Journal of Radiological Science and Technology, 43(6), 503-509 ISSN 2384-1168(On-line) ISSN 2288-3509(Print)

<원저>

머신러닝 기반 신체 계측정보를 이용한 CT 피폭선량 예측모델 비교

홍동희

신한대학교 방사선학과

Comparison of CT Exposure Dose Prediction Models Using Machine Learning-based Body Measurement Information

Dong-Hee Hong

Dept. of Radiological Science, Shinhan University

Abstract This study aims to develop a patient-specific radiation exposure dose prediction model based on anthropometric data that can be easily measurable during CT examination, and to be used as basic data for DRL setting and radiation dose management system in the future. In addition, among the machine learning algorithms, the most suitable model for predicting exposure doses is presented. The data used in this study were chest CT scan data, and a data set was constructed based on the data including the patient's anthropometric data. In the pre-processing and sample selection of the data, out of the total number of samples of 250 samples, only chest CT scans were performed without using a contrast agent, and 110 samples including height and weight variables were extracted. Of the 110 samples extracted, 66% was used as a training set, and the remaining 44% were used as a test set for verification. The exposure dose was predicted through random forest, linear regression analysis, and SVM algorithm using Orange version 3.26.0, an open software as a machine learning algorithm. Results Algorithm model prediction accuracy was R^2 0.840 for random forest, R^2 0.969 for linear regression analysis, and R^2 0.189 for SVM. As a result of verifying the prediction rate of the algorithm model, the random forest is the highest with R^2 0.986 of the random forest, R^2 0.973 of the linear regression analysis, and R^2 of 0.204 of the SVM, indicating that the model has the best predictive power.

Key Words : Machine learning, Random forest, Linear regression, Support vector machine, Exposure

중심 단어 : 머신러닝, 랜덤포레스트, 선형회귀분석, 서포트벡터머신, 피폭선량

Ⅰ. 서 론

전산화단층촬영(Computed tomography; CT)이 발전을 거듭하면서 사용의 보편화와 다양한 부위 검사까지 중요한 역할을 하고 있으며 검사 빈도가 크게 증가하고 있다[1]. CT 검사는 다른 검사에 비해 고선량이 사용되며 빈번한 검사 횟수와 진단 영상의 질적 향상을 위해 피폭선량에 대한 연구는 꾸준히 지속되어 왔다[2]. 특히, 검사자 및 환자에 대한 피폭 저감화 방안에 대한 연구가 가장 많으며 산란선 차폐를 위한 연구 와 선량 및 화질에 관한 연구 등의 피폭선량 관련 연구들이 계속되고 있다[3]. 그러나, 다양한 환경에서 피폭선량에 대한 예측을 활용해 적용시키는 연구는 부족한 상황이다.

환자권고선량((Dignosus reference level; DRL)은 의료피폭에 적용된 방사선 방어의 최적화로서 영상의학 검사에서 환자가 받는 피폭선량을 측정하여 진단에 참고하도록 권고하는 선량 준위이다[4-5]. 최근 CT 검사에 의한 방사선 노출로 암을 유발 할 수 있다는 연구들이 발표되며 방사선

This work was supported by the Shinhan University Research Fund, 2020
Corresponding author: Dong-Hee Hong, Department of Radiological Science, Shinhan University, 95, Hoam-ro, Uijeongbu-si, Gyeonggi-do, 11644, Republic of Korea / Tel: +82-31-870-3415 / E-mail: hansound2@hanmail.net
Received 11 December 2020; Revised 22 December 2020; Accepted 24 December 2020

피폭에 대한 관심이 증대되고 있으며, DRL에 대한 중요성이 대두되고 있다[6]. DRL 기준값 설정 시 전국 CT 장비의 실제 데이터를 기반으로 설정하지만 환자마다 갖는 신체적 특성이 다르고 검사 장비마다 갖는 프로토콜이 다르기 때문에 실질적인 피폭선량과 차이가 발생하게 된다. 그러므로 CT의 기본 원리인 방사선 감약 이론을 바탕으로 환자의 신체 계측 정보를 활용하여 각 병원마다 실제 사용되는 장비의 데이터 값을 머신러닝시켜 피폭선량 값을 예측 적용하는 과정이 필요하다.

최근 머신러닝 및 데이터마이닝은 의학 분야에서 질병 예측 및 식별을 위한 연구에 널리 이용되고 있으며[7-8] 인체계측정보를 이용한 고콜레스테롤 혈증 예측 모델에 관한 연구도 보고되었다[9]. 인공지능의 한 분야인 머신러닝은 빅데이터 기반 인공지능 학습법으로 알고리즘을 이용해 대량의 데이터를 분석하고 패턴을 인식하여 결과를 예측하는 방법이다. 크게 기계학습의 지도학습(Supervised modeling)과 비지도학습(Unsupervised modeling)으로 나뉘며, 이중 지도학습의 경우 이미 잘 알려진 데이터를 학습시켜 분류 및 예측에 활용된다[10]. 대부분 기계학습 알고리즘은 분류와 결과값 추정을 동시에 수행하지만 결과값 추정에 대표적인 알고리즘으로 선형회귀분석(Linear regression), 의사결정나무(Decision tree analysis), 신경망(Neural network), SVM(Surport vector machine), 랜덤포레스트(Random roest) 등이 있다. 본 연구는 CT 검사 시 쉽게 측정 가능한 신체계측 자료를 기반으로 환자맞춤 방사선 피폭선량 예측 모델을 개발하고 추후 DRL 설정과 방사선량 관리 시스템의 기초 자료로 활용되고자 한다. 또한, 기계학습 알고리즘 중 데이터마이닝에 최적인 선형회귀분석, SVM, 랜덤포레스트를 사용하여 분석하고 피폭선량 예측에 가장 적합한 모델을 비교 제시하고자 한다.

II. 연구방법

1. 연구대상 및 데이터 셋

본 연구에 사용한 데이터는 서울 소재 종합병원의 흉부 CT 검사 자료로써 환자의 개인정보를 제외하고 신체 계측 자료가 포함된 데이터를 기준으로 데이터 셋(data set)을 구성하였다. 데이터의 전처리와 샘플 선별에 있어, 전체 샘플 수 250개 중 조영제 사용 없이 흉부 CT 검사만 진행된 샘플을 추출하였고 이 중 키와 몸무게 변수가 포함된 샘플 110개를 추출하였다. 추출된 110개 샘플 중 66%는 훈련 셋으로 사용하고, 나머지 44%는 검증을 위한 테스트 셋으로 사용하였다. 변수(feature) 선정 시 피폭선량에 관계되는 신체 계측 자료 즉, 키와 몸무게를 포함한 실제 관전압(kVp), 관전류(mAs), CTDIvol(CT dose index volume), 조사시간(TI), 절편두께(Slice thickness), 성별, 나이를 독립변수로 DLP(dose length product)를 목표변수로 사용하였다(Table 1).

2. 예측모델 및 분석 방법

본 연구는 신체 계측에 따른 피폭선량 예측을 위해 다음 Fig. 1과 같이 예측모델을 구성하였다. 구성된 데이터 셋을 기준으로 신체 계측에 따른 피폭선량 예측모델을 개발하고자 하였다. 다음으로 랜덤포레스트, 선형회귀분석, SVM 알고리즘을 통해 예측모델을 생성하여 예측률 및 정확도를 파악하고자 하였다.

본 연구의 예측 모델 개발에 사용되고 있는 머신러닝 알고리즘으로 오픈 소프트웨어인 Orange version 3.26.0을 사용하여[11] 랜덤포레스트, 선형회귀분석, SVM 알고리즘을 통해 피폭선량을 예측하였다. Orange는 오픈 소스 Python 기반(Anaconda Mini 버전포함)의 데이터 시각화, 머신러닝 및 데이터 마이닝 툴킷이며 탐색적 데이터 분석 및 대화식 데이터 시각화를 위한 시각적 프로그래밍 프로트 엔드가 특징이다. 머신러닝 알고리즘은 분류 및 판별과 결과값 추정으로 나눌 수 있는데 본 연구는 결과값을 추정하기에 분류와 결과값 추정이 모두 가능한 랜덤포레스트과 SVM을 사용하였고, 범주형 목표 변수에 최적인 선형회귀분석을 사용하여 3가지 알고리즘의 정확도와 예측률을 비교하였다.

3. 머신러닝

기계학습은 지도학습과 비지도학습으로 구분되며 지도학습은 목표변수가 있는 경우이고, 비지도학습은 목표변수가 없는 경우이다. 알고리즘의 목적에 따른 구분에서 분류 및 판별과 결과값 추정의 알고리즘은 지도학습에 속하며, 연관성 규칙과 군집화는 비지도학습에 속한다.

본 연구는 지도학습에 속하며 특히 결과값 추정 알고리즘이 적합하다. 이에 해당하는 알고리즘은 선형회귀분석(linear regression analysis), 의사결정나무(decision tree analysis), 신경망(neural network), SVM(support vector machine), 랜덤포레스트(random forest)가 해당된다. 이에 영상분석보다 데이터마이닝에 많이 쓰이는 선형회귀분석, SVM, 랜덤포레스트를 대상으로 분석하고자 한다.

Table 1. Variable information and characters (*n*=110)

Independent variable	Mean	stedv	Median
Height(cm)	161.70	±9.10	162.0
Weight(kg)	60.90	±12.05	60.0
age	71.60	±10.90	73.5
kVp	120.00	±0.00	120.0
mAs	64.52	±29.30	58.5
TI(s)	0.50	±0.00	0.5
slice thickness(mm)	0.60	±0.00	0.6

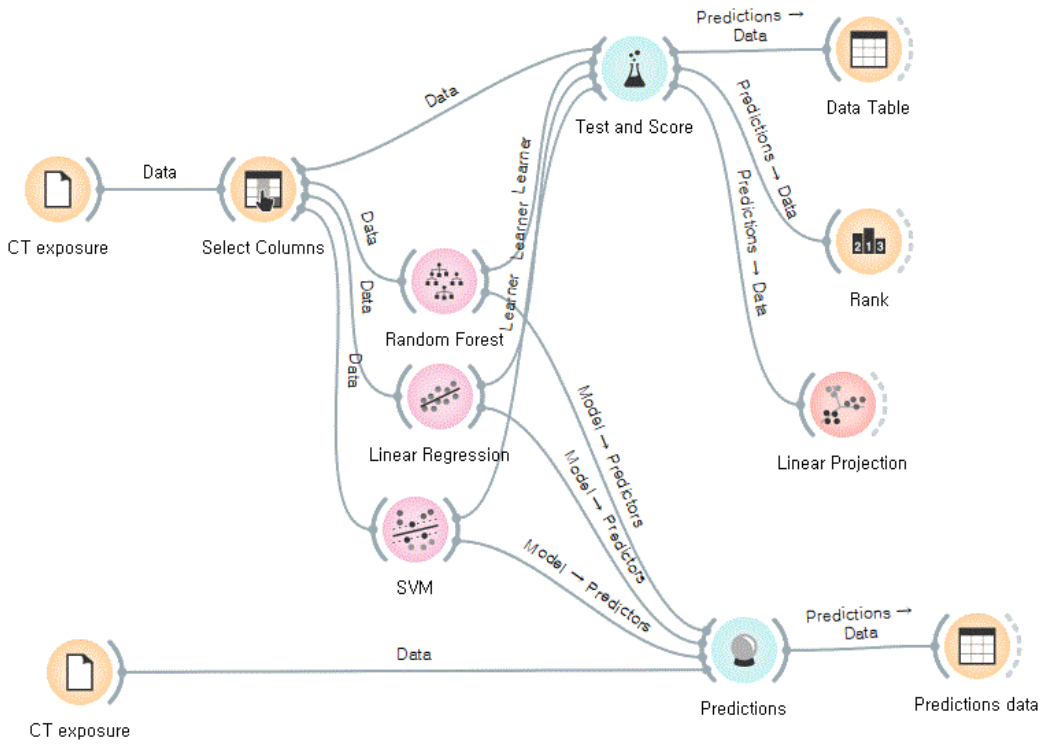

Fig. 1. Predicted model

1) 랜덤포레스트

랜덤포레스트는 의사결정나무분석 중 CART 알고리즘과 앙상블 모형 중 배깅 알고리즘을 적용한 알고리즘이다. 학습방법은 신경망의 MLP·RBF·SVM과 같은 지도학습이다. 이는 CART를 기반으로 하고 있어 분포에 대한 가정이 없고 목표변수와 입력변수의 타입에도 자유로워 제약조건이 거의 없다. 또한, 의사결정나무의 약점인 과대적합(over-fitting) 문제를 해결하고 앙상블 모형의 장점인 예측 정확도를 높인 알고리즘이다. 예측력이 가장 좋으며 결측지 자료에 유용하고, 변수변환이 필요 없으며 과대적합하지 않는다. 그러나, 학습시간이 과다하게 소요되며 데이터셋에 레코드와 변수가 적은 경우 모형 적합도가 높지 않을 수 있다[10].

2) 선형회귀분석

선형회귀 분석은 변수들 사이의 상관관계를 분석하는 데 사용하는 방법이다. 선형회귀 분석은 독립변수와 종속변수 사이의 관계를 모델링하는 방법이다[10]. 두 변수 사이의 관계일 경우 단순 선형회귀라고 하며, 여러 개의 변수일 경우 다중 선형회귀도 있다. 다중공선성은 독립변수들 사이에 상관관계가 발생하는 현상이며 이때 회귀계수에 대한 해석이 불가능해진다.

3) SVM

SVM은 분류 및 판별과 추정을 할 수 있는 분석 알고리즘으로 지도학습에 해당한다. 목표변수가 존재할 경우 분류

및 판별 예측이나 연속형 값(점추정) 예측을 할 때 사용하는 알고리즘이다. 기본적으로 두 범주를 갖는 관측값들을 분류하는 방법으로 주어진 데이터들을 멀리 2개의 집단으로 분리시키는 최적의 초평면(hyperplane)을 찾는데 중점을 둔다. 예측력이 우수하고, 모형 산출에 대한 가정이 없으며 변수타입에 자유롭지만 파라미터 C(단위비용)와 커널 선택에 따라 모형이 민감한 단점을 지닌다[10].

Ⅲ. 결 과

1. 알고리즘 모델 예측 정확도

SVM은 분류 및 판별, 추정을 할 수 있는 알고리즘으로 커널(kernel)은 시그모이드 커널(sigmoid kernel)을 설정하였고, 회귀정도를 나타내는 엡실론(epsilon)은 0.1로 설정하여 진행하였다. 랜덤 포레스트 알고리즘은 10개의 붓스트랩을 기준으로 진행하였고 선형회귀분석은 정규화를 위해 회귀계수를 Ridge방법을 이용하여 축소 진행하였다.

이러한 랜덤포레스트, 선형회귀분석, SVM의 알고리즘을 활용한 예측변수의 정확도 및 중요도를 살펴보기 위해 5개를 층화시켜 교차타당성을 분석하였다〈Table 2〉. 평가모델의 정확도는 평균제곱오차(MSE), 평균제곱근오차(RMSE), 평균절대오차(MAE)가 작을수록 좋은 모델이고, 설명력 지수인 R2가21에 가까울수록 좋은 모델로 평가한다. 분석결과, 랜덤포레스트의 MSE는 1026.416, RMSE는 32.038, MAE는 15.677, 선형회귀분석의 MSE는 199.107, RMSE는 14.111, MAE는 10.299, SVM의 MSE는 5208.721, RMSE는 72.171, MAE는 40.748로 선형회귀분석이 가장 낮게 분석되었다. 또한, 랜덤포레스트의 R2는 0.840, 선형회귀분석의 R2는 0.969, SVM의 R2는 0.189로 선형회귀분석이 가장 높은 것으로 나타나 가장 좋은 모델로 분석되었다.

예측변수의 중요도는 Table 3, Fig. 2와 같다. 랜덤포레스트는 관전류(0.205), 몸무게(0.192), 나이(0.140), 키(0.122), 성별(0.024)순으로 나타났고, 선형회귀분석은 몸무게(0.173), 관전류(0.170), 나이(0.153), 키(0.129), 성별(0.041) 순으로 중요도가 높게 나타났으며 SVM은 관전류(0.214), 몸무게(0.201), 나이(0.180), 키(0.151), 성별(0.043)순으로 중요도가 높게 나타났다. 관전압, 조사시간, 절편두께는 0.000으로 같게 나타났다.

2. 알고리즘 모델 예측률 검증

랜덤포레스트, 선형회귀분석, SVM의 알고리즘을 활용한 본 연구의 모형 예측력 검증결과는 Table 4와 같다. 평가모델의 정확도와 마찬가지로 평균제곱오차(MSE), 평균제곱근오차(RMSE), 평균절대오차(MAE)가 작을수록, 설명력 지수인 R2가 1에 가까울수록 예측력에 대한 신뢰도가 높은 것으로 평가한다. 분석결과, 랜덤포레스트의 MSE는 89.433, RMSE는 9.457, MAE는 6.406, 선형회귀분석의 MSE는 176.220, RMSE는 13.275, MAE는 9.802, SVM의 MSE는 5110.783, RMSE는 71.490, MAE는 39.026으로 랜덤포레스트가 가장

Table 2. Results of model accuracy comparison

Model	MSE	RMSE	MAE	R2
Random forest	1026.416	32.038	15.677	0.840
Linear regression	199.107	14.111	10.299	0.969
SVM	5208.721	72.171	40.748	0.189

Table 3. Prediction importance

Predictors	Random forest	Linear regression	SVM
Height	0.122	0.129	0.151
Weight	0.192	0.173	0.201
Age	0.140	0.153	0.180
kVp	0.000	0.000	0.000
mAs	0.205	0.170	0.214
TI	0.000	0.000	0.000
Slice thickness	0.000	0.000	0.000
Sex	0.024	0.041	0.043

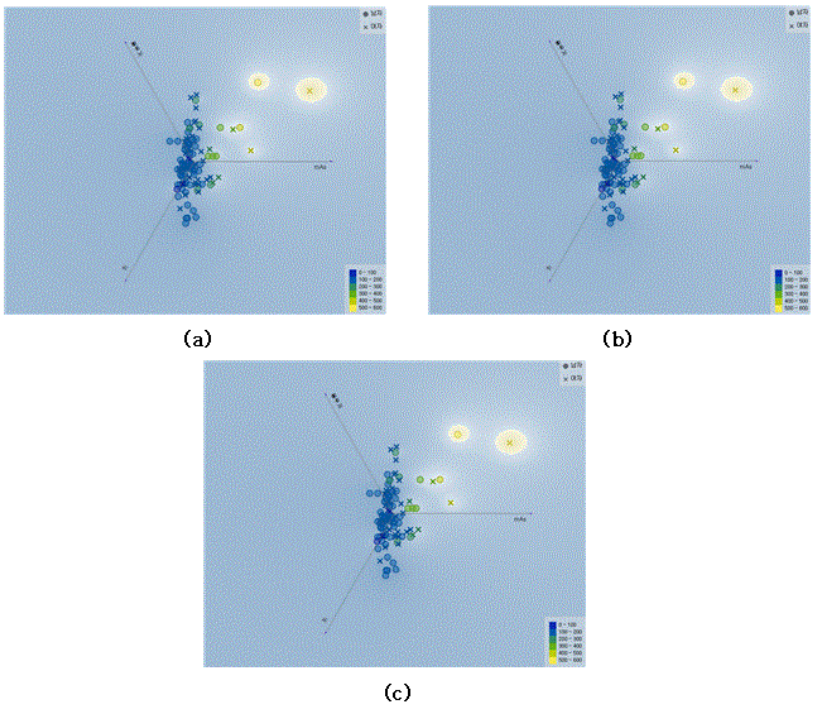

Fig. 2. Scopre plots of prediction importance (a) Random forest (b) Linear regression (c) SVM

Table 4. Results of prediction accuracy comparison

Model	MSE	RMSE	MAE	R2
Random forest	89.433	9.457	6.406	0.986
Linear regression	176.220	13.275	9.802	0.973
SVM	5110.783	71.490	39.026	0.204

낮게 분석되었다. 또한, 랜덤포레스트의 R2는 0.986, 선형회귀분석의 R2는 0.973, SVM의 R2는 0.204로 랜덤포레스트가 가장 높은 것으로 나타나 가장 예측력이 좋은 모델인 것을 알 수 있다.

Ⅳ. 고 찰

영상의학과 검사는 환자의 질병 진단에 도움이 되는 반면, 방사선 피폭의 위험성이 존재한다. 미국에서 CT로 발생되는 암이 전체 암 발생 중 2%를 차지한다고 발표하였다[12]. 이는 CT 발전으로 암의 조기 진단 및 치료에 큰 도움이 되었으나 CT 검사로 인한 피폭선량이 잘 관리 되지 못하는 것에 대한 사회적 우려를 증가시키고 있다. 이러한 관심과 우려가 증가되고 있는 현 상황에 맞춰 환자피폭선량관리는 반드시 필요한 상황이며[13], 특히 CT 검사 시 권고 선량인 DRL은 환자 개인 특성과 각 장비 프로토콜에 맞춘 값이 아니기에 실질적인 DRL 재정립이 필요하다.

선행연구에서 환자의 검사 프로토콜이 다르기 때문에 피폭선량도 다양할 수밖에 없으며, 국제기준의 환자 선량 권고량을 고려하여 합리적으로 낮추기 위한 체계화된 프로토콜 제정이 필요하다고 하였다[2, 14-15]. 그러기 위해선 다양한 환자와 장비 특성을 고려하여 각 병원마다 예측된 피폭선량을 기반으로 정립되어야 실제 임상에서 활용할 수 있는 의미 있는 기준이 될 수 있으리라 여겨진다.

또한, 임상에서 CT 검사 진행 시 평소보다 많은 선량이 피폭된 것을 종종 확인하게 된다. 이는 환자에게 과잉 노출된 경우이며 투여된 선량이 적정수준인지 여부를 판단할 수 있는 근거가 필요하다. 그러나 현재 임상에서 근거로 제시하는 DRL은 특수성을 반영하지 못한 결과이기에 특수한 경우의 수를 반영한 피폭선량 예측자료가 필요하다. 그러므로 특수한 경우의 수를 반영한 예측 피폭선량 값을 알아보기

위한 다양한 머신러닝 기법을 활용하였고 피폭선량에 가장 적합한 모델링을 제시하고자 하였다.

앞서 연구한 체지방 측정 정보를 이용한 고콜레스테롤혈증 예측은 CFS 기반 naive bayes 모델을 이용하였고 변수들을 통합하여 머신러닝 적용함으로써 예측력을 높였다[2]. 본 연구에서 사용된 모델은 연속형 예측 모델이기에 선행연구에서 사용된 분류 모델과 차이는 있지만 가장 변수 중요도가 높게 분석된 몸무게와 관련된 변수를 추가하고 러닝 기법을 달리한다면 충분히 예측력을 높일 수 있을 것이라 예상된다.

현재 CT 검사에 활용하는 머신러닝 기법은 많지 않지만 우상근 연구[16] 결과에서 실제 조영제 사용 없이 영상생성 머신러닝 수행만으로 조영 증강된 영상을 획득함으로써 불필요한 피폭선량을 최소화하는 등의 연구들이 활발히 연구들이 시도되고 있다[17]. 본 연구는 시간과 고도의 기술이 필요한 영상 생성 머신러닝보다 쉽게 접할 수 있는 데이터 마이닝에 기반한 머신러닝 모델을 활용한 것이며 팬텀이 아닌 실제 환자 데이터를 바탕으로 피폭선량을 최소화하는데 노력하였다.

본 연구에 사용된 머신러닝 모델 중 랜덤포레스트와 선형회귀분석은 높은 정확도를 보인 반면 SVM은 굉장히 낮은 정확도를 보였다. 비록 피폭선량의 활용 모델로 SVM은 어려울 것으로 분석되었지만 이는 SVM의 단점인 변수의 민감도가 반영된 결과라고 여겨지며 추후 연구에 독립변수와 데이터 셋의 양을 늘린다면 SVM의 정확도도 올라가 충분히 활용가치가 있을 것이라 예상된다. 앞으로 본 연구 결과를 토대로 피폭선량 예측에 머신러닝 활용이 활발히 이용될 수 있는 기초자료로 활용될 것이라 기대된다.

V. 결 론

본 연구는 신체계측 자료를 기반으로 CT 검사 시 피폭선량을 예측하여 산출할 수 있는 최적의 머신러닝 알고리즘을 제시하고 기초자료로 사용하고자 하였고, 그 결과는 다음과 같다.

연구에 사용된 머신러닝 알고리즘은 랜덤포레스트, 선형회귀분석, SVM 알고리즘이며 이중 모형 정확도는 선형회귀분석이 가장 정확하였고 예측값에 대한 정확도가 가장 높은 알고리즘은 랜덤포레스트였다. 각 모델의 예측에 가장 큰 요인으로 몸무게, 관전류, 키였으며 이는 쉽게 CT 검사 시 획득할 수 있는 자료이므로 임상에 적용 시 활용도가 높을 것으로 예상된다. 또한, 랜덤포레스트와 선형회귀분석의 모델 정확도와 예측 정확도 차이가 크지 않아 두 모델의 단점을 보완하여 사용한다면 둘 다 피폭선량을 예측 활용하는데 신뢰도 높은 알고리즘이 될 것으로 판단된다.

REFERENCES

[1] UNSCEAR. Sources and effects of ionizing radiation. UNSCEAR 2010 Report, New York, United Nations; 2010.

[2] Lee SY, Kim KL, Ha HK, et al. Evaluation of radiation exposure dose for examination purposes other than the critical organ from computed tomography: A base on the Dose Reference Level (DRL). Journal of the Korean Society of Radiology. 2013;7(2):121-9.

[3] Mo KH. Analysis of exposure dose according to chest and abdomen combine CT exam method [Dept. of Radiology]. Graduate School of Health Science, Eulji University; 2016.

[4] Fukushima Y, Tsushima Y, Takei H, et al, Diagnostic reference level of computed tomography (CT) in Japan. Radiat Prot Dosimetry. 2012;151(1):51-7.

[5] IAEA. International basic safety standards for protection against ionizing radiation and for the safety of radiation sources. IAEA Safety Series No. 115; 1996.

[6] Brenner DJ, Hall EJ. Computed tomography: An increasing source of radiation exposure. N Engl J Med. 2007;357(22):2277-84.

[7] Lee BJ, Kim JY. Identification of the best anthropometric predictors of serum high- and low-density lipoproteins using machine learning. IEEE J Biomed Health Inform. 2015;19(5):1747-56. doi:10.1109/JBHI.2014.2350014.

[8] Lee BJ, Kim JY. Indicators of hypertriglyceridemia from anthropometric measures based on data mining. Comput Biol Med. 2015;57:201-11. doi:10.1016/j.compbiomed.2014.12.005.

[9] Lee BJ, Prediction model of hypercholesterolemia using body fat mass based on machine learning. The Journal of the Convergence on Culture Technology. 2019;5(4):413-20.

[10] Cho YJ. Big Data, New SPSS Analysis Technique;

Neural Network, SVM, Random Forest. Hanarae Academic; 2018.

[11] Carlos MR, Hilario ML, Data mining for the study of the Epidemic (SARS-CoV-2) COVID-19: Algorithm for the identification of patients speaking the native language in the Totonacapan area - Mexico. Munich Personal RePEc Archive. 2020;102039:1-14.

[12] Brenner DJ, Hall EJ. Computed tomography: An increasing source of radiation exposure. N Engl J Med. 2007;357(22):2277-84.

[13] Lee CH. Individualized and intelligent radiation dose exposure guide and management system with clinical test operation. Health Technology R&D Project; 2017.

[14] Kalender WA. Computed tomography. John Wiley and Sons, New York; 2000.

[15] Dougeni E, Faulkner K, Panayiotakis G. A review of patient dose and optimisation methods in adult and paediatric CT scanning. Eur J Radiol. 2012; 81(4):e665-83.

[16] Woo SK, Synthesis of contrast CT image using deep learning network. Proceedings of the Korean Society of Computer Information Conference. 2019;465-7.

[17] Hong JY, Jung YJ. Evaluation of deep-learning feature based COVID-19 classifier in various neural network. Journal of the Korean Society of Radiology. 2020;43(5):397-404.

구분	성명	소속	직위
단독	홍동희	신한대학교 방사선학과	조교수