변수 중요도를 이용한 설명 가능한 인공지능 기법의 시각화에 대한 고찰과 보건정보 자료에의 응용

정혜린¹, 박정훈¹, 이영섭², 임창원³

¹중앙대학교 통계학과 석사과정, ²동국대학교 통계학과 교수, ³중앙대학교 응용통계학과 교수

Visualization of Explainable Artificial Intelligence Techniques Using Variable Importance with Its Applications to Health Information Data

Hyerin Jeong¹, Junghoon Park¹, Yungseop Lee², Changwon Lim³

¹Graduate Student, Department of Statistics, Chung-Ang University, Seoul; ²Professor, Department of Statistics, Dongguk University, Seoul; ³Professor, Department of Applied Statistics, Chung-Ang University, Seoul, Korea

Objectives: Deep learning techniques have been actively used in the medical field where precise diagnosis and results are very important. Deep neural network-based models utilizing big data from medical records are supporting medical opinions and are revolutionizing the medical industry. In addition, the convolutional neural network model shows excellent performance in analyzing image data and are used for image classification and X-ray/CT image reconstruction. **Methods:** In this paper, we conducted a visualization study using structured and unstructured data in the medical field. **Results:** In order to determine input variables affecting mortality and to evaluate their importance, a total of five techniques, namely, the augmented neural network model with multi-task learning, random forest, extra tree, gradient boosting and xgboost are applied to the intensive care unit data. Variable importance is calculated for each technique, and these indicators are all converted to ratios in consideration of the differences considering the patient group as a stratification variable. The converted values are shown in three graphs, a lollipop graph, a bubble chart graph, and a heat map graph. Through the visualization, it was easy to see which variables were relatively important for each technique and to what extent. InceptionResnetV2 was used as a classification model for skin cancer image data, and LIME and Grad-CAM were applied to the model to easily identify the characteristics of each cancer. **Conclusions:** Through this study, we apply several explainable artificial intelligence techniques to medical data to enhance understanding of the results of analysis and to help identify and visualize important input variables and features.

Key words: Deep learning, Explainable AI, Visualization, Variable importance

서론

최근 딥 러닝 기법을 적용한 분류 및 예측 모델은 각종 분야에서 우수한 결과를 도출하고 있고, 그에 따라 많은 각광을 받고 있다. 그러나 그 예측 결과에 대한 각 입력변수들의 기여도를 알 수 없기 때문에 블

랙박스 모델로도 불리며 모델 신뢰성의 문제를 안고 있다[1,2]. 이를 해결하고자 최근 딥 러닝 모델로 얻은 결과에 대해 시각화, 설명 및 해석을 가능하게 하는 설명 가능한 인공지능(Explainable Artificial Intelligence, XAI) 기법들이 연구되고 있다[3]. XAI 기법은 미국의 국방고등연구계획국(Defense Advanced Research Projects Agency)에서 중요한

Corresponding author: Changwon Lim

84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea
Tel: +82-2-820-5547, E-mail: clim@cau.ac.kr

Received: August 12, 2020 Revised: October 13, 2020 Accepted: October 18, 2020

*This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7083281).

*This is a summary of Hyerin Jeong's master degree thesis of Chung-Ang University.

How to cite this article:

Jeong H, Park J, Lee Y, Lim C. Visualization of explainable artificial intelligence techniques using variable importance with its applications to health information data. J Health Info Stat 2020;45(4):317-334. Doi: <https://doi.org/10.21032/jhis.2020.45.4.317>

© It is identical to the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permit unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2020 Journal of Health Informatics and Statistics

No potential conflict of interest relevant to this article was reported.

안전으로 선정될 만큼 주목받고 있다[4].

딥 러닝 기법을 적용해 데이터를 학습한 모델에 XAI 기법을 적용하고 시각화를 하면 결과에 대한 해석력을 높일 수 있다. 최근 연구된 XAI 기법들 중 Local Interpretable Model-agnostic Explanations (LIME) [5]과 Gradient-weighted Class Activation Mapping (Grad-CAM) [6]이 가장 널리 쓰이고 있다. 또한, Feng et al. [7]은 다중작업학습 증강신경망(Augmented neural network with multi-task learning, ANN-MTL)으로 변수 중요도 추출에 관한 연구를 하였다. 추출된 변수 중요도를 통해 각 입력변수들이 결과에 얼마나 영향을 미치는지를 알 수 있다.

의학 분야에서도 마찬가지로 딥 러닝 기법을 적용함으로써 환자에게 높은 정확도를 갖춘 결과를 제공할 수 있으나, 모델의 투명성 결여는 결정적인 단점이 될 수 있다. 따라서 의학 분야에 적합한 XAI 기법들이 다양하게 연구되고 있다[8,9]. Lamy et al. [8]은 실제 유방암 데이터를 기존의 사례 기반 추론(Case-Based Reasoning, CBR) 방법에 자동 알고리즘을 사용하여 시각적인 정보를 제공하거나 시각적인 추론을 위해 사용자의 인터페이스에 시각적으로 제시 될 수 있는 새로운 CBR 방법을 제안하였다. 이는 쿼리와 유사한 사례 간의 유사성을 나타내는 시각적 인터페이스를 통해 결과를 설명할 수 있다는 장점이 있다. 같은 맥락으로 Sabol et al. [9]은 조직 병리학적 이미지로부터 유방암 검출을 위해 의학적 영역에서 의미적으로 설명 가능한 분류기인 Cumulative Fuzzy Class Membership Criterion (CFCMC)를 제안하였다. 그들은 CFCMC와 일반적으로 쓰이는 3가지 분류기인 Convolutional Neural Network (CNN), Stacked Auto-Encoder, Deep Multi-Layered Perceptron을 비교하였다. CFCMC는 결과에 대한 시각적인 정보를 제공함으로써 오분류 가능성에 대한 의미적인 설명을 가능하게 함을 알 수 있었다.

본 논문에서는 인공지능 기법을 사용하여 의학 분야에서의 정형·비정형 데이터를 분석한 결과에서 변수 중요도를 기반으로 시각화하는 방안에 대해 살펴본다. 첫 번째는 사망 유무에 대한 중환자실 환자들의 상태를 측정한 정형 데이터이다. 사람들은 환자가 중환자실 내에서 사망했다는 결과보다 왜 사망을 했는지에 대한 이유에 주목하기 때문에 수치적인 근거와 함께 올바른 예측 결과 정보를 제공하고자 한다. 중환자실 데이터 중 환자 유형을 나타내는 변수를 층화 변수로 삼고, 각 환자군마다 ANN-MTL과 머신러닝 기법들을 데이터에 적용한 후 각 입력변수들의 중요도를 추출하여 기법별로 결과에 중요한 영향을 미치는 변수들이 무엇인지 비교하였다. 나아가 추출된 변수 중요도를 시각화하여 결과에 영향을 미치는 설명변수들을 종합적으로 판단할 수 있게 하였다.

두 번째는 일곱 가지 종류의 피부암을 분류하기 위한 이미지 데이터이다. 일반적으로 환자들이 피부암의 종류를 육안으로 구별하는 것은 어렵다. 이러한 이유로 이미지 분류를 위한 CNN 구조 모델에 이미지

데이터를 적용한 후, 해당 모델에 LIME과 Grad-CAM을 적용하여 많은 픽셀들 중 분류 결과에 영향을 미치는 픽셀들의 집합은 무엇인지에 대해 중요한 피쳐(feature)들의 시각화에 관해 연구하였다. 그리고 시각화된 결과를 바탕으로 피부암 종류에 따라 어떤 부분이 좋은 분류 결과를 도출하는데 중요한 영향을 미쳤는지에 대한 해석을 제공하고자 한다.

연구 방법

LIME

Ribeiro et al. [5]은 설명하고자 하는 데이터 위주로 국소적인 범위의 데이터에 근사하여 분류기 또는 회귀 예측을 설명할 수 있게 하는 알고리즘인 LIME을 제안하였다. 그들은 결과 신뢰의 문제를 각각 예측 결과의 신뢰 문제와 모형의 신뢰 문제로 정의하였다. 예측 결과의 신뢰 문제란 개별의 예측 결과를 통한 의사결정 가능 여부이고, 모형의 신뢰 문제는 학습 데이터에 과적합 되는 경우가 많으므로 모형 자체를 신뢰할 수 있는지에 대한 것이다. Figure 1으로 예를 들면, 예측 결과의 신뢰도가 높은 의사가 환자에게 어떤 요소들로 인해 감기에 걸렸는지를 신뢰도가 높은 모델을 통해 나온 결과로 설명한다면 전체적인 진단 과정 및 결과에 대한 신뢰도를 높일 수 있다는 것이다.

블랙박스 모델들을 해석하기 위해서는 실제로 모델에 입력되는 값인 피쳐는 사람이 해석할 수 있는 데이터 형태로 표현되어야 한다. 이를 해석가능한 데이터 표현(interpretable data representation)이라 한다. 예를 들어 텍스트 데이터의 경우 숫자의 의미를 알기 힘든 임베딩 벡터가 아닌 단어의 유무를 나타내는 원핫 인코딩(one-hot encoding), 이미지 데이터의 경우 픽셀의 유무를 나타내는 이진 벡터를 뜻한다.

원 데이터를 $x \in R^d$ 라고 할 때, 해석가능한 데이터 표현인 $x' \in \{0,1\}^d$ 으로 나타낸다고 하자. 해석 가능한 데이터 표현으로 원 데이터를 변환하는 작업은 디리클레 할당(Dirichlet allocation), 주성분 분석 등을 사용하여 높은 차원을 갖는 복잡한 데이터의 한 포인트에 대하여 변수 변환을 한 뒤, 변환된 변수를 사람이 얼마나 잘 원래 데이터와 같은지 판단할 수 있는가에 대한 여부를 실험을 통하여 변수 변환법을 선정하고 해석 가능하게 변형하는 방법으로 사용한다. 본 연구에서는 해석 가능하게 데이터를 표현하기 위해서 데이터의 차원이 변할 수 있으므로 차원을 d' 로 표기한다. 위에서 설명한 바와 같이 LIME은 설명하고자 하는 데이터 위주로 국소적인 범위의 데이터에 근사하기 때문에 국소의 공간에서 설명력을 높일 수 있다. 하지만 데이터 포인트의 전체적인 충실도(fidelity)를 고려하게 되면 설명력과 국소적 충실도가 작아지는 단점이 생기며 이를 충실도-해석력 트레이드오프라고 한다. 따라서 국소적 충실도와 설명력을 충족시키기 위한 설명자(explainer)인 g 는

다음과 같이 찾을 수 있다.

$$\xi(x) = \arg \min_{g \in G} \{L(g, f, \pi_x) + \Omega(g)\}$$

우선 정의된 비용함수 L 은 범위 인지 로스(Locality-aware loss)로, 다음과 같은 계산식을 갖는다.

$$L(g, f, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) (f(z) - g(z'))^2$$

또한 모델 f 는 $f: R^d \rightarrow R$ 의 함수로 데이터를 학습한 모형이자 블랙박스 모델이다. 예측 결과의 설명자인 g 는 $g \in G$ 로 정의되며 G 는 선형 회귀모델, 의사결정 나무와 같이 설명 가능한 모델들의 집합이다. π_x 는 설명하고자 하는 데이터인 x 와 다른 데이터인 z 간의 유사도를 나타내는 유사도 측도이고, $\Omega(g)$ 는 모형의 복잡도를 나타낸다. 유사도 측도는 대표적으로 벡터로 표현된 데이터들의 유클리디안 거리(Euclidean Distance), 맨해튼 거리(Manhattan Distance), 코사인 유사도(Cosine Similarity) 등으로 측정되며, 모형의 복잡도란 특정 모델이 갖는 파라미터

의 수 등으로 결정되는 모델의 크기에 대한 측도이다.

이러한 국소적 설명자를 생성하기 위해서는 설명하고자 하는 데이터를 기준으로 주변의 데이터들을 샘플링 해야 한다. 이 때 LIME은 f 를 특정한 모형으로 가정하지 않고도 국소적 설명자를 생성하여 설명을 가능하게 하는 것이 목적이다. $x' \in \{0, 1\}^d$ 에서 0이 아닌 값을 랜덤하게 뽑은 뒤 샘플을 만든다. 이렇게 생성된 샘플 $z' \in \{0, 1\}^d$ 은 해석 가능한 데이터 표현이기 때문에 예측 값을 얻기 위해 다시 모델에 입력되는 피쳐의 형태로 변환되어 $z \in R^d$ 로 정의된다. 여기에서 피쳐 z 를 통해 얻은 $f(z)$ 의 결과와 샘플 z' 로 얻은 $g(z')$ 와의 차이가 작을수록 유사도의 지표인 $\pi_x(z)$ 의 값이 커진다.

Grad-CAM

Selvaraju et al. [6]은 Class Activation Mapping (CAM)을 구할 때 예측 이미지에서 중요한 부분을 강조하는 지역 맵(local map)을 생성하기 위해 마지막 합성곱 층으로 흐르는 정답 클래스에 대한 기울기를 이용하는 Grad-CAM을 제안하였다. CAM은 이미지 분류 task에서 관심 있는 영역을 확인할 수 있도록 식별 가능 부분을 히트맵 형태로 나타내는 딥 러닝 시스템 해석 방법 중 하나로서 Zhou et al. [10]이 제안하

Figure 1. Principle of LIME (Local Interpretable Model-agnostic Explanations)(Ribeiro et al., 2016).

Figure 2. Architecture of Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017).

Table 1. Variable types and names for intensive care unit (ICU) data

Variable type	Variable name
GCS (Glasgow Coma Scale)	GCS._get_slope GCS._get_last_datapoint GCS._get_mean GCS._get_max GCS._get_min
HCO3	HCO3._get_last_datapoint HCO3._get_max HCO3._get_mean HCO3._get_min
BUN (blood urea nitrogen)	BUN._get_last_datapoint BUN._get_max BUN._get_min
WBC (white blood cell) count	WBC._get_last_datapoint WBC._get_max WBC._get_mean WBC._get_min
Temp (temperature)	Temp._get_last_datapoint Temp._get_max Temp._get_mean Temp._get_min
Glucose	Glucose._get_max Glucose._get_mean Glucose._get_min
Na	Na._get_max Na._get_mean Na._get_min
HR (heart rate)	HR._get_max HR._get_mean HR._get_min
K	K._get_max K._get_mean K._get_min
HCT (hematocrit)	HCT._get_max HCT._get_mean HCT._get_min
RespRate (respiration rate)	RespRate._get_max RespRate._get_mean RespRate._get_min
SysABP (invasive systolic arterial blood pressure)	SysABP._get_last_datapoint SysABP._get_max SysABP._get_mean SysABP._get_min
MechVent (mechanical ventilation)	MechVent._get_max
Lactate	Lactate._get_last_datapoint Lactate._get_max Lactate._get_mean Lactate._get_min
Urine	Urine._get_sum
O2	O2._get_ratio
Gender	Gender._get_identity
Age	Age._get_identity
Height	Height._get_identity
Weight	Weight._get_identity

였다. Figure 2에서는 Grad-CAM 기법의 구조를 보여주고 있다. Grad-CAM 기법은 이미지 분류, 이미지 캡셔닝, 시각 질의응답 task에 대해서 적용할 수 있으며 본 논문에서는 이미지 분류 task에 적용하였다.

아래는 분류 문제에 있어서 클래스를 구분하는데 있어서 중요한

Figure 3. Example of augmented neural network (Feng et al., 2018).**Table 2.** Selected variable types and their variance inflation factor (VIF) for intensive care unit (ICU) data

Variable type	Variable name	VIF
GCS (Glasgow Coma Scale)	GCS._get_last_datapoint	2.3311
	GCS._get_max	1.3858
HCO3	HCO3._get_last_datapoint	1.2076
BUN (Blood Urea Nitrogen)	BUN._get_last_datapoint	4.6472
	BUN._get_min	1.7191
	BUN._get_mean	4.9614
WBC (white blood cell) count	WBC._get_last_datapoint	2.5807
	WBC._get_max	4.0058
	WBC._get_min	3.9879
Temp (temperature)	Temp._get_mean	1.7821
Glucose	Glucose._get_mean	2.0490
Na	Na._get_min	4.2650
HR (heart rate)	HR._get_mean	3.0250
	HR._get_min	5.5341
K	K._get_max	2.2552
HCT (hematocrit)	HCT._get_mean	1.4327
RespRate (respiration rate)	RespRate._get_mean	5.1890
	RespRate._get_min	1.1844
SysABP (invasive systolic arterial blood pressure)	SysABP._get_last_datapoint	1.0980
MechVent (mechanical ventilation)	MechVent._get_max	4.2346
Lactate	Lactate._get_mean	1.1824
Urine	Urine._get_sum	1.2708
O2	O2._get_ratio	3.7321
Gender	Gender._get_identity	2.1208
Age	Age._get_identity	1.9622
Height	Height._get_identity	9.2872
Weight	Weight._get_identity	5.5832

부분을 표시해 주는 맵(class-discriminative localization map)인 Grad-CAM $L_{Grad-CAM}^c \in R^{u \times v}$ 을 얻기 위한 과정이다[11]. u 와 v 는 각각 특성 맵 A^k 의 너비와 높이이다. 우선 뉴런의 중요도 가중치(neuron importance weight) α_k^c 를 다음과 같이 구한다.

$$\alpha_k^c = \frac{\overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}}}{\underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}}$$

여기에서, y^c 는 소프트맥스 함수를 거치기 전 단계의 정답 클래스 c 에 대한 레이블 y 의 점수이며 역전파(backpropagation) 알고리즘을 거쳐 k 번째 특성맵(feature map)의 (i,j) 위치에 있는 $A_{i,j}^k$ 에 대한 기울기 값을 얻는다. Z 는 k 번째 특성맵의 크기이고 모든 위치에서의 기울기 값에 대해 Global Average Pooling을 함으로써 α_k^c 를 구한다.

이렇게 얻은 가중치를 통해 정답 클래스 c 에 대한 k 번째 특성 맵 A^k 의 중요도를 포착할 수 있다. k 개의 각 뉴런의 중요도 가중치 α_k^c 와 특

성 맵 A^k 을 곱해 선형 결합을 한 후 나온 맵의 픽셀 값들로 중요한 부분을 강조하는 지역 맵을 생성하기 위해 픽셀별로 ReLU함수를 씌워 양의 중요도 값을 갖는 픽셀들을 제외한 나머지 음수는 0값으로 만든다[11].

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

다중작업학습 증강신경망

Feng et al. [7]은 설명변수들의 다양한 부분집합에 대해 효율적인 계산이 가능한 손실함수와 ANN-MTL을 새롭게 제안하였다. 그들은 데이터를 적합할 때 데이터 생성 메커니즘(data-generating mechanism)의 함수로서 변수 중요도를 고려하는 것이 특정 예측 알고리즘의 사용보다 유용하다고 하였다. 이러한 원리를 기반으로 신경망을 이용한 예측 모델을 세울 때 변수 중요도 추정량을 도출할 수 있는 효율적인 방법을 제시하고 있다. 그들은 단일의 증강신경망으로도 설명변수들의 다양한 부분집합들의 중요도 추정이 동시에 가능하도록 다중작업학습

Table 3. Variable importance (%) based on different methods for intensive care unit (ICU) data of coronary artery intensive care unit

Input variable	Variable importance (%)				
	ANN-MTL	Random forest	Extra tree	Gradient boost	Xgboost
Weight_get_identity	0.001	1.984747	2.700235	0.907925	2.039848
WBC_get_min	0.001	7.102743	8.036932	10.247638	7.334815
WBC_get_max	0.206412	6.067605	6.255920	7.068826	11.962324
WBC_get_last_datapoint	3.428973	1.700929	3.704210	0.480767	0.830201
Temp_get_mean	0.512520	3.026570	3.042546	1.344442	1.674072
SysABP_get_last_datapoint	0.814416	6.109344	3.933065	3.869344	4.195320
RespRate_get_min	0.714252	4.801497	3.964689	6.152663	4.214692
RespRate_get_mean	3.941493	4.398003	3.707564	2.005488	3.334712
O2_test_ratio	2.899602	4.217877	3.510046	2.613592	2.931038
Na_get_min	6.039317	4.807719	3.948315	6.978437	3.818661
MechVent_get_max	5.528200	4.208337	3.688343	7.860636	4.614375
Lactate_get_min	4.598643	4.069973	3.089717	0.923987	2.494430
Urine_get_sum	3.515563	4.112411	3.432908	6.115355	2.390260
K_get_max	2.719401	2.811379	3.071433	1.059791	4.870373
HR_get_min	3.515563	2.365891	2.712851	0.884194	2.919153
HR_get_mean	5.225837	3.379394	3.640571	3.651095	2.776185
Height_get_identity	4.994149	2.323724	2.798093	2.035993	1.862700
HCT_get_mean	5.453311	2.643528	2.884201	2.498373	2.485545
HCO3_get_last_datapoint	4.837351	1.533862	3.102590	3.561430	2.542626
Glucose_get_mean	5.225837	1.180669	3.764013	4.642934	7.333767
Gender_get_identity	4.837351	5.454393	3.407977	5.276186	3.740742
GCS_get_max	5.453311	3.369225	3.268667	3.343644	1.467233
GCS_get_last_datapoint	5.528200	2.318257	2.729552	1.694452	4.518577
BUN_get_min	4.598643	4.626013	4.078473	4.345094	3.123457
BUN_get_mean	5.602621	4.379347	3.433827	5.523975	2.258781
BUN_get_last_datapoint	4.355722	3.146328	4.006811	3.844906	5.561736
Age_get_identity	5.453311	3.860236	4.086444	1.068835	2.704380

이 가능함을 보였다[7].

어떤 특정한 분포 P 에서 도출된 n 개의 독립적인 관측치 $O_1, O_2, \dots, O_n$ 가 있다고 가정한다. 그리고 각 관측치 O_i 는 설명 변수와 종속 변수인 $(\vec{X}_i, Y_i)$ 의 형태로 이루어져 있다. 이때 $\vec{X}_i$ 는 입력변수의 행벡터로 $\vec{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip}) \in R^p$ 이고, Y_i 는 관측치로 $Y_i \in R$ 이다. 그리고 분포 P 를 따르는 $s \subseteq \{1, \dots, p\}$ 에 속하는 입력변수 $\{X_i\}_{i \in s}$ 의 중요도를 입력변수들의 전체집합과 부분집합 s 를 제외한 축소 집합의 조건부 평균을 통해 측정하고자 한다. 입력변수들의 전체집합으로 얻은 조건부 평균은 $\mu_P(x)$ 라 하고 부분집합 s 를 제외한 축소 집합의 조건부 평균은 $\mu_{P,s}(x)$ 라고 할 때 다음과 같이 정의한다.

$$\mu_{P(x)} := E_P(Y | X = x)$$

$$\mu_{P,s}(x) := E_P(Y | X_{(-s)} = x_{(-s)})$$

두 값 간의 차이가 크면 입력변수들의 부분집합 s 의 중요도는 큰 것이고 작으면 s 의 중요도는 작다. 이를 통해 $\{X_i\}_{i \in s}$ 의 중요도를 나타내는

지표는 다음과 같다.

$$\Psi_s(P) := \frac{\int \{\mu_P(x) - \mu_{P,s}(x)\}^2 dP(x)}{\text{Var}_P(Y)}$$

이때 n 개의 관측치 $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ 가 알려지지 않은 분포 P_0 로부터 관측되었다고 가정하고, μ_{P_0} 와 $\mu_{P_0,s}$ 의 추정량인 $\hat{\mu}$ 와 $\hat{\mu}_s$ 로 이루어진 $\hat{P}_n$ 가 P_0 의 일치추정량이라고 한다면, 중요도의 추정량을 다음과 같이 표현할 수 있다.

$$\Psi_s(\hat{P}_n) := \frac{\frac{1}{n} \sum_{i=1}^n (\hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)}))^2}{\frac{1}{n} \sum_{i=1}^n (Y^{(i)} - \frac{1}{n} \sum_{i=1}^n Y^{(i)})^2}$$

위 추정량을 구하는 식은 부분집합 s 의 중요도 $\psi_{0,s}$ 를 추정하기 위해 P_0 와의 관련 구성요소를 추정한 $\hat{\mu}$ 와 $\hat{\mu}_s$ 를 사용하여 Ψ_s 로 매핑하는 과정을 거치며 중요도의 추정량을 계산할 수 있다. 즉, $\psi_{0,s} = \Psi_s(P_0)$ 이고,

Table 4. Variable importance (%) based on different methods for intensive care unit (ICU) data of Cardiac intensive care unit

Input variable	Variable importance (%)				
	ANN-MTL	Random forest	Extra tree	Gradient boost	Xgboost
Weight_get_identity	0.0001	5.959468	4.524703	1.121654	3.506411
WBC_get_min	2.138305	7.450499	8.414269	17.780270	11.932611
WBC_get_max	5.270900	5.833208	6.704595	1.659070	9.340221
WBC_get_last_datapoint	0.00001	0.640560	1.711124	0.0001	4.593001
Temp_get_mean	2.299458	3.605929	4.050904	1.899135	3.124144
SysABP_get_last_datapoint	0.00001	4.122259	3.566776	3.236429	3.765896
RespRate_get_min	1.976677	3.419955	3.285464	6.427495	3.359787
RespRate_get_mean	4.659616	3.250535	2.637879	5.491383	1.412965
O2_test_ratio	2.460134	3.383015	2.862100	4.370873	2.880383
Na_get_min	3.414195	2.421272	2.831085	0.615576	1.594466
MechVent_get_max	3.884799	3.146988	2.814208	3.677039	1.501748
Lactate_get_min	2.620335	6.406952	4.392777	10.470476	6.654652
Urine_get_sum	2.939307	2.551934	2.942353	0.601345	1.146474
K_get_max	4.196154	5.399280	4.692628	3.435476	3.977400
HR_get_min	6.024294	3.699771	3.655368	2.005268	3.104613
HR_get_mean	3.884799	2.661642	3.085963	10.257143	3.481806
Height_get_identity	3.728407	3.669415	3.638732	4.096971	3.987188
HCT_get_mean	3.098079	4.416741	3.321754	0.174189	1.910782
HCO3_get_last_datapoint	9.474909	1.580238	2.754314	1.260308	3.670819
Glucose_get_mean	4.505605	0.418093	2.355160	0.0001	0.0001
Gender_get_identity	6.322319	4.200228	3.888120	6.354813	3.345132
GCS_get_max	6.322319	3.005077	3.129361	0.0001	2.814240
GCS_get_last_datapoint	5.118793	2.739699	3.634824	0.0001	2.155007
BUN_get_min	5.270900	5.887796	3.380698	9.600376	5.989433
BUN_get_mean	0.0001	2.727353	3.516871	5.148142	3.884172
BUN_get_last_datapoint	5.118793	2.804205	3.465983	0.156768	4.083366
Age_get_identity	5.270900	4.597888	4.741987	0.159803	2.783292

몇 가지 계산을 거치면 추정량 $\hat{\psi}_{n,s}$ 을 아래의 식으로 구한 값으로 사용하게 된다[7]:

$$\hat{\psi}_{n,s} = \frac{\sum_{i=1}^n A_{i,n,s} \{\hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)})\}}{\sum_{i=1}^n \left(Y^{(i)} - \frac{1}{n} \sum_{i=1}^n Y^{(i)} \right)^2}$$

이 때 $A_{i,n,s} := 2Y_i - \hat{\mu}(X^{(i)}) - \hat{\mu}_s(X^{(i)})$ 이다.

ANN의 모든 모수들의 집합이 $\Theta^{(q)}$ 이고 입력층의 노드가 q 개, 출력층의 노드가 1개라 하자. 이 때 모수 $\theta \in \Theta^{(q)}$ 는 $f(\cdot; \theta)$ 로 표현한다. 훈련 데이터 셋은 n 개의 관측치이며 $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ 와 같이 순서쌍의 형태를 갖는다. 증강신경망의 구조는 Figure 3과 같다.

ANN 구조에서 입력 층은 각각 입력변수 task인 x_i 와 조건부 평균 task인 m_i 로 이루어져 있다. 모델의 핵심은 이러한 입력 층의 구조로 인해 위의 두 가지 task로 동시에 학습하는 다중작업학습을 통하여 조건부 기댓값을 추정함으로써 변수 중요도 값을 구할 수 있다는 것이

다. 이진벡터 $e_s \in \{0,1\}^p$ 로 입력 노드의 개수를 입력변수 개수의 2배만큼 늘린다. p 개의 입력변수 $X_1, \dots, X_p$ 가 모두 주어졌을 때, 조건부 평균 $E(Y|X_1, \dots, X_p)$ 를 얻으려면 $m_1, \dots, m_p$ 에 모두 0을 넣는다. 그리고 관심 있는 입력변수를 $X_j, 1 \leq j \leq p$ 로 하고 X_j 를 제외한 나머지 입력변수들이 주어졌을 때, 조건부 평균 $E(Y|X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p)$ 을 얻으려면 m_j 에 1을 넣고 나머지 m_i 들에는 0을 넣는다. 이렇게 두 가지 task로 인해 발생하는 모델의 multi-loss는 다음과 같이 정의하고 이를 최소화하는 θ 를 찾는다.

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta^{(q)}} \frac{1}{n} \sum_{i=1}^n \left[\{y^{(i)} - f(x^{(i)}, 0; \theta)\}^2 + \sum_{s \in S} E_{W_s} \left[\{y^{(i)} - f(\xi(x^{(i)}, W_s; s), e_s; \theta)\}^2 \right] \right] \quad (2.1)$$

여기에서 W_s 는 실수로 이루어진 공간인 $R^{|s|}$ 에 속한 확률변수이다. Feng et al. [7]은 실험을 통해 확률적 경사하강법(stochastic gradient descent)의 매 단계에서 적합한 W_s 를 찾아 대체 하는 것이 loss를 더욱 줄일 수 있는 면에서 나음을 보였다. 그러나 본 연구에서는 머신러닝 기법과도 비교하기 때문에 고정된 값으로 W_s 를 대체하였다. $\{\xi(x, w; s)\}_{(-s)} = x_{(-s)}^{(i)}$ 이고

Table 5. Variable importance (%) based on different methods for intensive care unit (ICU) data of internal medicine intensive care unit

Input variable	Variable importance (%)				
	ANN-MTL	Random forest	Extra tree	Gradient boost	Xgboost
Weight_get_identity	1.313213	5.111148	3.657771	3.606611	3.550381
WBC_get_min	0.0001	5.482862	6.264846	8.487957	13.527627
WBC_get_max	2.728389	2.286591	3.206881	1.632727	2.913420
WBC_get_last_datapoint	2.937520	1.505589	3.422733	0.682628	2.097240
Temp_get_mean	4.980232	4.176438	4.059628	3.838539	2.994533
SysABP_get_last_datapoint	8.071779	6.905809	4.356698	9.360910	4.460533
RespRate_get_min	2.539217	3.468726	3.982990	4.158083	3.743716
RespRate_get_mean	1.700814	4.756374	3.993136	4.533939	2.864446
O2_test_ratio	3.321890	3.625559	3.755108	3.257326	2.253265
Na_get_min	4.411125	4.346511	3.643005	3.596672	3.227546
MechVent_get_max	8.421295	3.760300	3.651645	4.281329	4.385176
Lactate_get_min	4.211780	4.035580	3.320316	4.484678	3.106723
Urine_get_sum	0.282746	3.527303	3.359838	2.800620	2.546731
K_get_max	1.495444	3.148612	3.424907	2.509678	3.565694
HR_get_min	2.775200	3.287242	3.669624	3.908434	3.468104
HR_get_mean	6.550448	3.191219	3.318066	2.109585	2.178870
Height_get_identity	2.914476	3.671482	3.513619	3.957206	1.954604
HCT_get_mean	4.701102	3.211956	3.520727	2.309994	3.186791
HCO3_get_last_datapoint	3.187821	2.667612	3.866917	1.455748	3.668066
Glucose_get_mean	1.624162	0.475035	2.079197	0.392247	2.876285
Gender_get_identity	5.421344	5.311181	4.546159	6.061881	4.686853
GCS_get_max	4.272090	4.227520	3.269199	4.616641	3.048196
GCS_get_last_datapoint	4.489512	1.190535	2.045394	1.233720	3.179331
BUN_get_min	5.106781	5.862623	4.532318	5.824200	4.761057
BUN_get_mean	3.734993	3.655422	3.842717	4.054883	4.850372
BUN_get_last_datapoint	4.048834	3.410322	3.679027	2.625454	3.133811
Age_get_identity	4.757796	3.700448	4.017533	4.218310	3.770621

$\{\xi(x, w; s)\}_s = W_s$ 로 정의하여 함수 $\{\xi(x, w; s)\}_s$ 는 부분집합 s 에 속한 변수들을 W_s 로 대체한다. 이를 통해 (x, w) 를 R^p 로 매핑하여 기존의 p 차원을 유지한다. 식(2.1)은 다음과 같이 다시 표현한다.

$$\hat{\theta} = \underset{\theta \in \Theta^{(p)}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left(\left\{ y^{(i)} - f(x^{(i)}, m=0; \theta) \right\}^2 + \sum_{s \in S} \left[y^{(i)} - f(\xi(x^{(i)}, w_s; s), m=e_s; \theta) \right]^2 \right)$$

따라서 모든 경우의 조건부 기댓값을 추정하는데 있어 손실함수 값의 합을 최소화하도록 이 증강신경망으로 학습한다. 그리고 해당 모델을 이용해 조건부 기댓값을 구함으로써 변수 중요도 값을 계산할 수 있다.

연구 결과

중환자실 데이터에 대한 변수 중요도의 시각화

데이터 설명 및 실험 목적

본 연구의 중환자실 수치형 데이터는 PhysioNet/CinC Challenge 2012로부터 공개되었다[12]. 중환자실 환자들의 사망률 예측 방법의

개선을 위해 입원했던 4,000명의 환자들을 대상으로 5개의 환자 고유 변수를 제외한 최대 36개의 변수에 대해 48시간 동안 시간마다 측정하였다. 이를 통해 각 변수의 수치 변화에 따라 위험하거나 안정적인 환자를 조기에 식별하는데 도움이 될 수 있다. 환자 고유변수는 환자 번호, 나이, 성별, 키, 몸무게이다. 그 외 36개의 변수 중 중요도 측정에 사용된 변수는 Table 1에 제시하였다.

또한 응급실을 내원한 환자들의 증상에 따라 분류한 값(Intensive Care Unit, ICU 타입)을 나타내는 변수를 층화 변수로 사용하였다. 환자군마다 응급실 내 사망에 대해 작용하는 변수들이 차이가 있고 그 결과로 머신 러닝 기법들의 변수 중요도로 나타날 것이라고 예상하였기 때문이다. ICU 타입은 총 4가지로, 1은 관상 동맥계 중환자실, 2는 심장계 중환자실, 3은 내과계 중환자실, 4는 외과계 중환자실이다. 사용된 변수 중 하나인 인공호흡기 Mechventilation (MechVent)는 차고 있지 않은 상태인 0이 정상임을 뜻한다. 48시간 동안 2번 이상의 측정 값을 가진 변수에 대해서는 이를 그대로 사용하지 않고 각 변수의 최솟값, 최댓값, 평균값, 합계, 변화량, 마지막 수치의 추가 변수를 생성하

Table 6. Variable importance (%) based on different methods for intensive care unit (ICU) data of surgical intensive care unit

Input variable	Variable importance (%)				
	ANN-MTL	Random forest	Extra tree	Gradient boost	Xgboost
Weight_get_identity	0.0001	3.650538	3.053597	4.618936	2.862142
WBC_get_min	0.0001	7.209325	8.104899	16.825437	14.341876
WBC_get_max	0.245416	6.837394	7.421712	2.103836	6.977782
WBC_get_last_datapoint	0.822838	2.093095	3.222705	1.584576	1.714369
Temp_get_mean	3.177637	4.115112	4.006885	5.472223	4.252891
SysABP_get_last_datapoint	8.919172	7.859078	4.778725	10.338187	4.691458
RespRate_get_min	1.748892	5.950134	5.678819	7.987829	6.282029
RespRate_get_mean	3.464816	3.569654	3.139229	2.583669	1.859348
O2_test_ratio	4.817297	3.757061	3.580982	2.046982	3.064959
Na_get_min	1.658329	3.208044	3.530784	4.102899	3.854127
MechVent_get_max	1.198707	4.002945	3.450060	3.780867	3.584676
Lactate_get_min	2.884897	3.404088	2.942227	2.502872	2.618749
Urine_get_sum	3.746434	3.814217	3.289525	4.465988	3.741363
K_get_max	4.139099	2.783826	3.071632	2.170050	3.934206
HR_get_min	5.799187	3.564722	3.467324	3.698861	3.146863
HR_get_mean	4.177741	3.406604	3.189839	2.311588	2.036769
Height_get_identity	4.743758	3.904514	3.534513	2.920932	2.487597
HCT_get_mean	5.178186	2.483274	2.644311	0.890918	2.412638
HCO3_get_last_datapoint	4.482794	3.142019	4.025327	3.109981	3.346658
Glucose_get_mean	5.389272	0.411850	1.935010	0.211530	1.396616
Gender_get_identity	3.177637	2.984321	3.098911	3.638714	3.334089
GCS_get_max	4.061474	3.125933	3.072938	2.426047	2.459663
GCS_get_last_datapoint	5.630375	1.930603	2.368201	0.542395	2.423868
BUN_get_min	6.477215	2.956192	3.194304	2.559336	1.842412
BUN_get_mean	5.389272	3.178690	3.848187	3.025548	4.875683
BUN_get_last_datapoint	4.963015	2.633018	3.014962	2.138782	3.182973
Age_get_identity	3.706543	4.023748	3.334392	1.941016	3.274183

였으며, 본 연구에 쓰인 변수는 총 55개이다. 더 나아가, 각 변수의 최솟값, 최댓값 등을 모두 사용하여 머신 러닝 기법에 적용한다면 변수 간 상관관계가 있는 경우 변수 중요도를 측정하는 데 있어서 잘못된 해석을 도출할 가능성이 매우 크다. 따라서 각 변수의 대푯값을 설정하기 위해, 최솟값, 최댓값, 평균값, 합계, 변화량, 마지막 수치를 모두 포함한 55개의 변수들에 대한 분산팽창지수를 계산하여 다중공선성이 일어나지 않도록 변수 선택법을 적용하였다. 선택된 변수는 총 27개였으며 각 변수 명과 계산된 분산팽창지수를 Table 2에 제시하였다. 타겟 변수는 중환자실 내에서의 사망 여부이고, 4,000개의 데이터 셋을 scikit-learn api [13]의 함수 중 `train_test_split`을 사용하여 랜덤하게 훈련 데이터 셋 3,200개, 테스트 데이터 셋 800개로 나누어 ANN-MTL로 학습하였다.

본 논문에서는 ICU 데이터를 중환자실 타입으로 분류한 뒤 머신러닝 기법인 랜덤 포레스트, 엑스트라 트리, 그라디언트 부스팅, Xgboost로도 학습하여 각 기법별로 응급실 내원 환자군에 따라 사망 여부에 영향을 미치는 변수들이 무엇인지, 그 변수들의 중요도는 어느 정도인

지 비교해보았다. 나이가 변수 중요도의 값들을 시각화를 하였으며 그 결과들을 종합적으로 판단하여 변수 중요도에 대한 결론을 내리고 결과에 대한 해석을 하였다.

실험 결과

분산팽창지수를 고려하여 선택된 27개의 중환자실 데이터 설명변수를 신경망 기반의 ANN-MTL과 랜덤 포레스트, 엑스트라 트리, 그라디언트 부스팅, Xgboost에 적용한 후 각 기법별 변수 중요도의 산출 기준이 다를 수 고려하여 비율로 변환하였다. 총 4개의 중환자실 환자군에 각각 기법들을 적용하였으며 이는 Tables 3에서 Table 6과 같다.

첫 번째 중환자실 타입인 관상 동맥계 중환자실의 경우, 랜덤포레스트, 엑스트라 트리, 그라디언트 부스팅의 기법들에서 백혈구의 수를 나타내는 단위의 변수인 WBC의 최솟값이 가장 중요한 변수로 공통적으로 나타났다. Xgboost의 경우도 또한 두 번째로 높은 변수로 나타나 관상 동맥계 중환자실의 경우 사망 여부에 백혈구의 최소 숫자가 가장 중요하게 작용한다는 것을 확인할 수 있었다. 또한 ANN-MTL의 경우

Figure 4. Lollipop plot of variable importance ratio for coronary artery intensive care unit.

나트륨의 최소값 변수인 Na_get_min 변수가 약 6.0%의 변수 중요도 비율로 가장 높게 나왔다.

두 번째 중환자실 타입인 심장계 중환자실의 경우, 관상 동맥계와 비슷한 결과를 보였다. 랜덤 포레스트, 엑스트라 트리, 그라디언트 부스팅, Xgboost 모두 WBC의 최소값 변수가 가장 중요한 변수로 작용했음을 알 수 있고, 그라디언트 부스팅의 경우 17.5% 이상의 중요도를 보이며 가장 높은 변수 중요도를 기록했다. ANN-MTL의 경우엔 중탄산 이온의 마지막 수치가 중환자의 사망 예측에 있어 가장 중요한 변수로 나타났다. 약 9.4%의 변수 중요도를 보였으며 두 번째로 높은 변수는 환자 의식 수준의 마지막 측정값인 GCS_get_last_datapoint로 약 6.3%로 나타났다.

세 번째 중환자실 타입인 내과계 중환자실은 엑스트라 트리와 Xgboost의 경우 WBC의 최소값이 가장 중요하게 작용하였고, 각각 6.2%와 13.2%로 나타났다. 랜덤 포레스트와 그라디언트 부스팅은 동맥 혈압인 SysABP의 마지막 기록 변수가 각각 약 6.9%와 9.3%, ANN-MTL의 경우는 인공 호흡기의 여부인 MechVnt_get_max의 값이 약 8%로

중환자의 사망 예측에 가장 중요하게 작용한 변수로 계산되었다.

네 번째 중환자실인 외과계 중환자실의 경우도 엑스트라 트리, 그라디언트 부스팅, Xgboost 3가지 모델에서 WBC의 최소값이 각각 8.1%, 16.8%, 14.3%로 가장 큰 변수 중요도 값을 나타내었다. 또한 WBC 가장 큰 값인 WBC_get_max 변수도 랜덤 포레스트, 엑스트라 트리, Xgboost 세 개의 모델에서 두 번째로 큰 변수 중요도를 보여 외과계 중환자의 경우 백혈구의 수가 사망 예측에 가장 큰 비중을 차지하는 것을 확인할 수 있었다. ANN-MTL의 경우엔 동맥 혈압의 마지막 측정치인 SysABP_get_last_datapoint가 약 8.9%로 중환자의 사망을 예측하는 데 가장 큰 변수 중요도 값으로 나타났다.

4가지 중환자실 유형에서 머신 러닝 기법들에 공통적으로 가장 큰 변수 중요도를 갖는 변수로 자주 선별된 것은 백혈구 수치이다. 백혈구의 수치는 광범위한 급성 또는 만성 질환의 지표로 널리 사용되는 것으로 알려져 있기 때문에[14], 혈관과 관련이 있는 관상 동맥계와 심장계 중환자실의 경우와 내과계, 외과계 중환자실 모두 사망 예측에 중요한 변수로 나타난 것이 타당하다고 보인다.

Figure 5. Lollipop plot of variable importance ratio for cardiac intensive care unit.

또한 총 4가지의 중환자실 타입별 비율로 변환한 변수 중요도를 통계 프로그램 R의 ggplot2 패키지[15]를 사용하여 다양한 형태로 시각화 하였다. 첫 번째와 두 번째 응급실 유형인 관상 동맥계와 심장계 중환자실은 Figure 4와 Figure 5로 롤리팝 플랏으로 나타내었다. 각 Figure의 x축은 변수 중요도의 비율이며 y축은 설명변수들이고 색깔은 변수 중요도를 비교한 각 기법이다. 또한 세 번째 유형인 내과계 응급실은 버블 차트를 사용하여 기법별로 변수 중요도에 따른 크기로 나타내었으며 Figure 6에서 확인할 수 있다. Figure 7은 네 번째 유형인 외과계 응급실의 히트맵으로, 큰 변수 중요도를 가질 수록 어두운 색으로 나타난다. 시각화한 그림들을 통해 각 기법마다 중요한 변수가 무엇인지 한 눈에 파악할 수 있었기 때문에 변수 중요도를 추출하여 이를 바탕으로 시각화한다는 것은 매우 중요하고 유용하다.

피부암 이미지 데이터를 활용한 피쳐 중요도의 시각화

데이터 설명 및 실험 목적

두 번째로 사용한 데이터인 피부암 이미지 데이터는 Tschandl et al.

[16]의 Human Against Machine with 10000 training images (HAM10000) 데이터 셋이다. HAM10000 데이터 셋은 총 10,015장이며 이미지 한 장 당 크기는 크기는 $600 \times 450 \times 30$ 이다. 피부암의 종류는 광선각화증(akiec), 기저세포암(basal cell carcinoma, bcc), 지루각화증(bkl), 피부섬유종(dermatofibroma, df), 악성흑색종(melanoma, mel), 비정형색소모반(Melanocytic nevi, nv), 혈관종(vascular lesion, vasc)으로 총 7가지이다. 각 병변별로 특징이 있지만 일반 환자들이 구별하기엔 무리가 있으며 초기에는 점과 비슷하게 보이기도 한다. 그러나 병변이 진행됨에 따라 위험도 높은 범주로 분류될 가능성이 있다. 또한 nv와 mel은 매우 비슷한 모양을 보이기도 한다. 그러므로 병변을 잘 분류하고 LIME과 Grad-CAM으로 시각화함으로써 피쳐 중요도를 나타내어 분류한 결과에 대한 설명을 용이하게 하고자 본 실험을 수행하였다.

실험 결과

다음은 피부암 이미지 데이터에 관한 실험 결과이다. 7가지 병변별로 LIME과 Grad-CAM 기법 적용 후의 이미지와 함께 생김새의 특징

Figure 6. Bubble chart of variable importance ratio for internal endoplasmic reticulum intensive care unit.

을 설명한다.

Figure 8은 LIME 기법을 적용하여 bcc에 속한 이미지에서 강조하고 있는 부분을 나타낸 것이다. 첫 번째 이미지는 분류하려는 이미지에서 가장 높은 확률을 갖고 도출해낸 레이블에 기여한 부분만 보였으며 두 번째 이미지는 해당 이미지 전체에 경계만 추가한 것이다. 세 번째

이미지에서 초록색은 1위의 레이블을 도출하는데 확률을 높이는 부분이며 빨간색이 확률을 낮추는 부분이다. Figure 9는 bcc에 속한 이미지에 Grad-CAM 기법을 적용하여 이를 이용해 원본 이미지에 히트맵(heatmap)을 더한 것이다. 가장 높은 확률을 갖는 레이블에서 강조하고 있는 부분이 노란색으로 나타났으며 원본 이미지에서는 그 부분이

Figure 7. Heatmap of variable importance ratio for surgical intensive care unit.

Figure 8. LIME (Local Interpretable Model-agnostic Explanations) result of bcc for skin cancer data.

빨간색으로 나타났다. 따라서 테스트 셋 이미지 데이터들 중 정답을 맞춘 이미지들의 다섯 가지 결과를 통해 각 병변의 고유한 특징 시각화를 하였다. 뿐만 아니라 1순위로 0.5에 이르는 확률의 오답 레이블, 2순위로 정답 레이블을 출력한 이미지는 출력된 두 순위의 특징이 비슷하기 때문이라 판단하여 이에 대해서도 LIME과 Grad-CAM을 적용했다.

akiec는 각화증인 만큼 표면이 건조하고 거칠며 적갈색의 원형모양이 1개에서 수십 개씩 생기고 기름기 있는 비늘과 딱지의 모양을 띤다. 초기에는 습진과 비슷하여 습진치료를 주로 받지만 잘 낫지 않기 때문에 초기에 조직검사를 통한 정확한 진단이 필요하다. Figure 10은 각각 거의 1에 가까운 0.99의 확률 값을 갖고 확실하게 분류하였으며 모두 akiec의 특징을 강조하고 있다. Figure 11은 0.57의 가장 높은 확률을 갖고 1순위로 mel이라 분류하였고 0.3의 확률을 갖고 2순위로 bkl이라 분류한 예시이다. 1순위 오답으로 mel이 가장 많았으며 mel로 분류하지만 이에 대한 1순위 확률 값이 0.5를 조금 넘는 정도이고 비슷한 확률 값으로는 2순위로 정답인 akiec를 도출했음을 통해 변색 된 부분에 대해 명확한 구분이 어려워 mel로 가장 많이 오분류하고 있음을 알 수 있었다.

bcc는 출혈상처와 비슷하게 생겼으며 색깔은 주로 붉은 색이고 검은

점을 특징으로 한다. 각각 1.0과 0.77의 확률 값으로 잘 분류하였으며 Figure 12를 통해 이러한 특징을 강조하고 있음을 알 수 있다. 그러나 오답은 mel, bkl, nv 순으로 가장 많았다. Figure 13을 통해 실례를 볼 수 있다. 각각의 상단에서는 0.95의 확률 값으로 mel이라 오분류하였다. 1순위를 0.53의 확률 값으로 mel, 2순위를 0.46의 확률 값으로 nv라 오분류 하였다. 이 경우에는 mel과 nv의 유사성에 의해 각각 0.5에 가까운 확률 값으로 정확한 분류를 하고 있지 못함을 알 수 있었다.

bkl은 검버섯과 비슷하게 생겨 저승꽃이라 불리기도 한다. 또한 작은 점들이 여러 개 있음을 알 수 있는데, Figure 14는 이런 특징들을 잘 강조하고 있음을 알 수 있다. 0.99의 확률 값으로 잘 분류하였다. 그러나 akiec와 마찬가지로 병변의 진행에 따라 검정색에 가까운 색으로 짙어지는데, 이런 이미지들은 mel과 nv 순으로 오분류하였으며 이는 Figure 15로부터 알 수 있다. 각각의 하단에서는 넓은 범위의 검정색에 가까운 진한 갈색의 모양을 띄고 있기 때문에 0.78의 확률 값으로 mel로 오분류 하고 있음을 알 수 있었다.

df는 멍이 든 것 같이 대체적으로 피부색과 비슷한 색을 띄거나 분홍, 회색, 적갈색으로 나타내지만 병변이 진행되면서 변색되기도 한다. 대체적으로 정답을 맞췄으며 Figure 16을 통해 이런 특징을 강조하고

Figure 9. Grad-CAM (Gradient-weighted Class Activation Mapping) result of bcc for skin cancer data.

Figure 11. Grad-CAM (Gradient-weighted Class Activation Mapping) result of akiec images which are incorrectly classified as mel for skin cancer data.

Figure 10. LIME (Local Interpretable Model-agnostic Explanations) result of akiec for skin cancer data.

Figure 12. LIME (Local Interpretable Model-agnostic Explanations) result of bcc for skin cancer data.

Figure 13. LIME (Local Interpretable Model-agnostic Explanations) result of bcc images which are incorrectly classified as mel for skin cancer data.

Figure 14. Grad-CAM (Gradient-weighted Class Activation Mapping) result of bkl for skin cancer data.

있음을 알 수 있었다. 상단과 하단에서는 각각 1.0, 0.7의 확률 값을 갖고 잘 분류함을 보였다. 그러나 Figure 17과 같이 검은 점이 있는 이미지에 대해서는 0.78의 확률로 흑색종(mel)이라 오분류했음을 알 수 있었다. 이 확률 값을 위에서 정답을 맞춘 그림이 도출한 확률인 0.70보다 더 큰 값이므로 이러한 특징을 보이는 df에 대해서 mel로 오분류할 가능성이 높음을 알 수 있었다.

mel은 전이의 확률 및 속도가 매우 높고 빠르기 때문에 피부암 중에

서도 고위험군으로 분류되고 있다. 다른 병변에 비해 6 mm 이상의 비교적 큰 크기로 발병, 울퉁불퉁한 경계, 비대칭을 특징으로 하고 있으며 Figure 18은 1.0의 확률 값으로 잘 분류하였고 앞서 기술한 mel의 특징들을 강조하고 있음을 알 수 있었다. 오분류된 이미지들을 분석해보니 가장 높은 확률을 갖고 1순위로 추출된 레이블들은 nv와 bkl 순이었다. 그러나 아래의 Figure 19의 경우 0.60 정도의 가장 높은 확률 값으로 nv라 도출하였으며 2순위는 0.40 정도의 확률 값을 갖고 정답인 mel로 분류하였다. 각 값이 0.5보다 조금 작거나 크기 때문에 mel을 nv로 오분류할 가능성이 높지만 mel과 nv가 굉장히 비슷한 모습을 보이고 있음을 알 수 있었다.

nv는 실제 mel 진행될 확률이 다른 병변에 비해 높고 모양도 mel과 비슷한 것이 특징이다. Figure 20은 1.0의 확률 값으로 정답을 맞추었고 이러한 특징을 강조하고 있음을 알 수 있었다. 오분류된 레이블들로는 bkl과 mel이 가장 많았다. 그리고 이런 이미지들을 분석해보니 Figure 21을 예로 들면, bkl이 1순위로 도출되었음에 불구하고 확률은 0.40이었다. 이어서 2순위인 mel의 확률은 0.34, 3순위로 정답인 nv의 확률은 0.25로 나타났다. 이러한 이유 때문에 어떠한 특징도 잡지 못하며 분류를 제대로 하고 있지 못함을 알 수 있다.

Figure 15. LIME (Local Interpretable Model-agnostic Explanations) result of bkl images which are incorrectly classified as mel for skin cancer data.

Figure 16. LIME (Local Interpretable Model-agnostic Explanations) result of df for skin cancer data.

Figure 17. LIME (Local Interpretable Model-agnostic Explanations) result of df image which is incorrectly classified as mel for skin cancer data.

마지막으로 vasc는 주로 혈관 내에서 발생하기 때문에 대체적으로 붉은 덩어리의 모양을 띄는데 Figure 22는 1.0의 확률 값으로 정답을 잘 맞췄고 해당 병변의 특징을 잘 강조하고 있음을 알 수 있었다. 대체적으로 분류를 잘 하였지만 Figure 23과 같이 검은 점도 있는 이미지는 0.78의 확률 값을 갖고 bcc, 넓은 부위에 검정색을 보이는 이미지는 0.92의 확률 값을 갖고 mel로 오분류했음을 알 수 있었다.

고찰 및 결론

본 논문에서는 의학 분야에서의 정형·비정형 데이터를 이용한 설명 가능한 인공지능 기법에 관한 연구를 수행하였다. 먼저 정형 데이터는 중환자실에 입원한 환자들을 대상으로 48시간 동안 측정된 기록 데이터이다. 종속변수는 중환자실 내 사망 여부이며 어떤 변수들이 중환자실에서 사망하는데 얼마나 영향을 미치는지에 대한 변수 중요도를 추

Figure 18. Grad-CAM (Gradient-weighted Class Activation Mapping) result of mel for skin cancer data.

Figure 20. Grad-CAM (Gradient-weighted Class Activation Mapping) result of nv for skin cancer data.

Figure 19. LIME (Local Interpretable Model-agnostic Explanations) result of mel images which are incorrectly classified as nv for skin cancer data.

Figure 21. LIME (Local Interpretable Model-agnostic Explanations) result of nv images which are incorrectly classified for skin cancer data.

출하였다. 사용된 기법들은 다중 작업 학습을 통한 증강 신경망, 랜덤 포레스트, extra tree, 그래디언트 부스팅, xgboost이다. 각 기법별로 변수 중요도 지표가 다를 것을 고려하여 모두 퍼센트 단위의 비율로 변환하였다. 변환한 데이터를 세 가지 방법으로 시각화하여 중요한 변수들을 한 눈에 파악할 수 있었다. 여러 가지 기법을 적용했기 때문에 오히려 시각화를 함으로써 각 기법별로 타겟변수에 영향을 미치는 입력변수들의 중요도를 기반으로 결과에 대한 해석을 더 잘 할 수 있었다.

이어서 비정형 데이터는 7가지 범주의 피부암 이미지 데이터이다. 설명 가능한 인공지능 기법 중 LIME과 Grad-CAM을 적용하여 가장 높은 확률의 레이블을 도출하는데 있어 기여한 부분을 강조하여 시각화하였다. 이를 통해 육안으로는 구별이 쉽지 않은 피부암의 특징들을 범주에 따라 한 눈에 파악할 수 있었다. 뿐만 아니라 bkl, mel, nv는 실제로도 매우 유사하게 생겼으며, 특히 nv는 병변이 진행됨에 따라 mel로 판정받을 가능성이 높다. 또한 모든 피부암은 병변이 진행되면서 질

Figure 22. Grad-CAM (Gradient-weighted Class Activation Mapping) result of *vasc* for skin cancer data.

은 갈색이나 검정색을 띄기 때문에, 실험 결과 위의 세 범주가 가장 높은 빈도의 응답으로 도출되었다. 이러한 정보들을 LIME과 Grad-CAM을 적용하여 더욱 잘 파악할 수 있었다. 그렇기 때문에 초기에는 일반적인 점과 비슷한 모양으로 발병하기 때문에 피부암이 의심될 경우 내원 후 조기에 병변과 알맞은 치료법으로 대처해야 할 필요가 있다.

환자 및 질병 데이터에 관한 분류 및 예측 결과는 생명이 직결되어 있는 만큼 높은 정확도도 중요하지만 왜 그러한 결과를 얻게 되었는지에 대한 설명을 통해 환자에게 신뢰를 안겨줄 필요가 있다. 이에 따라 본 논문에서는 의학 데이터의 변수 중요도를 이용한 분류 및 예측의 시각화 기법에 대한 연구를 수행하였다. 인공지능 기법 기반의 XAI 기법인 다중작업학습 증강신경망과 여러 가지 머신러닝 기법들에 수치형 데이터를 적용하여 각 기법별 변수 중요도를 추출할 수 있었다. 추출된 변수 중요도를 시각화하여 변수 관점에서 사망 결과에 대한 원인을 알 수 있었다. 그리고 LIME, Grad-CAM에 이미지 데이터를 적용하여 분류된 범주의 특징을 강조하는 부분의 피처의 중요도를 나타낼 수 있었다. 이를 통해 분류 및 예측에 있어 왜 그러한 결과가 나오게 되었는지 설명을 가능하게 해주고 해석을 용이하게 할 수 있다는 이점이 있었다. 따라서 본 연구를 통해 환자들이 합리적인 의사결정에 도모할 수 있을 것이라 기대하는 바이다.

또한 본 연구에서는 모델 적합에 사용된 변수 각각의 중요도에 대해서만 실험하였다. 기존의 머신러닝 기법들뿐만 아니라 다중작업학습 증강신경망 또한 변수들의 부분집합에 대한 중요도를 구할 수 있다[17]. 일례로 Gregorutti et al. [17]는 트리 모형에서의 기존 피루테이션 방법으로 계산하던 변수 중요도를 변수의 그룹으로 계산하는 방법을 제안하였다. 따라서 본 연구의 후속 연구도 마찬가지로 변수 각각이 아닌 변수들의 부분집합에 대한 중요도를 구하여 비교할 수 있는 연구의 필요성이 있다. 또한 수치형 데이터에는 LIME도 적용할 수 있으므로 함께 고려해볼 필요가 있다. 그리고 본 연구에서 수행한 의학 데이터뿐만 아니라 다른 도메인 데이터에도 XAI 기법을 적용해 볼 수 있

Figure 23. Grad-CAM (Gradient-weighted Class Activation Mapping) result of *vasc* image which is incorrectly classified for skin cancer data.

며 계속해서 연구개발되고 있는 최신 XAI 기법들도 함께 적용해야 할 것이다.

REFERENCES

1. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Model* 2004;178(3-4):389-397. DOI: 10.1016/j.ecolmodel.2004.03.013
2. Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint* 2018;arXiv:1610.01644.
3. Samek W, Wiegand T, Muller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv preprint* 2017;arXiv:1708.08296.
4. Gunning D. Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA); 2017.
5. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016, p. 1135-1144. DOI: 10.1145/2939672.2939778
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*; 2017, p. 618-626. DOI: 10.1109/ICCV.2017.74
7. Feng J, Williamson B, Simon N, Carone M. Nonparametric variable importance using an augmented neural network with multi-task learning. In *International Conference on Machine Learning*; 2018, p. 1496-1505.
8. Lamy JB, Sekar B, Guezennec G, Bouaud J, Seroussi B. Explainable ar-

- tificial intelligence for breast cancer: a visual case-based reasoning approach. *Artif Intell Med* 2019;94:42-53. DOI: 10.1016/j.artmed.2019.01.001
9. Sabol P, Sinčák P, Ogawa K, Hartono P. Explainable classifier supporting decision-making for breast cancer diagnosis from histopathological Images. In 2019 International Joint Conference on Neural Networks (IJCNN); 2019, p. 1-8. DOI: 10.1109/IJCNN.2019.8852070
10. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016, p. 2921-2929. DOI: 10.1109/CVPR.2016.319
11. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *proceedings of the IEEE International Conference on Computer Vision*; 2017, p. 618-626. DOI: 10.1109/ICCV.2017.74
12. Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting in-hospital mortality of ICU patients: the physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*; 2012, p. 245-248.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikitlearn: machine learning in Python. *JMLR* 2011;12(85):2825-2830.
14. Soo Y. Affects of white blood cell count for cardiovascular disease [dissertation]. Yonsei University Graduate School; Korea, 2006.
15. Wickham H. *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag; 2016.
16. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 2018;5:180161. DOI: 10.1038/sdata.2018.161
17. Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. *Comput Stat Data Anal* 2015;90:15-35.