER) had an overlapping region that yields uncertainty on the classification.

We identified two main reasons by analyzing the low-performing signals. One is the extreme noise that is caused by movement in measuring the ECG signal. The other is the failure of R peak detection in extracting the fiducial feature. For more precise noise cancellation, empirical mode decomposition (EMD) [196], wavelet decomposition [197],or sparse derivative [198] could be used. Non-fiducial features that do not need to detect certain characteristic points could be expected to increase robustness [199]. Although our approach is based on shallow learning, deep learning approaches such as convolutional neural networks (CNN) [200] or recurrent neural networks (RNN) [201] could

be used to further improve performance.

Our signals were acquired in a resting state; however, ECGs have characteristics that are influenced by the degree of stress or excitation. When we consider multiple sessions, the performance is usually degraded [199]. Our approach should thus be verified by multi-session ECG signals with various conditions.

One of the difficulties we faced during our study lay in data acquisition. The difficulty was to set some thresholds or hyper-parameters such as the bandpass width at noise cancellation, fiducial peak range at feature extraction, or kernel types and the sigma value fr SVM.

# Chapter 7

# Conclusions

While machine learning methodologies, including deep learning, are updating their state-of-art in a variety of areas, there are a number of issues to be solved in order to successfully apply to real-world application. In this dissertation, one of the issues, the imbalanced data learning, has been addressed along with the proposed GAN-based solution. In addition, throughout application researches using in-house real data, it has been discussed how to solve the imbalanced data distribution. The contributions of this dissertation are summarized as follows.

- A GAN-based imbalanced data learning method has been proposed, in which the classifier, the discriminator, and the generator are jointly trained. To this end, we have designed a three-player (the generator, the discriminator, and the classifier) utility function and provide proof of the existence of an equilibrium point of the utility function. In the training scheme to search an equilibrium point of the utility function, the main contribution is the cooperative training loop of the generator and the classifier, where the generator and the classifier are trained alternately to gradually expand the decision region of the minority class. An additional contribution is the use of batch-wise balancing, which balances the majority and minority samples for every mini-batch in training

the classifier. The validity of the proposed method was verified experimentally via ablation study, visualization analysis, and comparison with existing methods.

- As an application of imbalanced data learning, a low-cost high-accuracy diagnose framework has been proposed for early diagnosis of dementia. To achieve low-cost high-accuracy diagnose performance for dementia using a neuropsychological battery, a novel framework was proposed using the response profiles of 2,666 cognitively normal elderly individuals and 435 dementia patients who have participated in the KLOSCAD. The key idea of the proposed framework is to propose a cost-effective and precise two-stage classification procedure that employed Mini Mental Status Examination as a screening test and the KLOSCAD Neuropsychological Assessment Battery as a diagnostic test using deep learning. In addition, an evaluation procedure of redundant variables has been introduced to prevent performance degradation. A missing data imputation method is also presented to increase the robustness by recovering information loss. The proposed DNNs architecture for the classification has been validated through rigorous evaluation in comparison with various classifiers. The k-nearest-neighbor imputation has been induced according to the proposed framework, and the proposed DNNs for two stage classification show the best accuracy compared to the other classifiers. Also, 49 redundant variables were removed, which improved diagnostic performance and suggested the potential of simplifying the assessment. Using this two-stage framework, we could get 8.06% higher diagnostic accuracy of dementia than MMSE alone and 64.13% less cost than KLOSCAD-N alone. The proposed framework could be applied to general dementia early detection programs to improve robustness, preciseness, and cost-effectiveness.

- A drowsiness detection framework has been proposed using brain-wave

signals. The socioeconomic losses caused by extreme daytime drowsiness are enormous in these days. Hence, building a virtuous cycle system is necessary to improve work efficiency and safety by monitoring instantaneous drowsiness that can be used in any environment. In this dissertation, a novel framework is proposed to detect extreme drowsiness using a short time segment ($\sim$ 2 s) of EEG which well represents immediate activity changes depending on a person's arousal, drowsiness, and sleep state. To develop the framework, multitaper power spectral density (MPSD) has been used for feature extraction along with extreme gradient boosting (XGBoost) as a machine learning classifier. In addition, a novel drowsiness labeling method has been suggested by combining the advantages of the psychomotor vigilance task and the electrooculography technique. By experimental evaluation, it has been verified that the adopted MPSD and XGB techniques outperform other techniques used in previous studies. Finally, we have identified that spectral components ($\theta$, $\alpha$, and $\gamma$) and channels (Fp1, Fp2, T3, T4, O1, and O2) play an important role in our drowsiness detection framework, which could be extended to mobile devices. Furthermore, we verify the applicability of the proposed framework for a mobile environment by using a wireless EEG with dry-sensors as well as a wired EEG with wet-sensors.

- A noisy ECG-based biometric authentication has been proposed. ECG signals from mobile sensors are expected to increase the availability of authentication in the emerging wearable device industry. However, mobile sensors provide a relatively lower-quality signal than conventional medical devices. This work provides a practical authentication procedure for ECG signals that collected via one-chip-solution mobile sensors. We designed a cascading bandpass filter for noise cancellation and suggest eight fiducial features. For classification-based authentication, the radial basis function kernel-based support vector machine has been used to

show the best performance among nine classifiers through experimental comparisons. In spite of noisy ECG signals in mobile sensors, we achieved 4.61% of EER on a single heart-beat, and 1.87% of EER on 15 seconds' testing time on 175 subjects, which is a reasonable result and supports the usability of low-cost ECGs for biometric authentication.

# Bibliography

[1] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[3] R. B. Rao, S. Krishnan, and R. S. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp. 3–10, 2006.

[4] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning*, ECML'04, (Berlin, Heidelberg), pp. 39–50, Springer-Verlag, 2004.

[5] E. A. Garcia and H. He, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, 12 2008.

[6] S. Clearwater and E. Stern, "A rule-learning program in high energy physics event classification," *Computer Physics Communications*, vol. 67, no. 2, pp. 159 – 182, 1991.

[7] S. J. Graves, G. P. Asner, R. E. Martin, C. B. Anderson, M. S. Colgan, L. Kalantari, and S. A. Bohlman, "Tree species abundance predictions in

a tropical agricultural landscape with a supervised classification model and imbalanced data," *Remote Sensing*, vol. 8, no. 2, 2016.

[8] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211 – 229, 2012.

[9] X.-M. Zhao, X. Li, L. Chen, and K. Aihara, "Protein classification with imbalanced data," *Proteins*, vol. 70, p. 1125—1132, March 2008.

[10] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems and Their Applications*, vol. 14, no. 6, pp. 67–74, 1999.

[11] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *Acm sigkdd explorations newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[12] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, (New York, NY, USA), pp. 155–164, ACM, 1999.

[13] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, pp. 659–665, May 2002.

[14] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97 – 110, 2013.

[15] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing* (D.-S. Huang, X.-P. Zhang, and G.-B. Huang, eds.), (Berlin, Heidelberg), pp. 878–887, Springer Berlin Heidelberg, 2005.

[16] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, June 2008.

[17] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5718 – 5727, 2009.

[18] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications*, vol. 91, pp. 464 – 471, 2018.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014.

[20] C. Wang, Z. Yu, H. Zheng, N. Wang, and B. Zheng, "Cgan-plankton: Towards large-scale imbalanced class generation and fine-grained classification," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 855–859, Sep. 2017.

[21] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and A. C. I. Malossi, "BAGAN: data augmentation with balancing GAN," *CoRR*, vol. abs/1803.09655, 2018.

[22] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Advances in Knowledge Discovery and Data Mining* (D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, eds.), (Cham), pp. 349–360, Springer International Publishing, 2018.

[23] Y. Zhang, "Deep generative model for multi-class imbalanced learning," *Open Access Master's Theses*, vol. 1277, 2018.

[24] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning: Bagging of extrapolation-smote svm," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–11, 01 2017.

[25] B. Rocca, "Handling imbalanced datasets in machine learning," 10 2019.

[26] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, pp. 20–29, June 2004.

[27] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] S. Rota Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[30] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Con-*

*ference on Machine Learning - Volume 70*, ICML'17, pp. 2642–2651, JMLR.org, 2017.

[31] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. PMID: 7063747.

[32] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), pp. 233–240, ACM, 2006.

[33] S. Barua, M. M. Islam, X. Yao, and K. Murase, "Mwmote–majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 405–425, Feb 2014.

[34] C. LI, T. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4088–4098, Curran Associates, Inc., 2017.

[35] U. Hwang, D. Jung, and S. Yoon, "HexaGAN: Generative adversarial nets for real world classification," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, (Long Beach, California, USA), pp. 2921–2930, PMLR, 09–15 Jun 2019.

[36] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge distillation with adversarial samples supporting decision boundary," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[37] K. Sun, Z. Zhu, and Z. Lin, "Enhancing the robustness of deep neural networks by boundary conditional GAN," *CoRR*, vol. abs/1902.11029, 2019.

[38] K. Lee, H. Lee, K. Lee, and J. Shin, "Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples," *arXiv e-prints*, p. arXiv:1711.09325, Nov 2017.

[39] R. G, M. Wainwright, and B. Yu, "Early stopping for non-parametric regression: An optimal data-dependent stopping rule," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, Allerton'11, (Berlin, Heidelberg), pp. 1318–1325, Springer-Verlag, 2011.

[40] A. Krizhevsky, "Learning multiple layers of features from tiny images," *University of Toronto*, 05 2012.

[41] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 3573–3587, Aug 2018.

[42] H.-S. Choi, J. Y. Choe, H. Kim, J. W. Han, Y. K. Chi, K. Kim, J. Hong, T. Kim, T. H. Kim, S. Yoon, and K. W. Kim, "Deep learning based low-cost high-accuracy diagnostic framework for dementia using comprehensive neuropsychological assessment profiles," *BMC Geriatrics*, vol. 18, p. 234, Oct 2018.

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, Dec 2015.

[44] L. van der Maaten, "Learning a parametric embedding by preserving local structure," in *Proceedings of the Twelth International Conference on*

*Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5 of *Proceedings of Machine Learning Research*, (Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA), pp. 384–391, PMLR, 16–18 Apr 2009.

[45] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, pp. 853–867. Boston, MA: Springer US, 2005.

[46] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[47] H. C. Rossetti, C. Munro Cullum, L. S. Hynan, and L. H. Lacritz, "The cerad neuropsychologic battery total score and the progression of alzheimer disease," *Alzheimer Dis Assoc Disord*, vol. 24, no. 2, pp. 138–42, 2010.

[48] E. H. Seo, D. Y. Lee, J. H. Lee, I. H. Choo, J. W. Kim, S. G. Kim, S. Y. Park, J. H. Shin, Y. J. Do, J. C. Yoon, J. H. Jhoo, K. W. Kim, and J. I. Woo, "Total scores of the cerad neuropsychological assessment battery: validation for mild cognitive impairment and dementia patients with diverse etiologies," *Am J Geriatr Psychiatry*, vol. 18, no. 9, pp. 801–9, 2010.

[49] W. R. Shankle, A. K. Romney, J. Hara, D. Fortier, M. B. Dick, J. M. Chen, T. Chan, and X. Sun, "Methods to improve the detection of mild cognitive impairment," *Proc Natl Acad Sci U S A*, vol. 102, no. 13, pp. 4919–24, 2005.

[50] M. E. Strauss and T. Fritsch, "Factor structure of the cerad neuropsychological battery," *J Int Neuropsychol Soc*, vol. 10, no. 4, pp. 559–65, 2004.

[51] J. S. Chang, Y. K. Chi, J. W. Han, T. H. Kim, J. C. Youn, S. B. Lee, J. H. Park, J. J. Lee, K. Ha, and K. W. Kim, "Altered categorization of semantic knowledge in korean patients with alzheimer's disease," *J Alzheimers Dis*, vol. 36, no. 1, pp. 41–8, 2013.

[52] Y. K. Chi, J. W. Han, H. Jeong, J. Y. Park, T. H. Kim, J. J. Lee, S. B. Lee, J. H. Park, J. C. Yoon, J. L. Kim, S. H. Ryu, J. H. Jhoo, D. Y. Lee, and K. W. Kim, "Development of a screening algorithm for alzheimer's disease using categorical verbal fluency," *PLoS One*, vol. 9, no. 1, p. e84111, 2014.

[53] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, "Clinical data mining: a review," *IMIA Yearbook 2009: Closing the Loops in Biomedical Informatics*, no. 1, pp. 121–133, 2009.

[54] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 3rd ed., 2011.

[55] J. Baek, B. Lee, S. Kwon, and S. Yoon, "Lncrnanet: long non-coding rna identification using deep learning," *Bioinformatics*, p. bty418, 2018.

[56] T. Moon, S. Min, B. Lee, and S. Yoon, "Neural universal discrete denoiser," in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 4772–4780, Curran Associates, Inc., 2016.

[57] S. Kwon and S. Yoon, "Deepcci: End-to-end deep learning for chemical-chemical interaction prediction," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, (New York, NY, USA), pp. 203–212, ACM, 2017.

[58] S. Park, S. Min, H.-S. Choi, and S. Yoon, "Deep recurrent neural network-based identification of precursor micrornas," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 2891–2900, Curran Associates, Inc., 2017.

[59] B. Lee, J. Baek, S. Park, and S. Yoon, "deeptarget: End-to-end learning framework for microrna target prediction using deep recurrent neural networks," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '16, (New York, NY, USA), pp. 434–442, ACM, 2016.

[60] H. K. Kim, S. Min, M. Song, S. Jung, J. W. Choi, Y. Kim, S. Lee, S. Yoon, and H. H. Kim, "Deep learning improves prediction of crispr–cpf1 guide rna activity," *Nature biotechnology*, vol. 36, no. 3, p. 239, 2018.

[61] S. Mani, W. R. Shankle, M. J. Pazzani, P. Smyth, and M. B. Dick, "Differential diagnosis of dementia: A knowledge discovery and data mining (kdd) approach," in *Proceedings of the AMIA Annual Fall Symposium*, p. 875, 1997.

[62] R. E. Leighty, *Statistical and Data Mining Methodologies for Behavioral Analysis in Transgenic Mouse Models of Alzheimer's Disease: Parallels with Human AD Evaluation.* Thesis, University of South Florida, 2009.

[63] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonca, "Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Res Notes*, vol. 4, p. 299, 2011.

[64] L. Lemos, *A data mining approach to predict conversion from mild cognitive impairment to Alzheimer's Disease*. Thesis, Master's thesis, IST, 2012.

[65] T. H. Kim, J. H. Park, J. J. Lee, J. H. Jhoo, B.-J. Kim, J.-L. Kim, S. G. Kim, J. Youn, S.-H. Ryu, D. Y. Lee, K. P. Kwak, D. W. Lee, S. B. Lee, S. W. Moon, S. M. Cha, J. Han, Y. s. So, H.-G. Jeong, and K. W. Kim, "Overview of the korean longitudinal study on cognitive aging and dementia," *Alzheimers Dement*, vol. 9, no. 4 suppl., pp. 626–627, 2013.

[66] J. H. Lee, K. U. Lee, D. Y. Lee, K. W. Kim, J. H. Jhoo, J. H. Kim, K. H. Lee, S. Y. Kim, S. H. Han, and J. I. Woo, "Development of the korean version of the consortium to establish a registry for alzheimer's disease assessment packet (cerad-k): clinical and neuropsychological assessment batteries," *J Gerontol B Psychol Sci Soc Sci*, vol. 57, no. 1, pp. P47–53, 2002.

[67] Y. Lecrubier, D. Sheehan, E. Weiller, P. Amorim, I. Bonora, K. H. Sheehan, J. Janavs, and G. Dunbar, "The mini international neuropsychiatric interview (mini). a short diagnostic structured interview: reliability and validity according to the cidi," *European Psychiatry*, vol. 12, no. 5, pp. 224 – 231, 1997.

[68] J. C. Morris, "The clinical dementia rating (cdr): current version and scoring rules," *Neurology*, vol. 43, no. 11, pp. 2412–4, 1993.

[69] D. Y. Lee, K. U. Lee, J. H. Lee, K. W. Kim, J. H. Jhoo, S. Y. Kim, J. C. Yoon, S. I. Woo, J. Ha, and J. I. Woo, "A normative study of the cerad neuropsychological assessment battery in the korean elderly," *J Int Neuropsychol Soc*, vol. 10, no. 1, pp. 72–81, 2004.

[70] D. Wechsler, *Wechsler Memory Scale-Revised*. New York: Psychological Corporation, 1987.

[71] T. H. Kim, Y. Huh, J. Y. Choe, J. W. Jeong, J. H. Park, S. B. Lee, J. J. Lee, J. H. Jhoo, D. Y. Lee, J. I. Woo, and K. W. Kim, "Korean version of frontal assessment battery: psychometric properties and normative data," *Dement Geriatr Cogn Disord*, vol. 29, no. 4, pp. 363–70, 2010.

[72] D. R. Royall, J. A. Cordes, and M. Polk, "Clox: an executive clock drawing task," *J Neurol Neurosurg Psychiatry*, vol. 64, no. 5, pp. 588–94, 1998.

[73] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–5, 2001.

[74] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for dna microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–98, 2005.

[75] C. C. Chiu, S. Y. Chan, C. C. Wang, and W. S. Wu, "Missing value imputation for microarray data: a comprehensive comparison study and a web tool," *BMC Syst Biol*, vol. 7 Suppl 6, p. S12, 2013.

[76] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. Published online 2014; based on TR arXiv:1404.7828 [cs.NE].

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[78] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, pp. 124–141, Jan 2001.

[79] B. Yin, M. Balvert, D. Zambrano, A. Schoenhuth, and S. Bohte, "An image representation based convolutional network for DNA classification," in *International Conference on Learning Representations*, 2018.

[80] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[81] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 07–09 Jul 2015.

[82] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *CoRR*, vol. abs/1707.07012, 2017.

[83] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[84] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, vol. 1996, pp. 148–156, 1996.

[85] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[86] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[87] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[88] P. McCullagh, "Generalized linear models," *European Journal of Operational Research*, vol. 16, no. 3, pp. 285–292, 1984.

[89] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[90] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[91] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27–66, 2012.

[92] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[93] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar, "The mini-international neuropsychiatric interview (m.i.n.i.): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10," *J Clin Psychiatry*, vol. 59 Suppl 20, pp. 22–33;quiz 34–57, 1998.

[94] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Association, 4th ed., 1994.

[95] J. W. Kim, D. Y. Lee, E. H. Seo, B. K. Sohn, S. Y. Park, I. Choo, J. C. Youn, J. H. Jhoo, K. W. Kim, and J. I. Woo, "Improvement of dementia screening accuracy of mini-mental state examination by education-adjustment and supplementation of frontal assessment bat-

tery performance," *Journal of Korean medical science*, vol. 28, no. 10, pp. 1522–1528, 2013.

[96] H. Brodaty, D. Pond, N. M. Kemp, G. Luscombe, L. Harding, K. Berman, and F. A. Huppert, "The gpcog: A new screening test for dementia designed for general practice," *Journal of the American Geriatrics Society*, vol. 50, no. 3, pp. 530–534, 2002.

[97] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, 2017.

[98] P. M. Krueger and E. M. Friedman, "Sleep duration in the united states: A cross-sectional population-based study," *American Journal of Epidemiology*, vol. 169, no. 9, pp. 1052–1063, 2009.

[99] J.-H. Kim, K. R. Kim, K. H. Cho, K.-B. Yoo, J. A. Kwon, and E.-C. Park, "The association between sleep duration and self-rated health in the korean general population," *Journal of Clinical Sleep Medicine*, vol. 9, p. 1057—1064, October 2013.

[100] J. K. Walsh, C. Coulouvrat, G. Hajak, M. D. Lakoma, M. Petukhova, T. Roth, N. A. Sampson, V. Shahly, A. Shillington, J. J. Stephenson, and R. C. Kessler, "Nighttime insomnia symptoms and perceived health in the america insomnia survey (ais)," *Sleep*, vol. 34, no. 8, pp. 997–1011, 2011.

[101] J. Kim, K. In, J. Kim, S. You, K. Kang, J. Shim, S. Lee, J. Lee, S. Lee, C. Park, and C. Shin, "Prevalence of sleep-disordered breathing in middle-aged korean men and women," *American Journal of Respiratory and Critical Care Medicine*, vol. 170, p. 1108—1113, November 2004.

[102] Y. Hwangbo, W. J. Kim, M. K. Chu, C. H. Yun, and K. I. Yang, "Habitual sleep duration, unmet sleep need, and excessive daytime sleepiness in korean adults," *Journal of Clinical Neurology*, vol. 12, p. 194—200, April 2016.

[103] S. Suh, H.-C. Yang, C. P. Fairholme, H. Kim, R. Manber, and C. Shin, "Who is at risk for having persistent insomnia symptoms? a longitudinal study in the general population in korea," *Sleep Medicine*, vol. 15, p. 180—186, February 2014.

[104] S. Joo, I. Baik, H. Yi, K. Jung, J. Kim, and C. Shin, "Prevalence of excessive daytime sleepiness and associated factors in the adult population of korea," *Sleep Medicine*, vol. 10, no. 2, pp. 182 – 188, 2009.

[105] S. M. W. Rajaratnam, M. E. Howard, and R. R. Grunstein, "Sleep loss and circadian disruption in shift work: health burden and management," *The Medical Journal of Australia*, vol. 199, p. S11—5, October 2013.

[106] M. Daley, C. M. Morin, M. LeBlanc, J.-P. Grégoire, and J. Savard, "The economic burden of insomnia: Direct and indirect costs for individuals with insomnia syndrome, insomnia symptoms, and good sleepers," *Sleep*, vol. 32, no. 1, pp. 55–64, 2009.

[107] D. R. Hillman and L. C. Lack, "Public health implications of sleep loss: the community burden," *The Medical Journal of Australia*, vol. 199, p. S7—10, October 2013.

[108] C.-H. Yun, H. Kim, S. K. Lee, S. Suh, S. H. Lee, S.-H. Park, R. J. Thomas, R. Au, and C. Shin, "Daytime sleepiness associated with poor sustained attention in middle and late adulthood," *Sleep Medicine*, vol. 16, p. 143—151, January 2015.

[109] Y. J. Lee, J. Park, S. Kim, S.-J. Cho, and S. J. Kim, "Academic performance among adolescents with behaviorally induced insufficient sleep

syndrome," *Journal of Clinical Sleep Medicine*, vol. 11, p. 61—68, January 2015.

[110] M. L. Zeek, M. J. Savoie, M. Song, L. M. Kennemur, J. Qian, P. W. Jungnickel, and S. C. Westrick, "Sleep duration and academic performance among student pharmacists," *American Journal of Pharmaceutical Education*, vol. 79, p. 63, June 2015.

[111] S. Y. Kim, S.-G. Kim, S. Sim, B. Park, and H. G. Choi, "Excessive sleep and lack of sleep are associated with slips and falls in the adult korean population: A population-based cross-sectional study," *Medicine*, vol. 95, p. e2397, January 2016.

[112] J. Connor, R. Norton, S. Ameratunga, E. Robinson, I. Civil, R. Dunn, J. Bailey, and R. Jackson, "Driver sleepiness and risk of serious injury to car occupants: population based case control study," *British Medical Journal (Clinical research ed.)*, vol. 324, p. 1125, May 2002.

[113] P. M. Forsman, B. J. Vila, R. A. Short, C. G. Mott, and H. P. V. Dongen, "Efficient driver drowsiness detection at moderate levels of drowsiness," *Accident Analysis and Prevention*, vol. 50, pp. 341 – 350, 2013.

[114] D. Sandberg, T. Akerstedt, A. Anund, G. Kecklund, and M. Wahde, "Detecting driver sleepiness using optimized nonlinear combinations of sleepiness indicators," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 97–108, March 2011.

[115] T.-H. Chang, C.-S. Hsu, C. Wang, and L.-K. Yang, "Onboard measurement and warning module for irregular vehicle behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, pp. 501–513, Sept 2008.

[116] S. Ftouni, T. L. Sletten, M. Howard, C. Anderson, M. G. Lenné, S. W. Lockley, and S. M. W. Rajaratnam, "Objective and subjective measures

of sleepiness, and their associations with on-road driving events in shift workers," *Journal of Sleep Research*, vol. 22, p. 58—69, February 2013.

[117] D. F. Dinges and J. W. Powell, "Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations," *Behavior Research Methods, Instruments, & Computers*, vol. 17, pp. 652–655, Nov 1985.

[118] D. F. Dinges, N. L. Rogers, J. Dorrian, and C. A. Kushida, *Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss.*, pp. 39–70. New York, NY: Marcel Dekker, Inc.„ 2005.

[119] J. Lim and D. F. Dinges, "Sleep deprivation and vigilant attention," *Annals of the New York Academy of Sciences*, vol. 1129, p. 305—322, 2008.

[120] M. Basner, D. Mollicone, and D. F. Dinges, "Validity and sensitivity of a brief psychomotor vigilance test (pvt-b) to total and partial sleep deprivation," *Acta Astronautica*, vol. 69, no. 11, pp. 949 – 959, 2011.

[121] T. B. F. A. Management, D. F. Dinges, G. Maislin, J. W. Powell, P. D, and M. M. Mallis, "Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management," 1998.

[122] M. L. Jackson, G. A. Kennedy, C. Clarke, M. Gullo, P. Swann, L. A. Downey, A. C. Hayley, R. J. Pierce, and M. E. Howard, "The utility of automated measures of ocular metrics for detecting driver drowsiness during extended wakefulness," *Accident Analysis and Prevention*, vol. 87, pp. 127 – 133, 2016.

[123] E. Aidman, C. Chadunow, K. Johnson, and J. Reece, "Real-time driver drowsiness feedback improves driver alertness and self-reported driving performance," *Accident Analysis and Prevention*, vol. 81, pp. 8 – 13, 2015.

[124] M. Corbett, "A drowsiness detection system for pilots: Optalert," *Aviation, Space, and Environmental Medicine*, vol. 80, p. 149, February 2009.

[125] M. Fabbri, F. Provini, E. Magosso, A. Zaniboni, A. Bisulli, G. Plazzi, M. Ursino, and P. Montagna, "Detection of sleep onset by analysis of slow eye movements: A preliminary study of mslt recordings," *Sleep Medicine*, vol. 10, no. 6, pp. 637 – 640, 2009.

[126] J. A. Horne and L. A. Reyner, "Counteracting driver sleepiness: effects of napping, caffeine, and placebo," *Psychophysiology*, vol. 33, no. 3, pp. 306–309, 1996.

[127] S. McGuire, U. Müller, E.-M. Elmenhorst, and M. Basner, "Inter-individual differences in the effects of aircraft noise on sleep fragmentation," *Sleep*, vol. 39, no. 5, pp. 1107–1110, 2016.

[128] L. N. Boyle, J. Tippin, A. Paul, and M. Rizzo, "Driver performance in the moments surrounding a microsleep," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 11, no. 2, pp. 126 – 136, 2008.

[129] Z. Mardi, S. N. M. Ashtiani, and M. Mikaili, "Eeg-based drowsiness detection for safe driving using chaotic features and statistical tests," *Journal of Medical Signals and Sensors*, vol. 1, no. 2, pp. 130–137, 2011.

[130] M. V. Yeo, X. Li, K. Shen, and E. P. Wilder-Smith, "Can svm be used for automatic eeg detection of drowsiness during car driving?," *Safety Science*, vol. 47, no. 1, pp. 115 – 124, 2009.

[131] A. G. Correa, L. Orosco, and E. Laciar, "Automatic detection of drowsiness in eeg records based on multimodal analysis," *Medical Engineering Physics*, vol. 36, no. 2, pp. 244 – 249, 2014.

[132] J. Faber, "Detection of different levels of vigilance by eeg pseudo spectra," *Neural Network World*, vol. 14, pp. 285–290, 01 2004.

[133] H. Han and K.-Y. Song, "Electroencephalogram-based driver drowsiness detection system using errors-in-variables (eiv) and multilayer perceptron (mlp)," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 39, no. 10, pp. 887–895, 2014.

[134] T.-P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski, "Estimating alertness from the eeg power spectrum," *IEEE Transactions on Biomedical Engineering*, vol. 44, pp. 60–69, Jan 1997.

[135] N. R. Pal, C.-Y. Chuang, L.-W. Ko, C.-F. Chao, T.-P. Jung, S.-F. Liang, and C.-T. Lin, "Eeg-based subject- and session-independent drowsiness detection: An unsupervised approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 192:1–192:11, Jan. 2008.

[136] C. Papadelis, Z. Chen, C. Kourtidou-Papadeli, P. D. Bamidis, I. Chouvarda, E. Bekiaris, and N. Maglaveras, "Monitoring sleepiness with on-board electrophysiological recordings for preventing sleep-deprived traffic accidents," *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, vol. 118, p. 1906—1922, September 2007.

[137] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Drowsiness detection during different times of day using multiple features," *Australasian Physical & Engineering Sciences in Medicine*, vol. 36, pp. 243–250, Jun 2013.

[138] A. Subasi, "Automatic recognition of alertness level from eeg by using neural network and wavelet coefficients," *Expert Systems with Applications*, vol. 28, no. 4, pp. 701 – 711, 2005.

[139] H. S. Choi, B. Lee, and S. Yoon, "Biometric authentication using noisy electrocardiograms acquired by mobile sensors," *IEEE Access*, vol. 4, pp. 1266–1273, 2016.

[140] S. Choi, S. Kim, J. Seo, J. Y. Park, and S. Yoon, "Wearable and wireless measurement system for evaluating penile tumescence," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, Oct 2015.

[141] H.-S. Choi, S. Kim, J. E. Oh, J. E. Yoon, J. A. Park, C.-H. Yun, and S. Yoon, "Xgboost-based instantaneous drowsiness detection framework using multitaper spectral information of electroencephalography," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, (New York, NY, USA), pp. 111–121, ACM, 2018.

[142] M. Y. Khitrov, S. Laxminarayan, D. Thorsley, S. Ramakrishnan, S. Rajaraman, N. J. Wesensten, and J. Reifman, "Pc-pvt: A platform for psychomotor vigilance task testing, analysis, and prediction," *Behavior Research Methods*, vol. 46, pp. 140–147, Mar 2014.

[143] C. François, J. Wertz, M. Kirkove, and J. G. Verly, "Evaluation of the performance of an experimental somnolence quantification system in terms of reaction times and lapses," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5820–5823, Aug 2014.

[144] A. A. Putilov and O. G. Donskaya, "Construction and validation of the eeg analogues of the karolinska sleepiness scale based on the karolinska drowsiness test," *Clinical Neurophysiology*, vol. 124, no. 7, pp. 1346 – 1352, 2013.

[145] C. E. D. Alloway, R. D. Ogilvie, and C. M. Shapiro, "The alpha attenuation test: Assessing excessive daytime sleepiness in narcolepsy-cataplexy," *Sleep*, vol. 20, no. 4, pp. 258–266, 1997.

[146] E. M. Whitham, K. J. Pope, S. P. Fitzgibbon, T. Lewis, C. R. Clark, S. Loveless, M. Broberg, A. Wallace, D. DeLosAngeles, P. Lillie, A. Hardy, R. Fronsko, A. Pulbrook, and J. O. Willoughby, "Scalp electrical recording during paralysis: Quantitative evidence that eeg frequencies above 20hz are contaminated by emg," *Clinical Neurophysiology*, vol. 118, no. 8, pp. 1877 – 1888, 2007.

[147] C.-T. Lin, R.-C. Wu, S.-F. Liang, W.-H. Chao, Y.-J. Chen, and T.-P. Jung, "Eeg-based drowsiness estimation for safety driving using independent component analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, pp. 2726–2738, Dec 2005.

[148] S.-H. Hsu and T.-P. Jung, "Monitoring alert and drowsy states by modeling eeg source nonstationarity," *Journal of Neural Engineering*, vol. 14, no. 5, p. 056012, 2017.

[149] P.-Y. Tsai, W. Hu, T. B. Kuo, and L.-Y. Shyu, "A portable device for real time drowsiness detection using novel active dry electrode system," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3775–3778, Sept 2009.

[150] G. Li and W.-Y. Chung, "Estimation of eye closure degree using eeg sensors and its application in driver drowsiness detection," *Sensors*, vol. 14, no. 9, pp. 17491–17515, 2014.

[151] M. Awais, N. Badruddin, and M. Drieberg, "A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability," *Sensors*, vol. 17, no. 9, 2017.

[152] D. Ribeiro, C. Teixeira, and A. Cardoso, "Eeg-based drowsiness detection platform to compare different methodologies," in *2017 4th Experiment@International Conference*, pp. 318–322, June 2017.

[153] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, Jun 1967.

[154] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, pp. 1055–1096, Sept 1982.

[155] M. J. Prerau, R. E. Brown, M. T. Bianchi, J. M. Ellenbogen, and P. L. Purdon, "Sleep neurophysiological dynamics through the lens of multitaper spectral analysis," *Physiology*, vol. 32, no. 1, pp. 60–92, 2017. PMID: 27927806.

[156] C. Adam-Bourdarios, G. Cowan, C. Germain-Renaud, I. Guyon, B. Kégl, and D. Rousseau, "The higgs machine learning challenge," *Journal of Physics*, vol. 664, no. 7, p. 072015, 2015.

[157] G. R. Poudel, C. R. Innes, P. J. Bones, R. Watts, and R. D. Jones, "Losing the struggle to stay awake: Divergent thalamic and cortical activity during microsleeps," *Human Brain Mapping*, vol. 35, no. 1, pp. 257–269, 2014.

[158] W. Ting, Y. Guo-zheng, Y. Bang-hua, and S. Hong, "Eeg feature extraction based on wavelet packet decomposition for brain computer interface," *Measurement*, vol. 41, no. 6, pp. 618 – 625, 2008.

[159] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, 2015.

[160] O. Omeni, A. C. W. Wong, A. J. Burdett, and C. Toumazou, "Energy efficient medium access protocol for wireless medical body area sensor networks," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 2, pp. 251–259, Dec 2008.

[161] M. H. Rehmani, A. Rachedi, S. Lohier, T. Alves, and B. Poussot, "Intelligent antenna selection decision in ieee 802.15. 4 wireless sensor networks: An experimental analysis," *Computers & Electrical Engineering*, vol. 40, no. 2, pp. 443–455, 2014.

[162] F. Akhtar and M. H. Rehmani, "Energy replenishment using renewable and traditional energy resources for sustainable wireless sensor networks: A review," *Renewable and Sustainable Energy Reviews*, vol. 45, pp. 769–784, 2015.

[163] B. Shrestha, E. Hossain, and S. Camorlinga, "Ieee 802.15. 4 mac with gts transmission for heterogeneous devices with application to wheelchair body-area sensor networks," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 15, no. 5, pp. 767–777, 2011.

[164] P. Honeine, F. Mourad, M. Kallas, H. Snoussi, H. Amoud, and C. Francis, "Wireless sensor networks in biomedical: Body area networks," in *Systems, Signal Processing and their Applications (WOSSPA), 2011 7th International Workshop on*, pp. 388–391, IEEE, 2011.

[165] D. P. Tobon, T. H. Falk, and M. Maier, "Context awareness in wbans: a survey on medical and non-medical applications," *Wireless Communications, IEEE*, vol. 20, no. 4, pp. 30–37, 2013.

[166] S. K. Ghosh, S. Chakraborty, A. Jamthe, and D. P. Agrawal, "Comprehensive monitoring of firefighters by a wireless body area sensor network," in *Wireless and Optical Communications Networks (WOCN), 2013 Tenth International Conference on*, pp. 1–6, IEEE, 2013.

[167] S. N. Ramli, R. Ahmad, M. F. Abdollah, and E. Dutkiewicz, "A biometric-based security for data authentication in wireless body area network (wban)," in *Advanced Communication Technology (ICACT), 2013 15th International Conference on*, pp. 998–1001, IEEE, 2013.

[168] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *Instrumentation and Measurement, IEEE Transactions on*, vol. 50, no. 3, pp. 808–812, 2001.

[169] M. Abo-Zahhad, S. M. Ahmed, and S. Abbas, "Biometric authentication based on pcg and ecg signals: present status and future directions," *Signal, Image and Video Processing*, vol. 8, no. 4, pp. 739–751, 2014.

[170] C. Ye, M. T. Coimbra, and B. Kumar, "Investigation of human identification using two-lead electrocardiogram (ecg) signals," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pp. 1–8, IEEE, 2010.

[171] T.-W. Shen, W. J. Tompkins, and Y. H. Hu, "Implementation of a one-lead ecg human identification system on a normal population," *Journal of Engineering and Computer Innovations*, vol. 2, no. 1, pp. 12–21, 2011.

[172] S. Poornachandra, "Wavelet-based denoising using subband dependent threshold for ecg signals," *Digital signal processing*, vol. 18, no. 1, pp. 49–55, 2008.

[173] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "Ecg to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[174] L. Sörnmo and P. Laguna, *Bioelectrical ignal processing in cardiac and neurological applications*. Academic Press, 2005.

[175] K. N. Plataniotis, D. Hatzinakos, and J. K. Lee, "Ecg biometric recognition without fiducial detection," in *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*, pp. 1–6, IEEE, 2006.

[176] F. Agrafioti and D. Hatzinakos, "Ecg based recognition using second order statistics," in *Communication Networks and Services Research Conference, 2008. CNSR 2008. 6th Annual*, pp. 82–87, IEEE, 2008.

[177] Y. Wan, J. Yao, *et al.*, "A neural network to identify human subjects with electrocardiogram signals," in *Proceedings of the world congress on engineering and computer science*, pp. 1–4, Citeseer, 2008.

[178] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[179] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 20, no. 3, pp. 45–50, 2001.

[180] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of Neuroengineering and Rehabilitation*, vol. 9, no. 1, p. 21, 2012.

[181] J. K. Cooper, "Electrocardiography 100 years ago," *New England Journal of Medicine*, vol. 315, no. 7, pp. 461–464, 1986.

[182] T. Lugovaya, "Biometric human identification based on electrocardiogram," *Master's Thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation*, 2005.

[183] S. K. Sahoo, T. Choubisa, and S. M. Prasanna, "Multimodal biometric person authentication: A review," *IETE Technical Review*, vol. 29, no. 1, pp. 54–75, 2012.

[184] C. Free, G. Phillips, L. Galli, L. Watson, L. Felix, P. Edwards, V. Patel, and A. Haines, "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review," *PLOS Medicine*, vol. 10, no. 1, p. e1001362, 2013.

[185] L. A. Saxon, A. Smith, S. Doshi, J. Dinsdale, and D. Albert, "Iphone rhythm strip–the implications of wireless and ubiquitous heart rate monitoring," *Journal of the American College of Cardiology*, vol. 59, no. 13s1, p. E726, 2012.

[186] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[187] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[188] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[189] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[190] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2nd ed., 1998.

[191] D. S. Broomhead and D. Lowe, "Radial basis functions, multi-variable functional interpolation and adaptive networks," Tech. Rep. 39, DTIC Document, 1988.

[192] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[193] S. Chaudhuri, T. D. Pawar, and S. Duttagupta, *Ambulation Analysis in Wearable ECG*. Springer Publishing Company, Incorporated, 1st ed., 2009.

[194] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, pp. 230–236, March 1985.

[195] B. J. Drew, R. M. Califf, M. Funk, E. S. Kaufman, M. W. Krucoff, M. M. Laks, P. W. Macfarlane, C. Sommargren, S. Swiryn, and G. F. Van Hare, "Practice standards for electrocardiographic monitoring in hospital settings an american heart association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young: Endorsed by the international society of computerized electrocardiology and the american association of critical-care nurses," *Circulation*, vol. 110, no. 17, pp. 2721–2746, 2004.

[196] M. Blanco-Velasco, B. Weng, and K. E. Barner, "Ecg signal denoising and baseline wander correction based on the empirical mode decomposition," *Computers in Biology and Medicine*, vol. 38, no. 1, pp. 1–13, 2008.

[197] M. Alfaouri and K. Daqrouq, "Ecg signal denoising by wavelet transform thresholding," *American Journal of Applied Sciences*, vol. 5, no. 3, pp. 276–281, 2008.

[198] X. Ning and I. W. Selesnick, "Ecg enhancement and qrs detection based on sparse derivatives," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 713–723, 2013.

[199] I. Odinaka, P.-H. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh, "Ecg biometric recognition: A comparative analysis,"

*Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 6, pp. 1812–1824, 2012.

[200] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[201] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

# 초 록

딥러닝을 포함한 기계학습 기법들은 여러 분야 전반에 거쳐 최고의 성능을 갱신하고 있다. 하지만 의료 진단 등과 같은 실제 문제에 적용하기 위해서는 해결해야할 사안들이 여전히 남아 있다. 대표적인 사안 중 하나로 본 논문에서 다루고자 하는 사안은 데이터 불균형성이다. 데이터 불균형성이란, 축적된 데이터들 중 특정군의 분포가 매우 많거나 매우 적은 상태를 지칭한다. 불균형성이 존재하는 데이터를 기반으로 학습이 진행될 경우, 다수 데이터에 치우친 학습 경향을 보이기 때문에 소수 데이터에 대한 성능이 저하될 위험성이 있다.

본 논문에서는 데이터 불균형성을 극복하고자 하는 기존 방법들을 논하고, 이들의 생성적 적대 신경망 기법을 활용하여 한계를 극복할 수 있는 새로운 방법론을 제시한다. 해당 기법은 생성 모델과 분류 모델의 유기적 학습을 통해 생성된 데이터들이 소수 데이터에 대한 성능 향상을 유도하도록 한다. 부가적으로 실 환경 데이터를 활용하여 바이오메디컬 분야의 세 가지 응용 연구를 수행하였다. 각각 연구의 유의성과 함께 데이터 불균형성의 영향과 해결 방안에 대해 논하였다. 각각의 응용 연구는 신경심리검사를 활용한 치매 조기 진단, 뇌파 기반의 극도 졸음 탐지, 심전도 기반 생체 인증에 해당한다. 요약하자면, 본 논문은 데이터 불균형으로 인한 학습의 어려움을 실제 응용 연구들을 통하여 확인하고, 이를 해결하기 위한 방법론들을 탐구하였다.

**주요어**: 불균형 데이터, 기계학습, 심층학습, 생성적 적대 신경망, dementia diagnosis, drowsiness detection, biometric authentication, electroencephalography, electrocardiogram

**학번**: 2013-23144