The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 4, pp.91-98, July 31, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.4.91

JCCT 2022-7-12

VAE(Variational AutoEncoder) 기반 머신러닝 모델을 활용한 체중 라이프로그 이상탐지에 관한 연구

Study on Lifelog Anomaly Detection using VAE-based Machine Learning Model

김지용*, 박민서**

Jiyong Kim*, Minseo Park**

요 약 웨어러블 기기를 통해 지속적으로 수집되는 라이프로그 데이터는 많은 이상값을 포함할 수 있으므로 데이터 품질을 향상시키기 위해서는 이상값을 찾아 제거하는 것이 필요하다. 일반적으로 이상치의 개수가 정상 데이터의 개수보다 적기 때문에 클래스 불균형 문제가 발생한다. 이러한 불균형 문제를 해결하기 위해 Variational AutoEncoder를 outlier에 적용하는 방법을 제안한다. 제안된 방법으로 이상치 데이터를 전처리한 후, 다수의 머신러닝 모델(분류)을 통해 검증한다. 체중 데이터를 이용한 검증 결과, 모든 분류 모델에서 성능이 향상됨을 확인하였다. 실험 결과를 바탕으로 라이프로그 체중 데이터 분석 시 본 연구에서 제안한 이상치 처리 방법을 이용하여 데이터를 전처리한 후 성능이 가장 좋은 LightGBM 모델을 적용할 것을 제안한다.

주요어 : 이상탐지, VAE 기법, 체중 라이프로그, 클래스 불균형, 트리기반 머신러닝 모델

Abstract Lifelog data continuously collected through a wearable device may contain many outliers, so in order to improve data quality, it is necessary to find and remove outliers. In general, since the number of outliers is less than the number of normal data, a class imbalance problem occurs. To solve this imbalance problem, we propose a method that applies Variational AutoEncoder to outliers. After preprocessing the outlier data with proposed method, it is verified through a number of machine learning models(classification). As a result of verification using body weight data, it was confirmed that the performance was improved in all classification models. Based on the experimental results, when analyzing lifelog body weight data, we propose to apply the LightGBM model with the best performance after preprocessing the data using the outlier processing method proposed in this study.

Key words : Anomaly Detection, Variational AutoEncoder, Body Weight, Imbalance Problem, Machine Learning

Ⅰ. 서 론

최근 코로나 팬데믹에 의해 다양한 웨어러블이 건강 관리부터 패션 아이템[1]까지 다양한 영역에서 깊게 자리 잡았다. 디지털 헬스케어 분야에서 인공지능을 활용한 라이프로그 연구 및 분석도 많이 진행되고 있으며[2, 3] 그 중 체중은 라이프로그 데이터 중 다양한 만성질환 및 질병 예방에 있어서 중요한 데이터이다[4][5]. 지금까지

*준회원, 광운대학교 수학과 학사 (제1저자)
**정회원, 서울여자대학교 데이터사이언스학과 조교수 (교신저자)
접수일: 2022년 5월 26일, 수정완료일: 2022년 6월 21일
게재확정일: 2022년 7월 2일

Received: May 26, 2022 / Revised: June 21, 2022
Accepted: July 2, 2022
**Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Seoul, Korea

체중 변동에 대한 다양한 연구가 시도되었다. Ornstein-Uhlenbeck 과정이 체중 변동에 가장 크게 영향을 미친다는 것을 제시한 연구[6], 계절적 변화에 따른 체중 변동[7][8] 그리고 주간 패턴 따른 주기적인 체중 변동[9][10]에 대한 연구와 같이 체중 변동에 대한 많은 연구가 시도 되었고, 체중이 주기적인 변동 성질을 가지고 있다고 제시되었다. 다양한 만성질환을 예방을 위해서는 이러한 체중의 변동성을 지속적으로 모니터링하는 것이 매우 효과적[11][12]이다. 최근에는 스마트 체중계가 보편화되면서 많은 사람들이 자신의 체중을 스마트폰과 연동해 손쉽게 확인 및 기록할 수 있게 되었다. 그러나, 이렇게 수집되는 데이터 중에는 사용자들이 양말을 신고 체중계에 올라가 측정을 하거나, 기기 오류, 통신 오류, 또는 원 사용자가 아닌 다른 사용자가 체중계에 올라가 측정하는 등의 이슈가 발생될 수 있다. 이는 결측값 또는 이상값을 발생시켜 데이터의 정확성과 일관성을 저해해 데이터 품질에 악영향을 끼칠 수 있다. 따라서 높은 품질의 라이프로그 데이터를 수집하기 위해서는 이러한 노이즈를 찾아 제거할 필요가 있다. 단순 반복적인 수작업 방식의 데이터 탐색을 통한 이상값 확인을 하기에는 데이터 수가 방대하므로 이상 탐지 알고리즘을 활용한 이상값 탐지가 필요하다.

이상값의 개수는 일반적으로 정상 데이터 수에 비해 적어 클래스 불균형이 발생된다. 클래스가 적은 데이터에 대한 데이터 불균형을 보완 및 해소하는 기법들이 다양하게 사용되고 있다. 체중의 주기성을 반영하여 데이터를 생성하는 방식[13]으로 해소할 수 있고, 증강 기법으로 일반적으로 많이 사용되는 SMOTE (Syntheticminority Oversampling Technique)와 GAN (Genrative Adversairal Networks) 등의 기법을 활용하여 데이터 증강을 통해 불균형 클래스 문제를 해소할 수 있다[14][15]. 그러나, SMOTE 기법의 경우 샘플 겹침(Sample Overlapping), 잡음 간섭(Noise Interference) 및 잘못된 이웃 선택과 같은 이슈가 발생할 수 있고, GAN 기법의 경우는 학습의 불안정(Instability)과 수렴하지 않는 문제(Non-Convergence)가 있을 수 있다[16][17].

대부분의 이상값 탐지 관련 데이터 증강 연구에서는 정상값을 증강하여 정상적인 데이터와 차이가 큰 값을 탐지하는 방식을 활용하고 있다. 이는 정상 데이터와 이상값 데이터의 차이가 비교적 큰 데이터 셋에서는 매우 효과적[18][19]이나, 일별 기준으로 변동성이 크지 않거나 주기성이 있는 체중과 같은 라이프로그 데이터에서는 관련 이상값을 탐지하는데 한계가 있을 수 있다. 따라서, 본 연구에서는 이상값을 증폭하여 클래스 불균형을 해소하여 비지도 학습 방식보다 정확도가 상대적으로 높은 지도 학습 방식의 머신러닝을 활용해 이상값 분류 방법론을 제안한다. 이때, SMOTE와 GAN 방식에 비해 상대적으로 이슈가 적고 이상값의 특성을 반영하여 데이터를 생성할 수 있는 VAE(Auto-Encoding Variational Bayes)[20]를 활용하는 방법을 제안한다.

제안하는 방법의 우수성을 증명하기 위해, 라이프로그 데이터 중 체중 데이터를 활용하여, 데이터 증강 처리 전과 후의 모델의 성능을 비교하였다. 데이터 불균형을 가지고 있는 증강 처리 전의 데이터 기반 실험에서도, 성능을 최대치로 높이기 위해 비지도 학습 방식의 Isolation Forest 모델을 활용해 성능 비교를 진행하였고, 제안 방식 적용 후의 데이터 셋(이상값이 충분히 생성되어 클래스 불균형 문제가 해소된 데이터 셋)에 대해서는 기 증명된 우수한 머신러닝 알고리즘을 다양하게 실험 비교하였다. 실험 결과, RandomForest 모델[21], Support Vector Machine 모델, LightGBM 모델, 그리고 XGBoost 모델과 같이 대부분의 머신러닝 알고리즘에서 우수한 성능을 보였다.

본 연구의 구성은 다음과 같다. 2장에서는 관련 선행연구를 기술한다. 데이터 불균형 해소 기법에 주로 사용되는 데이터 증강 기법에 대해 검토한다. 3장에서는 본 연구에서 제안하는 이상값에 VAE를 적용하여 데이터 불균형을 해소한 후 라이프로그 데이터를 분석하는 알고리즘을 기술한다. 데이터 셋, 데이터 탐색, 데이터 모델링, 성능 지표, 그리고 실험 결과 순으로 설명한다.

4장에서는 결론 및 향후 개선 방향에 관하여 기술한다.

II. 선행연구

1. SMOTE

SMOTE 알고리즘은 KNN(K-Nearest Neighbor) 최근접 알고리즘을 활용해 합성 데이터를 생성하는 방식으로 데이터의 개수가 적은 클래스의 임의의 데이터 포인트 X에 대해 가장 가까운 K 개의 이웃 포인트를 탐색하고, 탐색한 K 개의 이웃 포인트와 데이터 포인트

X 사이에 임의의 새로운 데이터 포인트 X'를 생성하는 것이다[22]. 대표적인 방법으로는 임의의 새로운 데이터가 생성하는 것 대신, 분류하기 어려운 클래스의 경계 쪽에 위치한 데이터들을 생성하는 Borderline-SMOTE [23]와 클래스의 데이터 밀도에 따라 합성 데이터를 생성하는 방식으로 작동해 소수 클래스의 저밀도 부분에 상대적으로 데이터를 더 많이 합성하는 ADASYN[24] 등이 있다.

2. VAE (Auto-EncodingVariational Bayes)

VAE(Auto-EncodingVariational Bayes)는 랜덤 노이즈로부터 원하는 데이터를 얻기 위해 데이터의 확률분포 내에서 랜덤 값을 생성하는 방식을 활용한다[25].

VAE는 AE 구조와 같이 두 개의 뉴럴 네트워크로 구성되었다. 구체적으로 encoder라는 뉴럴 네트워크에서는 데이터를 입력받아 잠재변수(latent variable) Z를 만들고, decoder라고 불리는 뉴럴 네트워크에서는 잠재변수 Z를 통해 데이터를 복원하여 입력 데이터와 유사하지만 비선형적으로 더 낮은 차원의 데이터를 얻을 수 있다. 이때, VAE는 다음과 같은 식의 Variational Inference 방법을 활용한다[26].

$$\log\ p(x) \geq E_{z\sim q(z)}[\log p(x|z)] - D_{KL}(q(z)||p(z)) \quad (1)$$

해당 방법은 원 데이터로 복원이 잘 되기 위해 주어진 데이터 x에 대해 Z를 잘 샘플링 할 수 있는 이상적인 확률 분포 p(Z|x)를 찾아야 하지만 이상적인 확률분포를 모르므로 이를 추정하기 위해 다루기 쉬운 분포 $q(z)$를 가정하고 해당 확률분포의 모수를 업데이트하면서 최대한 근사하게 만들어 이상적인 확률 분포 대신 사용하는 방법이다. 여기서 $q(z)$는 다음과 같이 KL Divergence를 활용해 구할 수 있다.

$$D_{KL}(P|Q) = E_{x\sim P}[\log\frac{P}{Q}] = E_{x\sim p}[-\log\frac{Q}{P}] \quad (2)$$

식(2)은 두 확률 분포의 차이를 계산하는 것으로 여기서 $p(z|x)$와 $q(z)$ 사이의 차이 계산에 적용하면 다음과 같다[14].

$$D_{KL}(q(z)||p(z|x)) = D_{KL}(q(z)\|p(z)) + \log p(x) - E_{z\sim q(z)}[\log p(x|z)] \quad (3)$$

즉, 두 확률 분포의 차이를 줄어드는 방향으로 $q(z)$를 업데이트하는 과정을 반복해 최종적으로 잘 근사하는 $q^*(z)$를 얻는다.

한편, 일반적으로 $q(z)$를 다음과 같은 정규분포로 가정한다.

$$q(z) = N(\mu_q(x), \sigma_q^2) \quad (4)$$

그러나, 입력되는 데이터가 고차원인 경우 학습이 어려워지는 단점이 있어 다음과 같이 설정한다[27].

$$z = \mu(x) + \sigma(x) \times \epsilon \quad (5)$$

$$\epsilon \sim N(0,1) \quad (6)$$

이러한 설정과 함께 VAE의 encoder에는 데이터를 받아 Z의 평균과 분산을 생성하는 두 개의 뉴럴 네트워크가 있어 해당 뉴럴 네트워크가 산출한 평균과 분산을 통해 Z를 생성하는 Reparameterization Trick을 활용한다. 즉, 정규 분포 형태로 나타내는 latent variable 분포를 통해 샘플링된 변수(variables)를 디코더에 입력하며 Reparameterization Trick을 활용해 미분 가능하게 하여 Back propagation을 통한 학습이 가능하도록 한다[28][29].

Evidence of Lower Bound(ELBO)는 두 분포의 차이를 최소화하기 위해 사용되며[26][27][30], VAE에서 기존 KL Divergence 식(3)에 대해 $q(z)$를 정규분포로 가정한 식(5)와 식(6)을 반영한 식은 다음과 같다.

$$\log p(x) = E_{z\sim q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)\|p(z)) + D_{KL}q((z|x)||p(z|x)) = ELBO + D_{KL}(q(z|x)||p(z|x)) \quad (7)$$

이때, 식 (7)의 우변에 있는 $D_{KL}q((z|x)||p(z|x))$은 항상 양수이므로 다음과 같은 부등식이 항상 성립한다.

$$\log p(x) \geq E_{z\sim q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z) = ELBO \quad (8)$$

식(7)과 식(8)을 통해 ELBO(evidence lower bound)를 최대화는 것은 두 확률 분포의 차이를 최소화하는 것임을 확인할 수 있다.

이에 대한 일반적인 손실함수 형태로 정리하면 다음과 같다.

$$L = -E_{z\sim q(z|x)}[\log p(x|z)] + D_{KL}(q(z|x)||p(z)) \quad (9)$$

3. 기존 VAE를 활용한 연구사례

VAE를 활용한 데이터 증강 관련 응용 연구로는 당뇨병 환자를 예측하기 위해 VAE를 활용해 전체 데이터를 증강하고 SAE(Sparse Autoencoder)를 활용해 성능을 높인 연구들[31][32], 호흡음을 통해 질환을 감지하는 연구[33] 등이 있는데, 이들 모두 VAE 기법을 활용해 데이터 불균형을 해소하여 높은 성능을 보였다. 각종 기법에 대한 비교 연구 중 사람의 말과 언어에 대해 CGAN(Conditional GAN), VAE 및 VAE-SGAN을 사용하여 데이터 증강 성능을 비교한 연구[34]에서 VAE의 성능이 가장 우수한 것으로 나타났다.

III. 제안하는 알고리즘

1. 데이터 셋

본 연구는 ㈜지아이비타 비타민즈앱과 삼성 갤럭시 워치(Samsung Galaxy watch active 2)을 통해 수집된 라이프로그 데이터 중에서 체중, 체지방(FAT), 근육량, 그리고 체수분과 같은 신체 데이터 셋과 신장, 성별 및 생년월일과 같은 사용자 정보 데이터 셋을 사용하였고, 개별적으로 70개 행 이상을 가진 사용자 282명의 라이프로그 데이터를 사용하였다. 신체 데이터 셋과 사용자 정보 데이터 셋을 USER_CODE(사용자 식별코드) 변수와 DATE(데이터 수집 일지) 변수를 기준으로 병합하여 최종 데이터 셋을 생성하였다. 시간에 대한 정보를 추가하기 위해 '2021년 12월 9일 (기준: 최대한 최근 날짜)를 기준으로 수집된 데이터의 날짜와의 차이인 DATE_DIFF를 파생 변수로 생성하였다. 파생한 변수를 포함한 전체 변수 설명은 표 1로 정리하였다.

표 1. 최종 데이터셋의 파생 변수 및 설명
Table 1. Features and description of the final data set

Feature	Description	Feature	Description
DATE	Date the data was collected	USER_ CODE	Identification code
WEIGHT	weight value	BMI	bmi value
FAT	body fat	WATER	amount of water
MUSCLE	body muscle mass	HEIGHT	height
GENDER	gender	AGE	age
DAY_DIFF	Number of days from data collection date to December9, 2021		

2. 데이터 탐색

본 연구에서 사용되는 라이프로그 데이터(체중) 셋의 학습 데이터와 테스트 데이터는 0.75 대 0.25로 나누어 실험 및 테스트 하였다. 원 학습 데이터(데이터 불균형 이슈 처리 전 데이터; 그림1), 정상값을 증강한 학습 데이터(그림2), 이상값 증강한 학습 데이터(그림3)를 사용한다. 세 가지 데이터 셋 모두 데이터를 증강하는 경우에도 아래 Figure 1에서 확인할 수 있듯이, Variance & Bias 이슈를 발생하지 않는다.

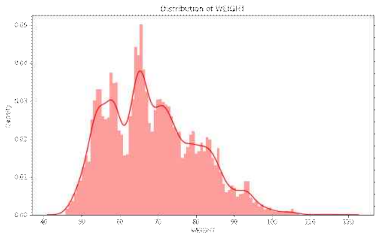

그림 1. 원 학습 데이터
Figure 1. RAW DATA

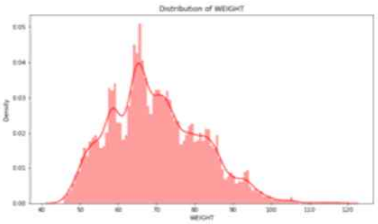

그림 2. 정상값을 증강한 학습 데이터
Figure 2. Data augmented for Normal Values

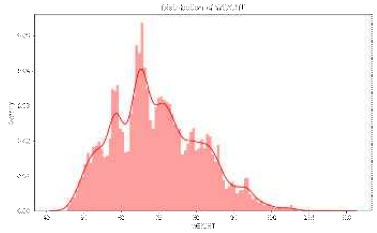

그림 3. 이상값 증강한 학습 데이터
Figure 3. Data augmented for Outlier.

3. 데이터 모델링

본 연구의 전체적인 모델링 프로세스와 제안 방법을 그림 4 로 정리하였다. 본 연구에서는 원 데이터 및 정상 데이터를 증강한 데이터 셋에 Isolation Forest(IF), 정상 데이터를 증강한 데이터 셋에 OneClassSVM을 활용해 성능을 확인하였고, 이상값을 증강한 데이터 셋에

머신러닝 알고리즘 Support Vector Machine(SVM),

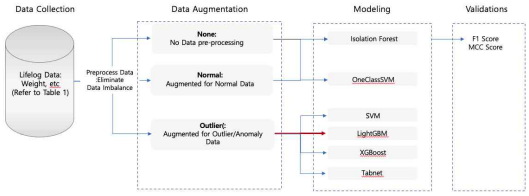

그림 4. 제안 방법론 요약 그래프
Figure 4. Proposal Summary Graph

LightGBM, XGBoost, 그리고 Tabnet 알고리즘을 적용해 성능을 비교 분석하였다. 각 모델의 하이퍼파라미터 튜닝은 조합 수에 따라 GridSearch 또는 Hyeropt을 활용하였고, 구체적은 각 모델의 하이퍼파라미터 값은 표 2와 같다.

표 2. 실험 모델들의 하이퍼파라미터 정리 표른 성능 표
Table 2. Hyperparameters of experimental models

Model	Hyperparameter
IF	contamination = 23931034482758617, max_features = 0.01
IF (normal)	contamination = 0.22482758620689652, max_features = 0.01
SVM (outlier)	C = 1000, gamma = 0.2, kernel = 'rbf'
XGBoost (outlier)	colsample_bylevel=0.6289801688467676 colsample_bytree=0.5848482288983499, gamma=3, learning_rate =0.020520818830002494, max_depth=2, min_child_weight=4, n_estimators = 803, scale_pos_weight = 3, subsample = 0.5101377375161377
IF(outlier)	contamination = 0.41310344827586204 max_features = 0.01
OneClassSVM	nu = 0.1, kernel = 'rbf', gamma = 'auto'
LightGBM (outlier)	colsample_bytree=0.542, min_child_samples=430, min_child_weight=10, n_estimators=1000, num_leaves=44, reg_alpha=1, reg_lambda=100, subsample=0.765
Tabnet (outlier)	mask_type = 'entmax', n_da = 60, n_steps = 3, gamma = 1.4, n_shared =3 lambda_sparse = 6.573608749970886 4e-06, patienceScheduler = 5, patience = 30, epochs = 19

4. 실험 평가를 위한 성능 지표

본 연구에서는 이진 분류 문제 중 불균형 클래스 관련 연구[35][36][37]에 일반적으로 사용되는 F1 score[38][39] 지표와 Matthews correlation coefficient [40] 지표를 사용하였는데, 각각의 공식은 다음과 같다.

$$\frac{2\times TP}{2\times TP+FP+FN} \tag{1}$$

$$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{2}$$

여기서 TP는 True Positive, FP는 False positive, FN은 False negative, TN은 True negative이다.

두 지표의 특징 및 차이점은 (1)과 (2)에서 확인할 수 있듯이 F1 score은 True negative의 수를 고려하지 않는 반면 MCC는 Confussion Matrix의 4개 요소를 모두 고려하여 성능을 산출한다.

Ⅳ. 실험 및 결과

데이터 전처리를 하지 않아 클래스가 불균형한 데이터 셋(None), 이상값 데이터를 증강하여 클래스 불균형을 해소한 데이터 셋(Outlier), 그리고 정상값 데이터를

표 3. 데이터 증강에 따른 성능 표
Table 3. Performance Validations(FI Score, MCC Score) according to Data Augmentation

Data Augmentation	class balance	Model	F1	MCC
None	data imbalance	IF	0.38	0.28
		OneClassSVM	0.28	0.14
		SVM	0.958	0.953
		LightGBM	0.840	0.962
		XGBoost	0.760	0.962
		Tabnet	0.954	0.949
Normal data Augmentaion	Extreme data imbalance	IF	0.37	0.26
		OneClassSVM	0.39	0.37
Outlier data Augmentation	Eliminate data imbalance	SVM	0.979	0.977
		LightGBM	**0.982**	**0.977**
		XGBoost	0.962	0.977
		Tabnet	0.969	0.965

증강하여 불균형을 극대화한 데이터 셋(Normal)에 다양한 알고리즘을 활용해 실험한 결과, 전통적인 이상탐지 알고리즘인 Isolation Forest 경우, 전처리하지 않은 데이터 셋에서 우수한 성능(0.38 (F1 score), 0.28 (MCCscore))을 보였다. 그러나, 전체적으로 성능이 매우 낮은 수준이다. OneClassSVM 경우 정상값을 증강한 데이터에 적용할 경우 상대적으로 우수한 성능(0.39(F1 score), 0.37(MCCscore))을 보였지만 역시 전체적으로 낮은 수준의 성능을 보였다. 마지막으로, 이상값 데이터에 전처리를 한 경우, 전처리를 하지 않은 데이터 셋에 비해 대부분의 머신러닝 알고리즘에서 매우 우수한 성능을 보였다. 그중에 LightGBM 알고리즘의 F1 score 0.982, MCC score0.977로 가장 우수한 것으로 나타났다.

Ⅴ. 결 론

본 연구에서는 일별 변동성이 작고 주기성을 보이는 라이프로그 데이터 (체중 데이터)에 대해 일괄적인 임곗값 바탕의 Rule-based Method가 아닌 머신러닝 알고리즘을 적용해 이상 탐지 및 모델링을 제안하였다. 라이프로그 분석 시 일반적으로 발생되는 클래스 불균형 이슈들을 이상값에 VAE(Auto-Encoding Variational Bayes)을 적용해 해소한 다음, 머신러닝 알고리즘을 적용하는 것을 제안하였다.

2장에서는 이상값 탐지에 사용되는 여러 가지 방법론과 불균형 클래스를 처리하는 방식에 대해 검토하였고, VAE에 대한 수학적 배경을 기술하였다. 3장에서는 본 논문에서 제안하는 기법을 기술하였는데 구체적으로 VAE 기법을 통해 이상값 증강 및 생성을 통해 클래스 불균형 문제를 해소하는 방법을 기술하였다. 이를 통해, 기 증명된 성능이 우수한 머신러닝 분류 알고리즘을 활용하여 학습할 수 있도록 데이터의 품질을 높였다. 웨어러블 장치로부터 수집되는 데이터 셋을 활용하여, 제안 방법의 성능을 비교 평가하였다. 실험 결과, 본 연구에서 제안하는 VAE 기법을 적용해 학습 데이터의 클래스 불균형을 해소한 후, 머신러닝 알고리즘에 적용할 경우, 라이프로그 데이터 분석 성능이 매우 우수한 것으로 나타났다. 따라서, 이상 탐지 관련한 특정 알고리즘을 사용하는 것도 좋지만 라이프로그 분석 및 이상 탐지 관련 분석에서는 데이터 불균형을 처리한 후, 머신러닝 알고리즘을 적용해 보는 것을 추천한다.

본 논문에서 사용된 알고리즘의 성능은 라이프로그 데이터에서 신규 사용자의 경우, 학습 데이터 자체가 없거나 매우 적은 경우 모델 성능의 차이가 발생할 수 있다. 따라서 추후, 실제 서비스 환경에서 흔히 발생 가능한 이상값 데이터 수 부족 관한 문제를 시간의 정보를 고려해 정교하게 생성하는 알고리즘 개발 및 최적화 연구를 진행할 필요가 있다.

References

[1] Park, H.-S., and Moon, P.-J. (2021). A Study on Fashion Items to Prevent COVID-19 Using Wearable Technology. International Journal of Advanced Culture Technology, 9(3), 277 - 282. DOI: 10.17703/IJACT.2021.9.3.277

[2] Park, M. (2022). Lifelog Analysis and Future using Artificial Intelligence in Healthcare. The Journal of the Convergence on Culture Technology, 8(2), 1 - 6. DOI: 10.17703/JCCT.2022.8.2.1

[3] Kim, Jiyong, Jiyoung Lee, and Minseo Park (2022). Identification of Smartwatch-Collected Lifelog Variables Affecting Body Mass Index in Middle-Aged People Using Regression Machine Learning Algorithms and SHapley Addictive Explanations. Applied Sciences 12.8. DOI: 10.3390/app12083819

[4] Sanyal, D., and Raychaudhuri, M. (2016), Hypothyroidism and obesity, *Indian journal of endocrinology and metabolism*, **20**(4), 554. DOI: 10.4103/2230-8210.183454

[5] Zheng, Y., Manson, J. E., Yuan, C., Liang, M. H., Grodstein, F., Stampfer, M. J., Willett, W. C., and Hu, F. B. (2017), Associations of weight gain from early to middle adulthood with major health outcomes later in life. *Jama*, **318**(3), 255-269. DOI: 10.1001/jama.2017.7092

[6] Yatabe, Z. and Asubar, J. T. (2021), Ornstein-Uhlenbeck process in a human body weight fluctuation, *Elsevier*, 582. DOI: 10.1016/j.physa.2021.126286

[7] Westerterp, K. R. (2020), Seasonal variation inbody mass, body composition and activity-induced energy expenditure: along-term study, *European Journal of Clinical Nutrition*, **74**, 135 - 140. DOI: 10.1038/s41430-019-0408-y

[8] Fahey, M. C., Klesges, R. C., Kocak, M., Talcott, G. W., and Krukowski, R. A. (2020), Seasonal

fluctuations in weight and self-weighing behavior among adults in a behavioral weight loss intervention. *Eating and weight disorders*, **25**(4), 921-928. DOI: 10.1007/s40519-019-00707-7

[9] Racette, S. B., Weiss, E. P., Schechtman, K. B., Steger-May, K., Villareal, D. T., Obert, K. A., and Holloszy, J. O. (2008), Influence of weekend lifestyle patterns on body weight. *Eat Weight Disord*, **16**(8), 1826 - 1830. DOI: 10.1038/oby.2008.320

[10]Orsama, A. L., Mattila, E., Ermes, M., van Gils, M., Wansink, B., and Korhonen, I. (2014), Weight rhythms: weight increases during weekends and decreases during weekdays. *Obesity facts*, **7**(1), 36 - 47. DOI: 10.1159/000356147

[11]Somes, G. W., Kritchevsky, S. B., Shorr, R. I., Pahor, M., and Applegate, W-B. (2002), Body mass index, weight change, and death in older adults: the systolic hypertension in the elderly program, *American journal of epidemiology*, **156**(2), 132 - 138. DOI: 10.1093/aje/kwf019

[12]Lin, M. Y., Liu, M. F., Hsu, L. F., and Tsai, P. S. (2017), Effects of self-management on chronic kidney disease: A meta-analysis. *International journal of nursing studies*, **74**, 128 - 137. DOI: 10.1016/j.ijnurstu.2017.06.008

[13]Kim, Jiyong and Minseo Park(2022). A New Body Weight Lifelog Outliers Generation Method: Reflecting Characteristics of Body Weight Data. Applied Sciences 12.9. DOI: 10.3390/app12094726

[14]Arslan, M., Guzel, M., Demirci, M. C., and Ozdemir, S. (2019), SMOTE and Gaussian Noise Based Sensor Data Augmentation. *4th Int. Conf. on Computer Science and Engineering*, 1-5. DOI: 10.1109/UBMK.2019.8907003

[15]Park, J., Ahn, G., and Hur, S. (2020). Oversampling Based on k-NN and GAN for Effective Classification of Class Imbalance Dataset, *Journal of the Korean Institute of Industrial Engineers*, **46**(4), 365-371. DOI: 10.7232/JKIIE.2020.46.4.365

[16]Jiang, Z., Pan, T., Zhang, C., and Yang, J. (2021), A New Oversampling Method Based on the Classification Contribution Degree. *Symmetry*, **13**(2), 194. DOI: 10.3390/sym13020194

[17]Kumar, S., and Dhawan, S. (2020), A Detailed Study on Generative Adversarial Networks, *5th Int. Conf. on Communication and Electronics Systems*, 641-645. DOI: 10.1109/ICCES48766.2020.9137883

[18]Chalapathy, R., and Chawla, S. (2019), Deep learning for anomaly detection: A survey, *arXiv: 1901.03407*. DOI: 10.48550/arXiv.1901.03407

[19]Wang, H., Bah, M. J., and Hammad, M. (2019), Progress in outlier detection techniques: A survey, *Ieee Access*, **7**, 107964-108000. DOI: 10.1109/ACCESS.2019.2932769.

[20]Zhang, C., Zhou, Y., Chen, Y., Deng, Y., Wang, X., Dong, L., and Wei, H. (2018), Over-sampling algorithm based on vae in imbalanced classification, *Int. Conf. on Cloud Computing*, **10967**, 334-344. DOI: 10.1007/978-3-319-94295-7_23

[21]Primartha, R., and Tama, B. A. (2017), Anomaly detection using random forest: A performance revisited, *Int. Conf. on Data and Software Engineering*, 1-6. DOI: 10.1109/ICODSE.2017.8285847

[22]Chawla, N., Bowyer, K., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, **16**, 321-357. DOI: 10.1613/jair.953

[23]Han, H., Wang, W. Y., and Mao, B. H. (2005), Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *Int. Conf. on intelligent computing*, 878-887. DOI: 10.1007/11538059_91

[24]He, H., Bai, Y., Garcia, E-A., and Li, S. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *Int. joint Conf. on neural networks*, 1322-1328. DOI: 10.1109/IJCNN.2008.4633969

[25]Kingma, D. P., and Welling, M. (2013), Auto-encoding variational bayes, *arXiv:1312.6114*. DOI: 10.48550/arXiv.1312.6114

[26]Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. (2019), Don't blame the Elbo! a linear Vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, **32**, 9408-9418. DOI: 10.48550/arXiv.1911.02469

[27]Yu, R. (2020), A Tutorial on VAEs: From Bayes' Rule to Lossless Compression., *arXiv:2006.10273*. DOI: 10.48550/arXiv.2006.10273

[28]Xu, M., Quiroz, M., Kohn, R., and Sisson, S-A. (2019), Variance reduction properties of the reparameterization trick, *22nd International Conference on Artificial Intelligence and Statistics*, 2711-2720. DOI: 10.48550/arXiv.1809.10330

[29]Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), Stochastic backpropagation and variational inference in deep latent gaussian models, *Int.*

Conf. on Machine Learning, **2**(2). DOI: 10.48550/arXiv.1401.4082

[30]Ovinnikov, I. (2019), Poincar\'e Wasserstein Autoencoder, DOI: 10.48550/arXiv.1901.01427

[31]García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., and García-Rodríguez, I. (2021), Diabetes detection using deep learning techniques with oversampling and feature augmentation, *Computer Methods and Programs in Biomedicine*, **202**, 105968. DOI: 10.1016/j.cmpb.2021.105968

[32]Moreno-Barea, F. J., Jerez, J. M., and Franco, L. (2020), Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications, 161*, 113696. DOI: 10.1016/j.eswa.2020.113696

[33]García-Ordás, M. T., Benítez-Andrades, J. A., García-Rodríguez, I., Benavides, C., and Alaiz-Moretón, H. (2020), Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data, *Sensors*, **20**(4), 1214. DOI: 10.3390/s20041214

[34]Mirheidari, B., Pan, Y., Blackburn, D., O'Malley, R., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2020), Data augmentation using generative networks to identify dementia, *arXiv:2004.05989*. DOI: 10.48550/arXiv.2004.05989

[35]Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009), An experimental comparison of performance measures for classification, *Pattern Recognition Letters*, **30**(1), 27-38. DOI: 10.1016/j.patrec.2008.08.010

[36]Sun, Y., Wong, A. K., and Kamel, M. S. (2009), Classification of imbalanced data: A review, *International journal of pattern recognition and artificial intelligence*, **23**(04), 687-719. DOI: 10.1142/S0218001409007326

[37]Branco, P., Torgo, L., and Ribeiro, R. (2015), A survey of predictive modelling under imbalanced distributions, *arXiv:1505.01658*. DOI: 10.48550/arXiv.1505.01658

[38]Fan, J., Sun, C., Chen, C., Jiang, X., Liu, X., Zhao, X., Meng, L., Dai, C., and Chen, W. (2020), EEG data augmentation: towards class imbalance problem in sleep staging tasks, *Journal of Neural Engineering*, **17**(5). DOI: 10.1088/1741-2552/abb5be

[39]Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013), Facing imbalanced data-recommendations for the use of performance metrics, *Humaine Assoc. Conf. on affective computing and intelligent interaction*, 245-251. DOI: 10.1109/ACII.2013.47

[40]Chicco, D., and Jurman, G. (2020), The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC genomics*, **21**(1), 1-13. DOI: 10.1186-/s12864-019-6413-7