JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 7, pp. 9-17, July 2021
https://doi.org/10.9708/jksci.2021.26.07.009

Classification models for chemotherapy recommendation using LGBM for the patients with colorectal cancer

Seo-Hyun Oh*, Jeong-Heum Baek**, Un-Gu Kang*

*Student, Dept. of Computer Engineering, Gachon University, Sungnam, Korea
**MD, Ph.D, Dept. of Surgery, Gil Medical Center, Incheon, Korea
*Ph.D, Dept. of Computer Engineering, Gachon University, Incheon, Korea

[Abstract]

In this study, we propose a part of the CDSS(Clinical Decision Support System) study, a system that can classify chemotherapy, one of the treatment methods for colorectal cancer patients. In the treatment of colorectal cancer, the selection of chemotherapy according to the patient's condition is very important because it is directly related to the patient's survival period. Therefore, in this study, chemotherapy was classified using a machine learning algorithm by creating a baseline model, a pathological model, and a combined model using both characteristics of the patient using the individual and pathological characteristics of colorectal cancer patients. As a result of comparing the prediction accuracy with Top-n Accuracy, ROC curve, and AUC, it was found that the combined model showed the best prediction accuracy, and that the LGBM algorithm had the best performance. In this study, a chemotherapy classification model suitable for the patient's condition was constructed by classifying the model by patient characteristics using a machine learning algorithm. Based on the results of this study in future studies, it will be helpful for CDSS research by creating a better performing chemotherapy classification model.

▸ **Key words**: Colorectal Cancer, Chemotherapy, Artificial Intelligence, Machine learning, classification model

[요 약]

본 연구는 대장암 환자의 치료방법 중 하나인 항암화학요법을 분류할 수 있는 시스템인 CDSS연구의 일환으로 시행되었다. 대장암 치료에서 환자의 상태에 맞는 항암화학요법의 선택은 환자의 생존 기간과 직결되기 때문에 매우 중요하다. 따라서 본 연구에서는 대장암 환자의 개인적, 병리학적 특성을 사용해 기저 모델, 병리학적 모델, 그리고 환자의 두 가지 특성을 모두 사용한 결합 모델을 만들어 머신러닝 알고리즘으로 항암화학요법을 분류하였다. Top-n Accuracy와 ROC 곡선, AUC로 모델의 예측 정확도를 비교한 결과, 결합 모델에서 가장 우수한 예측 정확도를 보였으며, LGBM 알고리즘의 성능이 가장 우수한 것을 알 수 있었다. 본 연구에서는 머신러닝 알고리즘을 이용해 환자 특성별 모델을 분류함으로써 환자의 상태에 맞는 항암화학요법 분류 모델을 구축하였다. 향후 연구에서 본 연구 결과를 기초한다면 더 좋은 성능의 항암화학요법 분류 모델을 만들어 CDSS 연구에 도움이 될 것이다.

▸ **주제어**: 대장암, 항암화학요법, 인공지능, 머신러닝, 분류 모델

• First Author: Seo-Hyun Oh, Corresponding Author: Un-Gu Kang
*Seo-Hyun Oh (dlsanf2000@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
**Jeong-Heum Baek (gsbaek@gilhospital.com), Dept. of Surgery, Gil Medical Center
*Un-Gu Kang (ugkang@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
• Received: 2021. 06. 10, Revised: 2021. 07. 01, Accepted: 2021. 07. 01.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

1. Introduction

우리나라의 암 발생은 2018년 기준 전체 암 발생 중 8.6%이고 대장암이 새로 발병한 환자는 15.3%이며 전체 암 발생률 중 2위에 해당한다[1]. 대장암의 원인으로는 50세 이상의 연령, 동물성 지방과 포화지방이 많은 음식 섭취, 신체활동 부족, 비만, 음주, 유전적 요인이 있으며, 한국뿐만 아니라 세계적으로 대장암은 증가하는 추세이다. 2018년 세계보건기구에 따르면 대장암 발생은 10.2%로 세계적으로 전체 암 발생률 3위이며, 사망률은 9.2%로 2위를 차지하였다[2].

대장암의 치료에서 가장 기본이 되는 치료법은 근치적 수술이고[3], 방사선 치료와 보조적 항암화학요법이 전이 및 재발을 줄이기 위해 사용되고 있다[4]. 치료 방법은 대장암의 종양 침윤정도, 림프절(Lymph node) 전이여부와 원격장기의 전이 여부 등에 따라 결정된다[5]. 대장암은 병기가 높을수록 재발률이 높으며 생존율이 낮아진다[6]. 대장암 1기는 재발의 가능성이 낮아 근치적 수술 후 경과를 관찰하며 항암화학요법을 실시하지 않는 경우가 대부분이다. 2기는 근치적 수술 후 재발 위험인자가 발견된 환자에게 항암화학요법이 실시된다[7]. 3기는 근치적 수술 후 미세잔존 종양, 전이를 차단하기 위해 항암화학요법을 실시하며, 실시하지 않을 경우 재발률이 50~60% 높게 나타난다고 보고된다[8-13]. 4기에는 증상 완화와 생존연장을 위해 항암화학요법을 시행하게 된다[14].

환자의 상태 및 병기에 따라 다르게 사용되는 항암화학요법은 5-FU와(Fluorouracil) 이를 기초로 한 병합 요법이 많이 사용되었다. 그 이후, Oxaliplatin이 개발되어 5-FU 제제와의 복합요법인 FOLFOX 등이 시도되었다[15-16]. 1980년대 이후 FOLFOX 도입 이전에, leucovorin(LV)이 도입되어 5-FU의 modulator로 근치적 수술 후 보조화학요법에서 좋은 치료 성과를 얻는데 기여하였다[17]. 환자의 상태에 따라 항암화학요법 선택의 폭도 넓어졌다. 선택의 폭이 넓어짐에 따라, 환자의 상태에 맞는 적절한 항암화학요법 사용도 중요한 문제로 대두되었다. 항암화학요법은 환자의 생존기간과 직결되는 문제이기에 적절하게 사용하는 것이 중요하다[18].

따라서 본 연구에서는 항암화학요법 분류를 위해 머신러닝 알고리즘을 이용하였다. 환자의 특성에 따라 항암화학요법별로 분류하는 머신러닝 알고리즘을 적용해 예측 정확도를 비교하고, 변수 중요도를 파악하여 항암화학요법 분류에 관한 임상 의사결정 지원시스템(Clinical Decision Supporting System: CDSS) 연구의 기초자료를 제시하고자 한다.

2. Purpose

본 연구의 목적은 대장암 환자의 수술 후 항암화학요법 추천을 위해 전자의무기록(EMR: Electronic Medical Record)을 활용하여 항암화학요법을 분류하는 머신러닝 알고리즘의 예측 정확도를 비교하는 것이다.

II. Related Works

1. Colorectal Chemotherapy

Hwa Jung Kim[12]의 연구에서는 대장암 근치적 수술 후 보조적 항암화학요법 사용이 환자의 생존기간에 어떤 관련이 있는지 알아보았다. 대장암 3기 환자에서 근치적 수술 후 보조적 항암화학요법의 지연 적용뿐만 아니라 조기 적용도 사망의 위험을 증가시킨다고 나타났다. 또한, 교란변수 보정 이후에도 관련성이 유지되었다고 말했다. 따라서, 3기 대장암환자 진료 시, 사망의 발생 예방을 위하여 근치적 수술 후 3~8주에 보조적 항암화학요법을 수행하도록 치료를 계획하는 것이 필요하다는 사실을 알 수 있었다.

위와 같은 선행 연구에서 항암화학요법 결정이 생존기간에 많은 영향을 미치고, 이는 곧 환자의 사망과 직결되는 문제라는 것을 알 수 있었다. 따라서 본 연구에서는 궁극적으로 환자의 생존기간에 영향을 미치는 항암화학요법을 Target 변수로 설정하여 분류 모델을 구축하였다.

2. CDSS(Clinical Decision Support System)

최근에는 EMR과 의료 정보기술 인프라의 구축으로 환자의 개인 건강 정보의 활용이 용이해졌다. 이에 본 연구와 같은 EMR 데이터 분석이 증가하는 추세이다[19].

EMR 데이터를 활용한 대표적인 사례 중 하나에는 CDSS가 있다. CDSS는 질병 진단 및 치료 시, 환자로부터 얻은 임상정보인 EMR을 바탕으로 의사결정을 도와주는 임상 의사결정 지원시스템이다[20].

Watson for Oncology(WfO)는 대표적인 CDSS 중 하나이다. WfO는 암 진단 및 치료 의사결정을 지원하는 시스템으로, 의료 빅데이터를 학습시킨 모델로 만들어졌다[21]. Koichi[22]의 연구에서는 WfO를 이용해 백혈병 환자 200명을 대상으로 진단 일치율을 보았고, 대장암과 자궁 경부암을 포함한 특정 암 종류에 대해 높은 진단 일치율을 나타내었다. 그러나, WfO는 외국 데이터를 기반으로

학습되었기 때문에 한국 데이터에서는 일치율이 떨어지는 현상을 보였다[23]. 따라서 본 연구에서는 가천대학교 길 병원 EMR의 데이터를 적용하여 항암화학요법 분류 모델을 구축하였다.

3. Machine Learning Model

3.1 kNN(k-Nearest Neighbors)

kNN은 값을 구하고자 하는 데이터로부터 가장 가까운 거리에 있는 데이터 포인트, 즉 '최근접 이웃(Nearest Neighbors)'을 찾는 알고리즘이다. 구하고자 하는 데이터와 최근접 이웃 데이터와의 거리를 계산하고 그에 비례한 가중치를 부여함으로써 분류한다[24-25].

3.2 SVM(Support Vector Machine)

SVM은 데이터가 어떤 카테고리에 분류될지 판단하는 이진 선형 분류 알고리즘이다. 학습하면서 분류 기준을 정의하고, 해당 기준을 바탕으로 데이터를 분류한다[26-27].

3.3 Decision Tree

DT는 분류 과정이 트리 구조로 표현되는 알고리즘이다. 트리 구조로 만들기 위해서 분류하는 목적과 데이터 종류에 알맞은 규칙을 적용하여 데이터를 분류한다[28].

3.4 Random Forest

RF는 학습을 통해 구성해놓은 다수의 의사결정트리로부터 분류 결과를 취합해서 정확도를 높이는 알고리즘이다[29-30].

3.5 LGBM(Light Gradient Boosting Machine)

LGBM은 맞지 않게 분류된 데이터에 가중치를 더해서 정확도를 높이는 알고리즘이다. Tree 기반 알고리즘이며, 틀린 부분에만 가중치를 더하기 때문에 leaf-wise 알고리즘이다[31-32]. LGBM의 구조는 Fig 1과 같다[33].

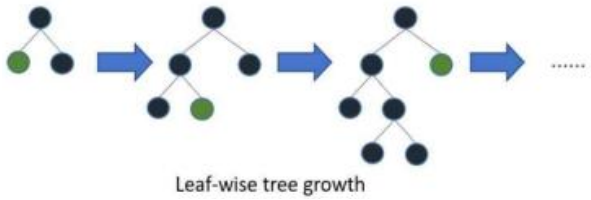

Fig. 1. Structure of LGBM

III. Method

1. Research Environment

본 연구에서는 항암화학요법 분류 모델 구축을 위해 구글 코랩(Google Colaboratory)을 사용하였다. 코랩은 구글에서 제공하는 클라우드 실험 환경으로 웹 브라우저에서 Python을 작성하고 실행할 수 있으며 컴퓨터 실험 환경에 관계없이 실행이 가능하다. 본 연구에서 사용한 Colab의 사양은 "Python 3.7.10, Ubuntu 18.04.5 LTS, Intel® Xeon® CPU @ 2.00GHz, MemTotal:13305368 kB"이다. 데이터 전처리 및 연산에 Numpy, Pandas 프레임워크를 사용하였으며 모델 학습에는 Scikit-learn 라이브러리를 사용하였다.

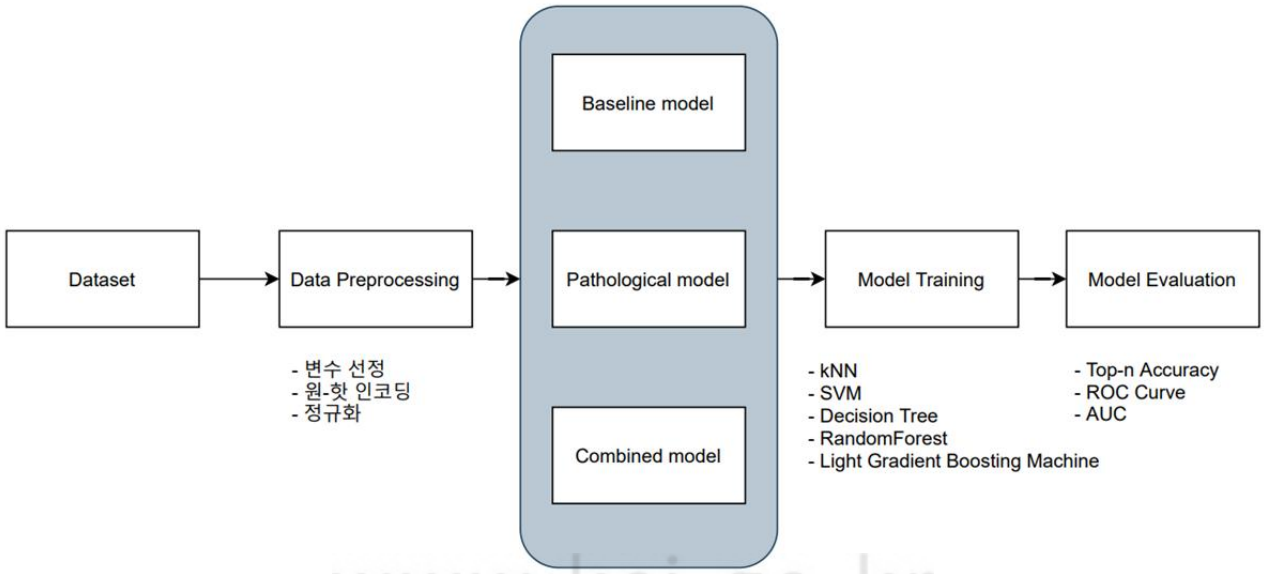

Fig. 2. Pipeline of Research

2. Research Method

본 연구의 파이프라인은 Fig 2이다. 먼저 Dataset을 머신러닝 모델 학습에 적합하게끔 변수 선정, 원-핫 인코딩, 정규화와 같은 데이터 전처리 과정을 거쳤다. 전처리 과정을 거친 후, 개인적 특성을 사용한 모델과 병리학적 특성을 사용한 모델, 두 가지 특성을 모두 사용한 결합 모델을 구축하였다. 그리고 분류한 모델들을 각각 머신러닝 알고리즘에 학습시켰다. 사용한 알고리즘은 kNN, SVM, DT, RF, LGBM 5가지이다. 학습 후, Top-n Accuracy, 변수 중요도, ROC 곡선, AUC를 통해 모델의 예측 정확도를 알아보는 순서로 진행하여 논문의 결과를 도출하였다.

2.1 Dataset

본 연구에서는 가천대학교 길병원에서 2004년부터 2018년 사이에 대장암 수술을 받은 3,321명의 환자 EMR의 데이터를 사용하였으며 최종적으로 변수 선정 및 결측치 처리 과정을 거쳐 3,086명에 대해 항암화학요법 분류 모델을 구축하였다.

2.2 Data Preprocessing

결측치 처리는 선정된 변수 중 결측치를 가진 행을 제거하는 방식을 택하였다. 또한, 결측치가 30%를 초과하는 변수는 전처리 과정에서 제거하였다.

범주형 변수는 모델에 적용하기 위해 원-핫 인코딩(One-hot Encoding)방법을 적용하여 전처리하였다. 원-핫 인코딩은 n개의 속성을 가진 변수에 대해서 n차원의 희소벡터 형태로 데이터를 변환하는 방법이다. 머신러닝 알고리즘은 입력값으로 문자열을 받지 않기 때문에 문자열을 숫자형의 벡터로 인코딩하는 원-핫 인코딩이 필요하다[34].

연속형 변수는 정규화 방법을 사용하여 범위를 0과 1 사이로 조정하였다. 정규화는 변수의 단위를 무시하고 모델 구축에 사용할 수 있도록 해주는 작업이다. 본 연구에서는 Standard Scaler를 사용해 평균을 0, 표준편차를 1로 변환하여 모든 변수들이 같은 범위를 갖게 했다. Standard Scaler(1)는 원래 변수에서 평균(2)을 뺀 값을 표준편차(3)로 나누어 계산하며, 아래 수식과 같다[35].

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \quad (2)$$

$$\alpha = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (3)$$

2.3 Feature Selection

전체 변수 중, 임상의가 선정한 항암화학요법 관련 변수들을 대상으로 변수를 선택하였다.

Feature는 개인적 특성과 병리학적 특성으로 구분할 수 있고 Table 1은 특성별로 변수를 기술한 표이다. 본 연구에서는 개인적 특성을 사용한 기저 모델(Baseline model)과 병리학적 특성을 사용한 병리학적 모델(Pathological model), 그리고 두 가지 특성을 모두 사용한 결합 모델(Combined model)을 구축하여 실험하였다.

종속 변수는 대장암 환자 치료법 중 하나인 항암화학요법 종류로 설정하였다. 본 연구에서는 보험적용이 가능하며 일반적으로 사용되는 항암화학요법 5종과 항암화학요법을 사용하지 않은 Surveillance를 대상으로 분석을 진행하였다. 항암화학요법 종류에는 5-FU/LV, Capecitabine, FOLFOX, XELOX, FOLFIRI를 분석에 사용하였다.

2.4 Chemotherapy classification model building

본 연구에서는 kNN, SVM, DT, RF, LGBM 알고리즘을 사용하여 환자의 특성별로 기저 모델과 병리학적 모델, 두 특성을 결합한 결합 모델을 구축하였다. 기저 모델은 환자의 개인적인 특성만을 가지고 어떤 대장암 항암화학요법을 사용할지를 예측하는 모델이다. 병리학적 모델은 환자의 임상적 정보를 담은 병리학적 특성만으로 대장암 항암화학요법을 예측하는 모델이다. 결합 모델은 개인적, 병리학적 특성을 모두 사용한 모델로 본 연구에서 제시하고자 하는 모델이다.

본 연구에서는 5개의 머신러닝 알고리즘을 사용하여 각 모델의 예측 정확도와 ROC 곡선을 비교하였다.

2.5 Performance Evaluation

본 연구에서는 알고리즘의 예측 정확도의 평가지표로 각 분류 모델의 정확도(Accuracy)와 ROC 곡선(Receiver Operating Characteristic Curve), AUC(Area Under Curve)를 사용하였다.

정확도는 해당 분류 알고리즘이 데이터를 얼마나 정확하게 분류하였는가를 나타내며, 높을수록 분류 정확도가 높아 성능이 좋다는 것을 의미한다. 본 연구에서는 Top-n Accuracy(Top-2 Accuracy)를 사용하여 예측 정확도를 비교하였다. Top-n Accuracy는 true 클래스가 모델에 의해 예측된 가장 가능성이 높은 클래스 n개 중 하나와 일치하는 정확도를 산출하는 방법이다[36]. Top-n Accuracy를 사용하여 결과를 산출하는 방법은 선행연구[37]에서 알 수 있듯이, 추천 모델의 성능을 산출할 때 많이 사용된다.

Table 1. Information of features by personal and pathological characteristics

Variable			
Individual Characteristics	Independent Variable	Age	Age of patiets
		Sex	man, women
		ASA	ASA Grade (1~5)
		BMI	Body Mass Index
		DM_history	Diabetic History
		Pulmonary_disease	Pulmonary disease history
		Smoking_history	History of smoking
		Prior_Dx_cancer	History of other cancers except colorectal cancer
Pathological Characteristics	Independent Variable	pT	Stage according to tumor size
		pN	Stage according to lymph node metastasis
		pM	Stage according to distant metastasis
		pTNM	Stage according to cancer progression
		LVI	Lymphatic vessel, Vascular vessel, Lymphatic/Vascular infiltration
		PNI	Infiltration around nerves
		Harvested_LN	Number of lymph nodes
		Positive_LN	Number of metastatic lymph nodes
		K-ras	not assessed, wild type, mutated
		N-ras	not assessed, wild type, mutated
		BRAF	not assessed, wild type, mutated
	Dependent Variable	Chemotherapy Regimen	Surveillance, 5FU/LV, capecitabine, FOLFOX, XELOX, FOLFIRI

Table 2. Parameters of kNN, SVM, DT, RF, LGBM

	kNN	SVM	DT	RF	LGBM
Parameters	n_neighbors:5 weights: uniform p : 2 metric: minkowski	penalty : l2 loss: squared_hinge dual : True tol : 1e-4 C : 1.0 multi_class : ovr fit_intercept: True intercept_ scaling:1 class_weight: None verbose : 0 random_state:None max_iter : 1000	criterion : gini splitter : best min_samples_split:2 min_samples_leaf:1 random_state : 0	n_estimators :10 criterion : gini min_samples_split:2 min_samples_ leaf:1 max_features:auto bootstrap : True verbose : 0	objective: multiclass boosting_type:goss learning_rate:0.014 max_depth : 100 lambda_l2 : 5 num_iteration:200 early_stopping_rou nds:100 min_data_in_leaf:2 eval_metric: multi_logloss

정확도 외에 모델의 성능을 평가할 수 있는 ROC 곡선과 AUC를 이용하여 모델의 예측 정확도를 평가하였다. ROC 곡선은 분류 모델의 성능을 보여주는 그래프로, x 축은 허위 양성 비율(False Positive Rate: False를 True라고 예측한 비율), y축은 참 양성 비율(True Positive Rate: True를 True라고 예측한 비율)이다. ROC 곡선의 참 양성 비율(y축)이 1.0에 가까울수록 좋은 모형이라고 할 수 있다[38]. AUC는 ROC 곡선의 하단 면적을 의미하며, 최소값은 0.5이고 성능이 좋은 모형일수록 1.0에 가깝다[39].

비교한 머신러닝 알고리즘 중 LGBM의 내장기능인 변수 중요도(Feature importance)를 이용하여 어떤 특성이 중요한지 알아보았다. LGBM의 변수 중요도는 데이터를 split할 때, 변수가 사용된 횟수가 포함된다. gain인 경우, 해당 변수의 정보를 잘 분류한 만큼 total gain이 포함된다[40]. 여기서, gain이란 새로운 변수를 기준으로 데이터를 분류할 때, 종속 변수에 대해 얼마나 설명할 수 있는지를 측정하는 기준이다.

2.6 Chemotherapy classification model parameters

본 연구에서 사용된 머신러닝 알고리즘의 하이퍼 파라미터는 Table 2와 같다. kNN, SVM은 기본 파라미터를 사용하였고, DT, RF, LGBM은 하이퍼 파라미터를 미세 조정하였다. DT는 random_state를 default None인 상태에서 0으로 바꾸었다. RF는 n_estimators를 default가 100인 상태에서 10으로 바꾸어 학습했다. LGBM은 objective, boosting, num_iteration, learning_rate 등을 변경하였다.

IV. Results

1. Accuracy of chemotherapy classification model for colorectal cancer patients

Table 3은 기저 모델과 병리학적 모델 그리고 결합 모델에 kNN, SVM, DT, RF, LGBM을 적용하여 Top-2 Accuracy를 사용하여 예측 정확도를 산출한 결과표이다. 기저 모델의 예측 정확도는 모두 75% 이하로 병리학적 변수를 포함한 모델에 비하여 약 10%가량 낮은 예측 정확도를 보여주었다.

결합 모델의 정확도를 살펴본 결과, LGBM에서 85.42%로 가장 높았다. 그다음으로 SVM의 정확도가 81.20%, RF가 80.89%, kNN이 77.21%, DT가 56.59%로 나왔다.

Fig 3은 결합 모델의 ROC 곡선을 그래프로 나타낸 그림이다. 그래프를 살펴본 결과, LGBM이 참 양성 비율(y축)의 1.0에 가장 가깝고, 그 다음으로는 RF, SVM, kNN, DT 순으로 가까운 것을 확인할 수 있었다. Table 4에서 보는 바와 같이 AUC 또한 LGBM이 0.8586으로 가장 크고, RF 0.7819, SVM 0.7381, kNN 0.7275, DT 0.6259 순으로 결과를 확인할 수 있었다.

Table 3. Top-2 Accuracy(%) of kNN, SVM, LGBM, DT, RF

	Baseline model	Pathological model	Combined model
kNN	65.87	76.57	77.21
SVM	73.54	81.75	81.20
DT	39.52	55.08	56.59
RF	68.25	77.43	80.89
LGBM	75.05	83.90	85.42

Table 4. AUC of kNN, SVM, LGBM, DT, RF

	kNN	SVM	DT	RF	LGBM
AUC	0.7275	0.7381	0.6259	0.7819	0.8586

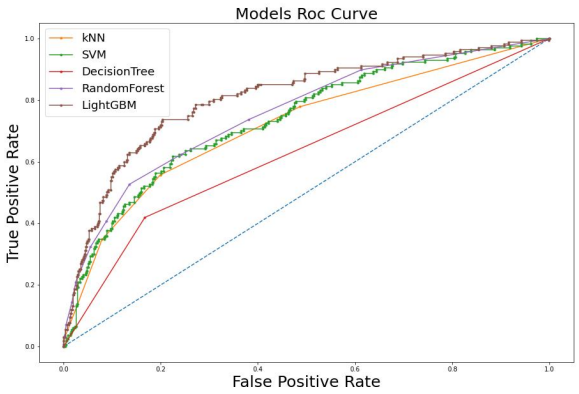

Fig. 3. ROC Curve of combined model

2. Feature importance of chemotherapy classification model for colorectal cancer patients

LGBM에서 변수 중요도 도출 결과, FIg 4에서 보는 바와 같이 획득한 림프절 개수(Harvested_LN)가 가장 높고, 그다음으로는 나이, 조직학적 병기(pTNM) 순으로 변수 중요도가 나타났다.

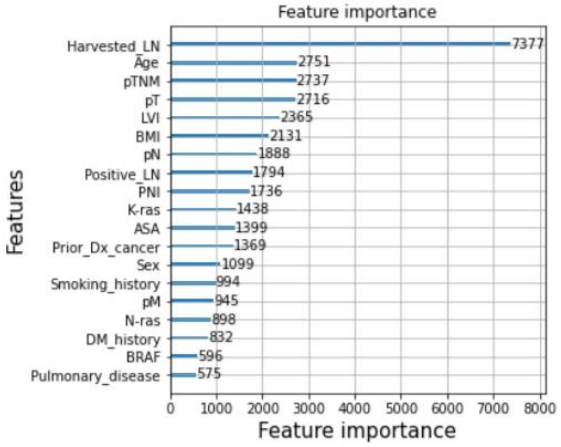

Fig. 4. Feature Importance of Colorectal Cancer in LGBM

V. Discussion

본 연구에서는 항암화학요법 분류 모델 구축을 위해 머신러닝 알고리즘을 사용하여 대장암 환자의 개인적, 병리학적 특성에 따른 항암화학요법 예측 정확도를 산출하고 비교하였다.

Top-2 Accuracy를 통해 예측 정확도 비교 결과, LGBM이 85.42%로 다른 모델들보다 예측 정확도가 높은 것을 확인할 수 있었다. 이를 보면 항암화학요법을 분류하는 모델을 사용하여 CDSS를 구축하는 데에도 정확도가 높은 알고리즘인 LGBM을 사용하는 것이 도움이 될 것이다.

ROC 곡선과 AUC 값을 비교한 결과, 예측 정확도가 가장 높았던 LGBM에서 가장 우수한 성능을 보였다. 정확도 비교 결과와는 다르게 RF, SVM, kNN, DT 순으로 ROC 곡선의 모형이 우수하며, AUC 값도 위 순서대로 크기가 산출되었다. 정확도와 다르게 ROC 곡선과 AUC 값 평가에서는 SVM이 RF보다 성능이 낮다고 평가되었다. 이는 SVM이 항암화학요법 예측률은 높았으나, 오답률도 같이 높은 등의 경과를 보였기 때문에 나온 결과라고 생각된다.

본 연구에서는 예측 정확도에서 가장 좋은 성능을 보인 LGBM의 변수 중요도를 알아보았다. LGBM으로 변수 중요도를 도출한 결과, 획득한 림프절 개수, 나이, 조직학적 병기 순으로 중요도가 높은 상위 3개 변수임을 확인할 수 있었다. 이는 병리학적 모델과 결합 모델의 정확도 차이를 설명할 수 있다. 비교적 변수 중요도가 높은 개인적 특성인 나이를 추가하였기 때문에 병리학적 모델보다 결합 모델에서 더 좋은 정확도를 얻을 수 있었다.

LGBM 모델의 정확도는 85.43%로 선행 연구에 비해 약간 낮은 수치이다[37]. 이는 Target 변수인 항암화학요법 종류의 데이터 불균형이 원인으로 사료된다. 항암화학요법 중 5-FU/LV, FOLFOX에서는 케이스 수가 각각 약 600명이고, 항암화학요법을 시행하지 않은 환자는 약 1300명이었다. 하지만 Capecitabine, XELOX의 경우에는 약 100명으로 비교적 케이스 수가 적었다. 이와 같은 데이터 불균형 문제는 동등한 모델 학습이 되지 않게 한다. 따라서 추후 연구에서는 데이터 불균형을 해결하고 학습을 한다면, 분류 성능이 향상될 것으로 보인다.

향후 연구에서는 추가적인 전처리를 통해 변수들을 추가하고 세부적인 파라미터를 조정하여 모델을 정교하게 튜닝하여 더 좋은 성능을 보이는 모델을 구축하는데 초점을 맞출 것이다.

VI. Conclusion

본 연구에서는 가천대학교 길병원 EMR의 대장암 환자 데이터를 이용하여 대장암 항암화학요법 추천에 있어 어떤 모델이 적합한지 분석하였다. 본 연구에서는 데이터를 전처리한 후 머신러닝 알고리즘을 사용하여 특성별 모델의 분류 결과를 비교하였다. 비교한 모델로는 기저 모델, 병리학적 모델, 결합 모델이 있었고, 머신러닝 알고리즘은 kNN, SVM, DT, RF, LGBM이 사용되었다. 예측 정확도는 Top-n accuracy와 ROC 곡선, AUC 지표를 이용하여 비교하였다. 연구 결과, 본 연구에서 제시한 결합 모델에서 우수한 성능을 보였고, 머신러닝 알고리즘 중에서 LGBM의 예측 정확도가 좋은 것으로 나타났다. 이는 CDSS에서 임상 결정을 내릴 때 적용되는 보조 도구로써 결합 모델을 사용한다면 좋은 예측 결과를 얻을 수 있을 것으로 판단된다.

REFERENCES

[1] World Health Organization, The global cancer observatory in Republic of Korea [Internet], http://gco.iarc.fr/today/data/factsheets/populations/410-korea-republic-of-fact-sheets.pdf

[2] World Health Organization. The global cancer observatory: Colorectal cancer [Internet], http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf

[3] Eun Ok Lee, Aeyong Eom, Rhayun Song, young Ran Chae, and Paul Lam, "Factors influencing quality of life in patients with gastrointestinal neoplasms", J Korean Acad Nurs Vol. 38, No. 5, pp. 649-655, September 2008. DOI: https://doi.org/10.4040/jkan.2008.38.5.649

[4] Kim JH, Choi KS, Kim TW, Hong YS, "Quality of life in colorectal cancer patients with chemotherapy-induced peripheral neurophthy", Journal of Korean Oncology Nursing, Vol. 11, No. 3, pp. 254-262, November 2011. DOI: https://doi.org/10.5388/jkon.2011.11.3.254

[5] Nam Kyu Kim, Jae Kun Park, Kang Young Lee, Seong Hyeon Yun, Seung Kook Sohn, Jin Sik Min, "Prognostic Factors Influencing the Recurrence Pattern and Survival Rates in Curatively Resected Colorectal Cancer", Annals of Surgical Treatment and Research, Vol. 62, No. 5, pp. 421-429, April 2002.

[6] No Kyung Kim, Guide to Cancer Management Patient Care and Palliative Treatment, ilchokak, 2005.

[7] Kim Mi-Sook, "The Change of Physical Function in Accordance with Rehabilitation Exercise Frequency for the Breast Cancer Survivors", The Korean Journal of physical education humanities and social science, Vol. 49, No. 4, pp. 315-323, July 2010.

[8] Yang-Sook Kim, Mi-Sook Kim, "The Change of Functional Fitness and Bone Mineral Density on a Long-Term Combined Exercise Intervention in Breast Cancer Survivors", Jounal of Life Science, Vol. 18, No. 7, pp. 968-973, July 2008.

[9] Won Ho Kim, Jae Hui Chun, Colorectal Cancer Guidebook, Kugil media, 2007.

[10] Yu Young Kim, Internal medicine guidelines, Korea Medical Book Publishing Company, 2002.

[11] Kim Eun Jung. "Hospitalization-related Stress and Coping Strategies of Colon Cancer Patients", Clinical Health Science The Graduate School of Ewha Womans University, 2007.

[12] Hwa Jung Kim, “Adjuvant chemotherapy timing after curative surgery in stage III colorectal cancer patients”, Seoul National University College of Medicine, August 2013.

[13] The Korean Society of Coloproctology, Chemotherapy of colon cancer. Medical Culture Publishing Company, 2004.

[14] National Cancer Information Center: Colorectal Cancer chemotherapy, https://www.cancer.go.kr/

[15] Kabbinavar FF, Schulz J, McCleod M, et al., “Addition of bevacizumab to bolus fluorouracil and leucovorin in first-line metastatic colorectal cancer: results of a randomized phase II trial”, J Clin Oncol, Vol 23, No. 16, pp. 3697-705, Jun 2005.

[16] Lévi F, Misset JL, Brienza S, et al., "A chronopharmacologic phase II clinical trial with 5-fluorouracil, folinic acid, and oxaliplatin using an ambulatory multichannel programmable pump. High antitumor effectiveness against metastatic colorectal cancer", Cancer, Vol. 69, pp. 893-900, February 1992.

[17] Madajewicz S, Petrelli N, Rustum YM, et al., "Phase I-II trial of high-dose calcium leucovorin and 5-fluorouracil in advanced colorectal cancer", Cancer Res, Vol 44, pp. 4667-4669, 1984.

[18] Seung-Taek Oh, Edwin R.Fisher, et al., "Clinical use of targeted therapies in colorectal cancer", Korean Journal of Clinical Oncology, Vol. 1, No. 2, pp. 54-58, December 2005.

[19] Youn-Tae Lee, Young-Taek Park, Jae-Sung Park, Byoung-Kee Yi. "Association between Electronic Medical Record System Adoption and Healthcare Information Technology Infastructure", Healthcare Informatics Research Vol. 24, No. 4, pp. 327-334, Oct 2018.

[20] D.H. Lee, H.Y. Jung, M.H. Kim, D.H. Kim, Y.W. Han, Y.W. Lee, J.H. Choi, S.H. Kim, "Trends of Clinical Decision Support System(CDSS)", Electronics and Telecommunications Trends, Vol. 31, No. 4, pp.77~85, August 2016.

[21] IBM Watson Hard At Work: New Breakthroughs Transform Quality Care for Patients [Internet] IBM News releases [cited 2021 Jan 27], http://www-03.ibm.com/press/us/en/pressrelease/40335.wss?i=1360645029661

[22] Koichi T, Kantarjian HM, Garcia-Manero G, Stevens RJ, Dinardo CD, Allen J, et al., “MD Anderson's Oncology Expert Advisor powered by IBM Watson: a web-based cognitive clinical decision support tool”, 2014 ASCO Annual Meeting, 2014 May 30-Jun 3

[23] Lee WS, Ahn SM, Chung JW, Kim KO, Kwon KA, Kim Y, et al., “Assessing concordance with Watson for oncology, a cognitive computing decision support system for Colon Cancer treatment in Korea”, JCO Clin Cancer Informatics, Vol. 2, pp. 1-8, 2018.

[24] GU Huiyan, DAI Limin, WU Gang, XU Dong, WANG Shunzhong, and WANG Hui "Estimation of forest volumes by integrating Landsat TM imagery and forest inventory data", Science in China Series E: Technological Sciences, Vol.49, No.1, pp. 54-62, June 2006.

[25] Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., “Efficient knn classification with different numbers of nearest neighbors,” IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 5, pp. 1774-1785, May 2018.

[26] Zaklouta, Fatin, and Bogdan Stanciulescu. "Real-time traffic-sign recognition using tree classifiers." Intelligent Transportation Systems, IEEE Transactions, Vol 13, No. 4, pp.1507-1514, Dec 2012.

[27] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, IEEE Computer Society Conference, Vol 1, July 2005.

[28] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J, “Classification and Regression Trees", Wadsworth International Group, 1984

[29] Lee, M., Y. Kim, Y. Jun, and Y. Shin, "Random forest based prediction of road surface condigion using spatio-temporal features", Journal of Korean Society of Transportation, Vol 37, No.4, pp. 338-349, August 2019.

[30] Han, D., Y. J. Kim, J. Im, S. Lee, Y. Lee, and H. Kim, "The estimation of arctic air temperature in summer based on Machine Learning approaches using IABP Buoy and AMSR2 satellite data", Korean Journal of Remote Sensing, Vol 34, No. 6-2, pp.1261-1272, Dec 2018.

[31] Aurelien Geron, “Hands-On Machine Learning with Scikit-Learn & TensorFlow”, Hanbit Media, 2018.

[32] Alexey Natekin, Alois Knoll, "Gradient boosting machines, a tutorial", Front Neurorobot., 2013

[33] Kim Seon Uk, Bang Jun Il, Hong Seong Eun, Kim Hwa Jong, "A Study on The Prediction of Missing Value in Data of Air Pollution Using LightGBM", Korea Institute Of Communication Sciences Conference, pp. 1029-1030, 2019.

[34] Brownlee, J, “Why one-hot encode data in machine learning, Machine Learning Mastery”, 2017.

[35] Scikit-learn machine learning in python, https://scikit-learn.org/stable/index.html

[36] Harald Steck, “Evaluation of recommendations: rating-prediction and ranking”, Association for Computing Machinery, pp. 213-220, 2013.

[37] Jin-Hyeok Park, Jeong-Heum Baek, Sun Jin Sym, Kang Yoon Lee, Youngho Lee, “A data-driven approach to a chemotherapy recommendation model based on deep learning for patients with colorectal cancer in Korea”, BMC Medical Informatics and Decision Making, Vol 241, 2020.

[38] Tom Fawcett, “An introduction to ROC analysis”, Pattern Recognition Letters, Vol 27, pp. 861-874, 2006.

[39] “A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis”, Intensive Care Medicine Vol 29, pp. 1043-1051, 2003

[40] LightGBM, Release 3.1.1.99, Microsoft Corporation, 2021.

Authors

Seo-Hyun Oh is a student of computer engineering at Gachon University. She is currently conducting research on De-identification of patients' Personal Health Informations using machine learning.

Her research interests include machine learning, deep learning, and data analysis.

Jeong-Heum Baek received Ph.D. degree in HanYang University in 2000. He is currently a Professor in Department of Surgery at Gachon University and clinician in Division of Colon and Rectal Surgery at Gil Medical Center. His primary research interests include Clinical Decision Support System, Healthcare Information, translational research and surgery for colorectal cancer.

Un-Gu Kang received Ph.D. degree in Computation Engineering from Inha University in 2001. He is currently a Professor in Department of Computer Engineering at Gachon University.

His primary research interests include Mobile Software, Healthcare Information, U-healthcare.