Journal of the Korea Industrial Information Systems Research Vol.28 No.2, Apr. 2023 :1-9
http://dx.doi.org/10.9723/jksiis.2023.28.2.001 ISSN:1229-3741

건강보험 청구 데이터를 활용한 머신러닝 기반 유방암 환자의 생존 여부 예측

(The Prediction of Survival of Breast Cancer Patients Based on Machine Learning Using Health Insurance Claim Data)

이 덕 규[1)*], 변 경 근[2)], 이 형 동[2)], 신 선 희[3)]
(Doeggyu Lee, Kyungkeun Byun, Hyungdong Lee, and Sunhee Shin)

요 약 유방암 관련 기존 AI 연구는 보조적인 진단 예측이나 임상적 요인에 따른 진료 결과를 예측하는 주제가 많았다. 또한 연구기관의 코호트 자료나 일부 환자 자료를 이용하는 경우가 대부분이었다. 본 논문에서는 건강보험심사평가원이 보유하고 있는 전 국민 유방암 환자의 전수 데이터를 활용하여 유방암 환자의 40~50대와 다른 연령대 간의 생존 여부 예측과 생존 여부에 미치는 요인의 차이점을 분석했다. 그 결과, 환자들의 생존 여부 예측 정밀도는 40~50대가 평균 0.93으로 60~80대 0.86 보다 높았으며, 요인에 있어서도 40~50대는 치료횟수(46%)가, 60~80대는 나이(32%)의 변수 중요도가 제일 높았다. 기존 연구와 성능 비교 결과, 평균 정밀도가 0.90으로 기존 논문의 정밀도 0.81보다 높았다. 적용 알고리즘별 성능 비교 결과, 의사결정나무(Decision Tree), 랜덤포레스트(Random Forest) 및 그래디언트부스팅(Gradient Boosting)의 전체 평균 정밀도는 0.90, 재현율은 1.0으로 연령대 그룹 내에서 동일하였으며, 다층퍼셉트론(Multi-Layer Perceptron)의 정밀도는 0.89, 재현율은 1.0 이었다. 심평원의 전 국민 심사청구 빅데이터 가치 활용을 제고하기 위해 비전문가용 머신러닝 자동화(Auto ML) 도구를 사용한 더 많은 연구가 진행되기를 바란다.

핵심주제어: 머신러닝, 유방암, 생존여부 예측, 건강보험심사평가원, 심사청구자료, 의사결정나무

Abstract Research using AI and big data is also being actively conducted in the health and medical fields such as disease diagnosis and treatment. Most of the existing research data used cohort data from research institutes or some patient data. In this paper, the difference in the prediction rate of survival and the factors affecting survival between breast cancer patients in their 40~50s and other age groups was revealed using health insurance review claim data held by the HIRA. As a result, the accuracy of predicting patients' survival was 0.93 on average in their 40~50s, higher than 0.86 in their 60~80s. In terms of that factor, the number of treatments was high for those in their 40~50s, and age was high for those in their 60~80s. Performance comparison with previous studies, the average precision was 0.90, which was higher than 0.81 of the existing paper. As a result of performance comparison by applied algorithm, the overall average precision of Decision Tree, Random Forest, and Gradient Boosting was 0.90, and the recall was 1.0, and the precision of multi-layer perceptrons was 0.89, and the recall was 1.0. I hope that more research will be conducted using machine learning automation(Auto ML) tools for non-professionals to enhance the use of the value for health insurance review claim data held by the HIRA.

Keywords: Machine Learning, Breast Cancer, Prediction of survival, Health Insurance Review Assment, Review Claims Data, Decision Tree

* Corresponding Author: leedg317@naver.com
Manuscript received February 12, 2023 / revised March 07, 2023 / accepted March 13, 2023

1) 숭실대학교 IT정책경영학과, 교신저자
2) 숭실대학교 IT정책경영학과
3) 강남대학교 교육학과

1. 서 론

4차 산업혁명의 발전과 코로나19 팬데믹에 의한 비대면 문화 활성화로 산업 생태계에 인공지능(AI)과 빅데이터를 이용한 데이터 분석이 많아지고 있다. 보건의료 분야에서도 질병 진단, 치료 및 약제 개발 등의 분야에서 AI와 빅데이터를 이용한 연구가 활발히 진행되고 있다. 유방암 관련 기존 AI 연구는 Choi et al.(2021), Lee(2020) 및 Yun et al.(2022) 등과 같이 유방암 보조적인 진단 예측이나 임상적 요인에 따른 진료 결과를 예측하는 주제가 많았다. 또한 연구 데이터는 특정 병원 환자의 데이터나 연구기관의 코호트 데이터를 이용하였다.

의료기관에서 건강보험심사평가원(이하 심평원 또는 HIRA)으로 청구한 심사청구 데이터를 활용한 선행 연구로는 Byun et al.(2022)이 사망률이 높은 80대 및 90대 노령자 대상 폐암 진단 후 84개월간의 사망률을 예측하고, 알고리즘별 성능을 비교하였다. 본 논문에서는 심평원 심사청구 데이터를 활용하여 국내 유방암 환자 전수 데이터를 전체 연령대, 40~50대, 60~80대 연령 그룹별로 생존 여부 예측 모델을 생성하고, 그룹별로 생존 여부에 영향을 미치는 요인의 차이점을 분석하였다. 알고리즘은 Kim et al.(2022)에 제시된 머시러닝 자동화(Auto ML) 도구인 와이즈프로핏에서 제공하는 의사결정나무(Decision Tree, 이하 DT), 랜덤포레스트(Random Forest, 이하 RF), 그래디언트부스팅(Gradient Boosting, 이하 GB)과 사회과학용 통계 패키지(Statistical Package for the Social Science, 이하 SPSS)의 분석 도구인 다층퍼셉트론(Multi-Layer Perceptron, 이하 MLP), 방사형기저함수(Radial Basis Function, 이하 RBF)를 이용하였다. 논문의 구성은 관련 선행 연구를 고찰하고, 연구대상 자료, 전처리 과정 및 모델 생성에 대해 설명하였다. 이어서 모델 성능 분석 및 선행 연구 결과를 이용한 성능 비교 평가를 하고, 연구 결과의 의의와 향후 과제를 제시한다.

2. 관련 연구

2.1 국내 암환자 현황

2021년 공표한 '암 등록 통계'에 따르면 2019년 기준 암 유병자[4]는 약 214만 명으로 남자 94만 명, 여자 120만 명으로 여자 암 유병자가 남자 대비 22% 더 많다. 암종별 유병자는 남자 위암 22만 명, 대장암 17만 명, 전립선암 11만 명 순이며, 여자는 갑상선암 38만 명, 유방암 26만 명, 대장암 11만 명 순이었다. 위암, 간암 및 폐암 발생률은 최근 10여 년간 감소 추세를 보이고 있으나 유방암의 발생률은 1999년 12.8(인구 10만 명당 기준)에서 2018년 32.9로 20년간 증가하는 추세이다. 암종별 연령별 암 발생에 있어서도 1위는 유방암 40대 7,717 명, 2위 유방암 50대 7,449 명, 다음으로 갑상선암 40대가 6,189 명으로 나타났다(The Yakup, 2022). 모든 암 사망률은 1997년 124.2(인구 10만 명당 기준)에서 2019년 74.2로 꾸준히 감소하고 있으나 유방암만 1997년 2.3에서 2019년 3.1로 지속적으로 증가하고 있다(National Cancer Center. 2021. *Cancer Monitoring Indicator*). 한 해 유방암으로 진단받는 환자 수가 2만 명에 달하며, 여성에게는 매우 흔한 암종 중 하나이다. 따라서 유방암은 여성에게 다른 암종보다 관리가 필요한 암종이다.

2.2 선행 연구

머신러닝 또는 딥러닝 기술을 이용한 유방암 보조적인 진단 예측이나 임상적 요인에 따른 진료 결과 예측 관련 연구는 Choi et al.(2021), Lee(2020) 및 Yun et al.(2022) 등 국·내외 저널에 다수 등재되어 있다. Choi et al.(2021)는 X선 이미지를 읽어 유방암 검진 보조로 유방암의 악성과 양성을 판별하기 위하여 합성곱 신경망인 Fully Convolutional Network(FCN) 모델을 활용하여 유방조영술 촬영본의 정확도를 향상시켰다. Lee(2020)는 병리 이미지를 이용한 딥러닝 기반의 삼중음성 유방암 환자의 예후 및 예측 분석 모델을 개발하였다. Yun et al.(2022)은 메

4) '99년 확진 후 '20.1.1. 기준, 치료 중이거나 완치된 사람

타브릭스 시험 참가자로부터 수집된 유방 종양에 관련된 데이터인 메타브릭스 데이터 세트를 이용하여 유방암 생존 예측에서 임상적 특성뿐 아니라 유전적 특성을 함께 고려하는 것이 중요하다는 것을 실험적으로 보였다. Adam et al.(2019)은 전통적 임상 유방암 위험 모델보다 더 정확한 딥러닝 기반 유방조영술 유방암 위험 모델을 개발하였다. Ayelet et al.(2019)은 건강기록 및 유방조영술과 연계한 딥러닝을 적용하여 초기 유방암 탐지의 정확성과 효율성을 평가하였다. Keping et al.(2021)은 딥러닝으로 멀리 떨어진 지역의 유방암 환자에 대한 원격 온라인 진단 정확도를 향상시킨 사례도 있었다. 기타 암종에 대한 딥러닝 선행 연구를 알아보면, Ryu et al.(2017)는 DT 알고리즘을 기반으로 국민건강보험공단의 노인 코호트 DB(ver 1.0)를 활용하여 고령자의 뇌졸중 질환 예측 방법론의 성능을 검증하였다. Kang et al.(2022)은 머신러닝을 이용해 고혈압 발병에 영향을 미치는 요인들의 생애주기별로 차이를 분석하였다.

심평원 심사청구 데이터를 활용한 선행 연구로는 서론에서 언급한 Byun et al.(2022)이 와이즈프로핏에서 제공하는 Auto ML인 DT 등 5개 알고리즘을 적용하여 사망률을 예측하고 성능을 비교하였다. 위에서 알아본 바와 같이 Byun et al.(2022)을 제외한 대부분 선행 논문은 코호트 자료 또는 시험 참가자를 대상으로 유방암 진단의 정확도를 높이거나 진단 단계에서 생존율 등을 예측한 논문들이었다. 본 논문에서는 의료기관에서 진료 후 심평원에 심사청구한 전 국민 유방암 전수 데이터를 활용해서 유방암 환자의 생존 여부 예측 및 그에 미치는 요인을 분석하였다.

3. 연구모형 및 방법

3.1 연구 모형

본 논문에서는 연구용 데이터셋으로 국내 유방암 환자 심사청구 데이터와 와이즈프로핏과 SPSS 분석 도구를 활용하였다. 훈련용 데이터는 연구용 데이터셋의 80%로, 검증용 데이터는 20%로 'Fig. 1'과 같이 연구 모델을 설계하였다. 모델 생성 후 유방암 환자의 생존 여부를 예측하고, 생존 여부에 영향을 미치는 요인을 분석하였다.

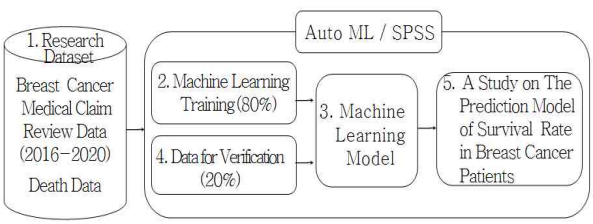

Fig. 1 A Research Model

3.2 연구대상 자료

연구대상 자료는 의료기관에서 심평원으로 심사청구 한 자료 중 2016년부터 2020년 12월까지 심사 결정한 자료이며 상병코드가 유방암(C50)인 데이터이다. 이 데이터는 환자식별번호(H-pin)[5], 상병명, 최초 진료개시일, 수술여부, 당월요양개시일, 최초입원개시일, 입원일수, 입원·외래구분 등 요양급여비용명세서[6]의 항목과 진료비 심사를 위해 행정안전부에서 제공받은 사망자 정보를 포함하고 있다. 5개 년도 요양급여비용명세서 건수는 5,870,274건이며, 연도별 건수는 'Table 1'과 같다. 환자식별번호로 병합(merge)한 결과 148,133명으로 통합되었다. 통합한 유방암 환자와 행정안전부 사망자 일치 건은 14,403명이었다.

3.3 전처리 작업

전처리 작업은 연도별 유방암 요양급여비용명세서를 환자식별번호로 통합한 후 'Table 2'와 같이 연구용 데이터셋 변수로 나이, 주상병코드, 치료기간, 진료횟수, 수술코드, 사망코드로 전환하였다. 전환처리 과정에서 변수들의 결측치 값

5) H-pin : 심평원에서 개인고유식별자 유·노출를 방지하기 위해 이용하는 주민등록번호 13자리를 의미 없는 다른 숫자로 변환·저장하는 일종의 암호화 대체키

6) 요양급여비용명세서 : 의료기관에서 심평원으로 보험 비용 청구를 위하여 환자단위로 1주 또는 월 단위로 진료내역을 기재하여 청구하는 자료

은 발생되지 않았다. 이는 심평원에서 요양급여 비용명세서 항목을 심사하여 의료기관에 진료 비용을 지급 결정하기 때문에 항목의 값이 없거나

Table 1 Number of Medical Claim

Year	Breast Cancer Statement Count	Number of patients	The death toll
2016	1,292,690	148,133	14,403
2017	1,225,413		
2018	1,192,427		
2019	901,539		
2020	1,258,205		

사전에 정의한 범위 외의 자료인 경우 의료기관으로 반송하는 제도가 있기 때문이다. 또한 이상치(outlier) 값은 아니지만 본 논문의 주제가 국내 여성들의 유방암에 관한 내용이기 때문에 환자의 성별 코드가 남자인 명세서와 외국인 명세서는 연구용 데이터셋 구축 시 제외하였다. 제외된 명세서 건수는 2020년도 자료 기준으로 남자는 4,309건, 외국인은 17,327건 이었다.

Table 2 Research Datasets Variable

Variable name	Variable Code	Variable Description
Age	1(Under 19), 2(20s), 3(30s) 4(40s), 5(50s), 6(60s), 7(70s), 8(80s), 9(90 years of age)	(Date of death or Final care) - (Date of birth)
Main Disease	C500, C501, C502, C503, C504, C505, C506, C508, C509	Table 3 Ref.
Treatment Period	1(0~12 Months), 2(13~24 Months), 3(25~36 Months), 4(37~48 Months), 5(Over 49 Months)	(Last date of treatment) - (Old medical date)
Number of Treatments	1(0~9), 2(10~29), 3(30~49), 4(50~69), 5(70~89), 6(90~99), 7(More than 100 times)	Cumulative number of Inpatietient or outpatient
Surgery Code	0(Unsurgical), 9(Surgery)	Surgery status
Death Code	0(Survival), 1(Death)	Death status

나이·치료기간·치료횟수는 'Table 2'의 변수 설명에 기술한 바와 같이 요양급여비용명세서의 환자생년월, 진료일자, 입원·외래구분 코드 등을 이용하여 변수를 재생성하였다. 치료횟수는 입원, 외래구분 없이 1건으로 누적한 횟수이다. 주상병 및 사망코드는 심평원 요양급여비용명세서 항목 자료를 변경 없이 사용하였다.

Table 3 Description of Main Disease Code

Code	Description
C500	Malignant neoplasm of nipple and areola
C501	Malignant neoplasm of central portion of breast
C502	Malignant neoplasm of upper-inner quadrant of breast
C503	Malignant neoplasm of lower-inner quadrant of breast
C504	Malignant neoplasm of upper-outer quadrant of breast
C505	Malignant neoplasm of lower-outer quadrant of breast
C506	Malignant neoplasm of axillary tail of breast
C508	Malignant neoplasm of overlapping lesion of breast
C509	Malignant neoplasm of breast unspecified

3.4 연구방법

2.1절 국내 암환자 현황에서 알아본 바와 같이 40대 및 50대 유방암 발생률이 국내 암종별 연령별 전체 암 발생율에 있어서 1위·2위를 차지하였다. 따라서 본 논문에서는 40~50대 유방암 환자들이 다른 연령 그룹간의 생존 여부 예측이나 변수 중요도의 차이점을 찾는 것이었다. 따라서 연구용 데이터셋을 세 그룹으로 나누어 모델 생성 후 상호 비교 분석을 하였다. 첫째는 유방암 환자의 전체 연령 그룹, 둘째는 40~50대 중년 그룹, 셋째는 60~80대까지 노년 그룹으로 나누었다. 노년 그룹에서 90대 이상은 연구대상 건수가 미비하여 제외하였다. 세 그룹 각각의 유방암 환자 생존 여부 예측을 위한 모델 생성 알고리즘은 DT, RF, GB 및 SPSS의 MLP, RBF 알고리즘을 활용하였다. 알고리즘별로 입력된 파라미터 값은 'Table 4'와 같다. 알고리즘별 파라미터의 기능은 Kim et al.(2022)에 따르면 다음과 같다. criterion은 분활 품질을 설정하는 기능으로서 트리 분리 기준으로 gini가 설정되어 있다. gini는 불순도 집합에 이질적인 것이

얼마나 섞여 있는가를 나타내는 척도이다. max_depth는 얼마나 깊게 트리를 만들 것인가의 기준으로 값이 클수록 모델이 복잡해진다. min_samples_leaf은 노드를 구성하는 최소한의 샘플 수이며, n_estimators는 트리의 수이다. learning_rate는 학습을 진행할 때마다 적용하는 학습률로 기본값이 0.1로, 순차적으로 오류값을 보정해 나가는데 적용하는 계수로 0~1 사이의 값을 가진다. subsample은 각 트리마다 데이터의

Table 4 Algorithm Description

Classify	Parameters(value)
DT	△criterion(gini) △max_depth(4) △min_samples_leaf(1)
RF	△criterion(gini) △max_depth(5) △min_samples_leaf(1) △n_estimators(10)
GB	△learning_rate(0.1) △max_depth(3) △subsample(1.0)
MLP	△hidden layers(1) units(7) △Activation function(hyperbolic tangent) △Resulting layer activation function(softmax) △Error function(cross entropy)
RBF	△hidden layer units(10), △Activation function(softmax) △Resulting layer activation function(equivalent function) △Error function(sum of squares function)

샘플링 비율로 범위는 0~1 사이의 실수값을 가진다. MLP와 RBF의 신경망 설계 옵션은 은닉층 단계에서 은닉층 수, 은닉층에서 노드의 수, 활성화 함수, 출력층 단계에서 활성화 함수 및 오차함수를 정의하였다. criterion 값은 entropy로, max_depth 값은 DT(5), RF(6), GB(4)로 변경하여 실행하였으며, 은닉층 수와 노드 수도 2와 8 등으로 변경하여 실행하였으나 유의미한 결과를 도출하지 못하여서 최종적으로 'Table 4'와 같이 설정하였다.

모델의 성능평가 지표는 'Table 5'와 같이 정밀도와 재현율을 사용하였다. 이는 연구용 데이터셋 및 연구 방법론이 유사한 Byun et al.(2022)의 연구 결과와 성능 비교 분석을 위해 동일한 성능 평가지표를 이용했다. 한편, 생존여부에 영향을 미치는 요인 분석은 와이즈프롯핏의 Feature Engineering인 Extra Tree 알고리즘 기법의 변수 중요도(이하 변수코드 중요도)와 SPSS에서 제공하는 독립변수 중요도(이하 독립변수 중요도)를 이용하였다.

4. 연구모델 성능 분석 결과

4.1 전체 연령대 분석 결과

전체 연령대(148,133명)의 5개 알고리즘 정밀도 예측 평균값은 'Table 6'과 같이 0.90 이었다. 재현율의 평균값은 1.0 이었다.

Table 5 Performance Evaluation Indicator(PEI)

A measured formula
△ Precision = $\frac{TP}{(TP+FP)}$ * Percentage of predicted values where actual values occur
△ Recall = $\frac{TP}{(TP+FN)}$ * Proportion of intercepts detected accurately by values classified in the model

Classify		Prediction	
		Positive	**N**egative
Actuality	**P**ositive	**T**rue **P**ositive(**TP**)	**F**alse **N**egative(**FN**)
	Negative	**F**alse **P**ositive(**FP**)	**T**rue **N**egative(**TN**)

정밀도는 DT, RF, GB, MLP 알고리즘이 동일하였고, 재현율은 DT, RF, GB, RBF 알고리즘이 동일하였다.

Table 6 The Whole Age Group PEI Results

Algorithms	Precision	Recall
DT	0.90	1.00
RF	0.90	1.00
GB	0.90	1.00
MLP	0.90	0.99
RBF	0.89	1.00
Average[7)]	0.90	1.00

생존 여부에 영향을 미치는 변수코드 중요도 분석 결과, 'Table 7'과 같이 80대에서 가장 큰 영향을 미쳤으며 그다음 치료횟수가 100회 이상인 경우, 나이가 90대, 70대, 60대 순이었으며, 치료기간은 12개월 이내 및 25~36개월, 수술한 경우 순이었다. 마지막으로 주상병 중에는 C501과 C504 순으로 확인되었다.

독립변수 중요도는 'Table 8'과 같이 나이가 가장 큰 영향을 미쳤으며 다음으로 치료횟수, 주상병 코드, 치료기간, 수술여부 순으로 나타났다.

Table 7 The Whole Age Group Analysis of Factors Affecting Survival Rate(DT)

Ranking	Variable Code	Importance
1	Age_8	0.37
2	Number of Treatments_7	0.17
3	Age_9	0.13
4	Age_7	0.09
5	Age_6	0.05
6	Treatment Period_1	0.04
7	Treatment Period_3	0.04
8	Surgery Code_9	0.03
9	Main Disease_C501	0.02
10	Main Disease_C504	0.01

Table 8 The Whole Age Group Independent Variable Importance(MLP)

Classify	Importance	Normalization Importance
Age	.40	100%
Number of Treatments	.32	80%
Main Disease	.14	34%
Treatment Period	.13	33%
Surgery Code	.02	5%

4.2 40~50대 분석 결과

40~50대(89,027명)의 알고리즘별 정밀도 예측 평균값은 'Table 9'와 같이 0.93이었으며 재현율의 평균값은 1.00이었다. 정밀도 예측 평균값은 4.1절에서 기술한 전체 연령대 0.90 보다 높았다. 정밀도는 DT, RF, GB 알고리즘이 0.93으로 동일했고, 재현율은 모든 알고리즘이 동일한 결과가 나왔다.

Table 9 40 to 59 years old Group PEI Results

Algorithms	Precision	Recall
DT	0.93	1.00
RF	0.93	1.00
GB	0.93	1.00
MLP	0.92	1.00
RBF	0.92	1.00
Average	0.93	1.00

변수코드 중요도 분석 결과는 40~50대 중년 그룹에서는 전체 연령대와 다르게 나이보다 치료횟수와 치료기간이 생존 여부에 미치는 중요 요인으로 나타났다. 'Table 10'과 같이 치료횟수가 100회 이상인 경우가 가장 큰 영향을 미쳤으며 다음으로 치료기간이 25~36개월, 치료횟수가 12개월 이내 등 순이었다.

Table 10 40 to 59 years old Group Analysis of Factors Affecting Survival Rate(DT)

Ranking	Variable Code	Importance
1	Number of Treatments_7	0.35
2	Treatment Period_3	0.19
3	Number of Treatments_1	0.09
4	Age_4	0.09
5	Surgery Code_9	0.07
6	Age_3	0.05
7	Main Disease_C504	0.04
8	Main Disease_C502	0.04
9	Surgery Code_0	0.04
10	Number of Treatments_6	0.02

주상병 중에는 C504 및 C502의 중요성이 유사하였다. SPSS의 독립변수 중요도 분석 결과는 'Table 11'과 같이 치료횟수가 가장 큰 영향

7) 'Table 6', 'Table 9', 'Table 12' 평균값(Average)은 소수점 이하 세 번째 자리에서 반올림함

을 미쳤으며 다음으로 치료기간, 주상병코드, 나이, 수술여부 순으로 변수의 중요도가 나타났다.

Table 11 40 to 59 years old Group Independent Variable Importance(MLP)

Classify	Importance	Normalization Importance
Number of Treatments	.46	100%
Treatment Period	.21	45%
Main Disease	.20	42%
Age	.10	22%
Surgery Code	.03	7%

4.3 60~80대 분석 결과

60~80대 그룹(50,749명)의 알고리즘별 정밀도 예측 평균값은 'Table 12'와 같이 0.86이었다. 재현율의 평균값은 1.00이었다. 정밀도는 모든 알고리즘이 0.86 이었고 재현율도 1.0으로 동일한 결과가 나왔다.

Table 12 60 to 89 years old Group PEI Results

Algorithms	Precision	Recall
DT	0.86	1.00
RF	0.86	1.00
GB	0.86	1.00
MLP	0.86	1.00
RBF	0.86	1.00
Average	0.86	1.00

변수코드 중요도 분석 결과는 'Table 13'과 같이 나이 80대가 가장 큰 요인이었으며, 그 다음 치료횟수가 100회 이상이었다. 독립변수별 중요도 분석 결과는 'Table 14'와 같이 나이가 가장 큰 영향을 미쳤으며 다음으로 치료횟수, 치료기간, 주상병코드, 수술여부 순이었다.

Table 13 60 to 89 years old Group Analysis of Factors Affecting Survival Rate(DT)

Ranking	Variable Code	Importance
1	Age_8	0.46
2	Number of Treatments_7	0.15
3	Age_6	0.09
4	Treatment Period_3	0.08
5	Treatment Period_1	0.06
6	Surgery Code_0	0.03
7	Main Disease_C504	0.02
8	Number of Treatments_5	0.02
9	Main Disease_C501	0.02
10	Number of Treatments_3	0.02

Table 14 60 to 89 years old Group Independent Variable Importance(MLP)

Classify	Importance	Normalization Importance
Age	.32	100%
Number of Treatments	.21	66%
Treatment Period	.17	55%
Main Disease	.16	51%
Surgery Code	.14	44%

4.4 기존 연구와의 성능 비교 분석

본 논문의 성능 결과를 최근에 발표된 80대 및 90대 폐암 환자를 대상으로 폐암 진단 후 사망률을 예측한 Byun et al.(2022)과 정밀도와 재현율 측면에서 비교하였다. 비교논문에서도 와이즈프로핏에서 제공하는 DT, RF, GB, XGBoost, Logistic Regression 5개 알고리즘을 사용하였으며, 변수는 치료횟수(Number of Treatments)만 제외하고 동일하였다. 'Table 15'와 같이 본 논문의 정밀도와 재현율이 0.90 및 1.00으로 비교 모델 보다 높았다. 이러한 차이는 본 논문에서 추가한 치료횟수의 변수 중요도가 40~50대 그룹에서 1위인 100%, 60~80대 그룹에서 2위인 66%를 차지했기 때문이라고 추정한다.

Table 15 PEI Comparision

Classify	Precision	Recall
This Paper[8)]	0.90	1.00
Byun et al.(2022)	0.81	0.94

5. 결 론

본 논문에서 2016년부터 2020년까지 60개월 동안 유방암 심사청구 데이터를 활용하여 전체 연령, 40~50대, 60~80대 그룹으로 나누어 유방암 생존 여부를 예측하기 위한 모델 생성과 생존 여부에 영향을 미치는 요인을 분석하였다. 5개 알고리즘 평균 정밀도는 40~50대가 0.93으로 60~80대 0.86에 비하여 높았다. 재현율에 있어서는 모두 1.0으로 동일하였다. 생존 여부에 미치는 변수 중요도에 있어서도 40~50대는 치료횟수가 46%로 제일 중요하였으며, 60~80대는 나이가 32%로 중요한 요인이었다. 기타 변수인 치료기간, 주상병코드 및 수술여부는 연령대 구분 없이 일정 부문 생존 여부에 영향을 미치는 요인이었다. 알고리즘별 성능은 DT, RF 및 GB의 전체 평균 정밀도(0.90)와 재현율(1.0)은 동일하였으며, MLP와 RBF의 전체 평균 정밀도(0.89)와 재현율(1.0)도으로 동일하였다. 본 논문과 기존 연구 Byun et al.(2022)의 모델 성능을 비교 분석한 결과, 본 논문에서 제안한 모델의 평균 정밀도는 0.90, 재현율은 1.00으로 기존 논문의 정밀도 0.81, 재현율 0.94보다 더 높았다.

본 논문의 한계점은 데이터셋 구성에 있어 생존(148,133건) 및 사망(14,403건)에 대한 데이터 불균형에 대한 연구방법 처리가 미흡했다.

향후 과제로는 연구대상 자료 범위를 확대할 필요가 있다. 요양급여비용명세서 중 암 병기를 기록한 특정내역, 각종 검사, 치료 및 처방 등 진료내역 범위까지 확대할 필요가 있다. 더 나아가 건강보험공단의 건강검진 내용과 연계하여 건강검진과 치료결과와 연계한 종합적 분석이 필요하다. 심평원의 전국민 심사청구 빅데이터의 가치 활용을 제고하기 위해서 비전문가용 머신러닝 자동화(Auto ML) 도구를 사용하여 더 많은 연구가 진행되기를 바란다. 또한 진료비 심사, 의료의 질 평가 및 보건의료정책 수립의 실증 자료로 활용되기를 기대한다.

8) 3개 연령 그룹의 정밀도와 재현율 평균값 임

References

Adam, Y., Constance, L., Tal, S., Tally, P. and Regina, B. (2019). A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction, *Radiology,* 292, 60-66. https://doi.org/ 10.1148/radiol.2019182716.

Ayelet, A. B. Michal, C., Yoel, S. ... Adam, S. (2019). Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms, *Radiology,* 292, 331-342.

Byun, K. K., Lee, D. G. and Shin, Y. T. (2022). A Study on the Prediction of Mortality Rate after Lung Cancer Diagnosis for the Elderly in their 80s and 90s Based on Deep Learning, *Annual Spring Conference of KIPS*, 29(1), 452-455.

Choi, H. N., Kim, M. R., Lee, J. H., Kim, H. J. and Kim J. E. (2021). AI development data research for breast cancer screening diagnosis, *Proceedings of KIIT Conference* 2021(11), 633-636.

Keping, Y. Liang, T. Long, L. Xiaofan, C. Zhang, Y. and Takuro, S. (2021). Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health, *IEEE wireless communications,* June, 54-61.

Kang, S. A., Kim, S. H., and Ryu, M. H. (2022). Analysis of Hypertension Risk Factors by Life Cycle Based on Machine Learning, *Journal of the KIISR*, 27(5), 73-82.

Kim, G. G., Lim, E. T. and Kim, J. H. (2022). *Easily develop AI models with the AutoML platform WisePropet,* Seoul, Cheongram.

Lee, M. S. (2020). *Development of a prediction model for prognosis of triple-negative breast*

cancer based on deep learning using pathology images, Ph.D. Thesis, Graduate School of Ulsan University, Ulsan.Ministry of Heath and Welfar. (2021). *Cancer regstry statistics*, Sejong.

National Cancer Center. (2021). *Cancer Monitoring Indicator.*

Ryu, J. H. Hong, S. H. Park, H. G. Kim, D. M. Kim, S. J. and Park, S. J. (2017). Application of Machine Learning Techniques to Predict Stroke Diseases in Older Adults, *The Spring Conference of KOSES,* 42-42.

The Yakup. (2022). *Analysis by age by cancer type in 2019*, https://www.yakup.com (Accessed on March. 11th, 2022)

Yun, S. O., Jung, J. G., Wo, H. G. and Kim, J. E. (2022). Breast Cancer Survival Prediction: Model Comparison and Effect of Genetic Features, *Database Research,* 38(1), 3-15.

이 덕 규 (Doeggyu Lee)

- 정회원
- 계명대학교 전자계산학과 공학사
- 가톨릭대학교 의료경영대학원 의료경영학과 석사
- (현재)숭실대학교 IT정책경영학과 박사과정, 건강보험심사평가원 실장
- 관심분야: 보건의료 관련 IT정책 및 정보화분야

변 경 근 (Kyungkeun Byun)

- (현재)숭실대학교 IT정책경영학과 박사과정
- 관심분야 : 정보보호, 자동차 S/W 보안, 양자암호, ML

이 형 동 (Hyungdong Lee)

- 서울시립대학교 전자공학과 학사
- 건국대학교 정보통신대학원 정보보호학과 석사
- (현재) 숭실대학교 IT정책경영학과 박사과정, 국가안보전략연구원 수석연구위원
- 관심분야 : 정보보호, 사이버보안 정책, ICT공급망 보안, ML

신 선 희 (Sunhee Shin)

- 한양대학교 교육공학과 학사
- 이화여자대학교 대학원 교육공학과 석사
- 한양대학교 대학원 교육공학과 교육학박사
- (현재) 강남대학교 교육학과 초빙교수
- 관심분야 : SMART 교육, Blended-Learning, Ubiquitous-Learning