# 머신러닝을 활용한 연령대 별 치주질환 예측모델 비교

최은선[1*], Kong Vungsovanreach[1], 조완섭[2], 차은종[3], 김경아[3], 손호선[4]

충북대학교 빅데이터협동과정[1] 충북대학교 경영대학 경영정보학과[2],

충북대학교 의과대학 의공학교실[3], 충북대학교 의학연구소[4]

# Comparison of Prediction Models of Periodontal Diseases by Age Group Using Machine Learning

Eun-Seon Choi[1*],Kong Vungsovanreach, Wan-Sup Cho[2], Eun Jong Cha[3], Kyung Ah Kim[3], Ho Sun Shon[4]

Department of Big Data Cooperative Course, ChungBuk National University, Korea[1]

Department of Management Information Systems, Chungbuk National University[2]

Department of Biomedical Engineering, College of Medicine, Chungbuk National University[3]

Medical Research Institute, College of Medicine, Chungbuk National University[4]

*tmxk147@gmail.com

**Abstract**

Gingivitis and periodontal disease, the second most common diseases for the elderly over 65 years of age for outpatient treatment, increase national medical expenses and become a burden on health insurance finances. In addition, since periodontal disease can adversely affect not only daily life but also general health, various studies have been conducted to derive variables affecting the prevalence of periodontal disease, but studies targeting the elderly population over 65 are insufficient. Therefore, in this study, a predictive model was built and performance was evaluated after confirming the variables affecting periodontal disease using data of the elderly population aged 65 years or older during the 6th period of the National Health and Nutrition Examination Survey. These results can be used as base data for future periodontal disease as well as medical data analysis.

## 1. 연구 배경

치주 질환은 치아를 둘러싸는 치조골이 점차 소실되는 질병이다. 치주질환을 제 때 치료되지 않으면 치아 지지조직의 염증을 악화시켜 치아손실 (tooth loss)을 가져올 수 있으며 [1] 이를 방치할 경우 저작능력의 저하로 음식의 선택이 제한되어 영양불균형을 초래하고, 일상생활은 물론 전신건강상태에도 악역향을 미칠 수 있다. 치주질환의 염증징후와 진전은 개인의 성향, 사회적 요인, 전신적 요인, 유전적 요인, 치아 상태, 치면세균막 미생물 등의 다양한 요인의 영향을 받는다[2]. 또한 치주질환은 삶의 질과 직접적으로 연관된 요인 중 하나로, 건전치주군에 비해 치주질환군에서 1.32배 삶의 질 저하가 될 위험이 더 높음을 확인한 연구 결과가 있다[3]. 치주질환과 인구 · 사회경제적 수준 및 구강건강행위 등과 관련된 연구들과 한국 성인의 치주질환 유병관련 위험요인에 대한 연구도 진행되었다[4]. 이러한 선행연구들은 주로 일부 위험요인과 치주질환의 관련성만을 살펴본 연구들이 대부분이며, 다양한 머신러닝 알고리즘을 적용하여 예측모델을 구축하는 연구는 미흡한 실정이다.

따라서 본 연구에서는 국민건강영양조사(6기)의 자료에서 나이 65세 이상 데이터를 추출하고, 로지스틱 회귀 (Logistic regression: LR), 서포트벡터머신 (Support Vector Machine: SVM), 랜덤 포레스트 (Random Forest: RF), 에아다부스트 (Adaboost: AB)를 사용하여 다양한 예측 모델을 구축하고 모델 별 정확도를 비교 분석하였다. 본 논문의 구성은 다음과 같다. 2절에서는 연구 방법에 대해 제시하고 3절에서는 연구의 결과를 제시하며 4절에서는 결론을 도출하였다.

머신러닝을 질병 예측에 적용한 기존연구에서는 2012년 국민건강영양조사 자료 중 40세 이상 성인의 대사증후군 유병 여부 예측에 영향을 미치는 변수를 추출하고 예측모형을 개발하였다[5]. 또 다른 연구에서는 XGBoost, 랜덤포레스트, SVM을 적용하여 백내장 예측모델을 구축하고 60세 이상 한국인의 식이 섭취와 백내장 유병의 관련성을 도출하였다[6].

## 2. 연구 방법

국민건강영양조사(The Korea National Health and Nutrition Examination Survey, KNHANES)는 국민건강증진법 제16조에 근거하여 시행하는 전국규모의 건강 및 영양조사이다. 본 연구에서는 국민건강영양조사 제6기(2013~2015)의 데이터 중 나이가 65세 이상인 데이터를 추출하여 사용하였고, 선택된 데이터는 51개의 변수와 1,921개의 행으로 구성되어 있다.

본 연구에서는 독립변수의 경우 선행 연구에 의해 추출된 치주질환과 관련된 유의한 변수들을 선택하고 종속변수로 치주질환 유병 여부를 사용하였다[7]. 독립변수는 인구사회학적특성, 건강행태관련 특성, 구강건강 특성으로 구분하였다. 인구사회학적 특성으로 지역, 성별, 나이, 민간의료보험 가입여부, 가구소득, 가구원수, 기초생활 수급여부, 주택소유여부, 결혼여부, 건강보험종류, 교육수준, 경제활동 상태를 사용하였다. 치주염과 다양한 질환의 관련성에 대해 보고되고 있으며[8] 건강관련 행태특성으로 고혈압, 이상지질혈증, 뇌졸중, 심근경색증, 협심증, 당뇨병, 비만 의사진단여부, 주관적건강상태, 건강검진 수진여부, 주관적 체형인식, 평생 음주경험, 평생 흡연여부, 하루 평균 수면 시간, 최종수축기혈압, 최종이완기혈압, BMI, 공복혈당, 당화혈색소, 총콜레스테롤, 전환식 HDL 콜레스테롤, 중성지방, 저 HDL 콜레스테롤 혈증 유병여부, AST, ALT, 백혈구, 요단백, 요당, 요케톤, 요빌리루빈을 사용하였다. 구강건강 특성으로는 우식영구치수, 영구치 우식유병여부, 본인 인지 구강건강상태, 최근 1년간 치통 경험여부, 교정치료 경험 여부, 씹기 문제, 저작불편 호소여부, 말하기 문제, 어제 하루 칫솔질 여부를 사용하였으며 또한 선행연구를 참고하여 칫솔질 횟수, 사용하는 구강용

품 수를 파생변수로 추가하였다[10]. 분석도구는 R (v4.0.5), Jupyter Notebook, Python(v3.8.5)을 사용했다. 로지스틱회귀로부터 도출된 결과는 혼돈행렬(confusion matrix)을 기반으로 정확도(accuracy), 정밀도(precision), 재현율(recall), F1 score를 기반으로 평가하였다.

## 3. 연구 결과

주성분 분석 결과 3개의 주성분으로 축소되었고 전체 데이터의 99.19%를 설명한다. 주성분 분석결과 요약은 <표 1>과 같다. 주성분 분석을 통해 도출된 변수는 다음과 같이 구성되어 있다. 제 1 주성분은 체질량지수와 강한 양의 관계를 가지고 있으며, 제 2 주성분은 백혈구 수와 강한 양의 관계를 가지고 있다. 제 3 주성분은 중성지방과 체질량지수와 강한 음의 관계를 가지고 있다

표 1. 노년 주성분 분석결과 요약

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| Standard deviation | 548.91 | 143.68 | 62.01 |
| Proportion of Variance | 0.92 | 0.06 | 0.01 |
| Cumulative Proportion | 0.92 | 0.98 | 0.99 |

본 연구에서는 치주질환 관련 유의한 변수 추출 후 기계학습 알고리즘을 사용하여 나이 별 모델의 분류 성능을 예측하였다. <표 2>는 노년의 주성분 분석과 오토인코더에서 추출된 변수를 기반으로 치주질환 유무에 따른 분류 모델을 비교 분석한 것이다.

표 2. 노년 모델 별 성능 평가

| Machine Learning | Evaluation | Feature Selection | PCA | AE |
|---|---|---|---|---|
| LR | Accuracy | 0.59 | 0.54 | 0.49 |
| | Precision | 0.57 | 0.53 | 0.49 |
| | Recall | 0.69 | 0.76 | **0.97** |
| | F1 Score | 0.63 | 0.62 | 0.65 |
| SVM | Accuracy | 0.53 | 0.51 | 0.53 |
| | Precision | 0.53 | 0.51 | 0.52 |
| | Recall | 0.58 | 0.87 | 0.82 |
| | F1 Score | 0.55 | 0.64 | 0.63 |
| RF | Accuracy | 0.58 | 0.51 | 0.55 |
| | Precision | 0.57 | 0.51 | 0.54 |
| | Recall | 0.70 | 0.55 | 0.59 |
| | F1 Score | 0.63 | 0.53 | 0.56 |
| AB | Accuracy | 0.56 | 0.54 | 0.48 |
| | Precision | 0.63 | 0.46 | 0.47 |
| | Recall | 0.54 | 0.56 | 0.27 |
| | F1 Score | **0.82** | 0.39 | 0.35 |

## 4. 결론

본 연구에서는 2013년부터 2015년까지 진행된 국민건강영양조사 6기 자료를 이용하여 노인 인구의 치주질환 여부 예측모델을 구축하고 모델 별 성능을 비교하였다. 예측모델을 구축하기 위해 로지스틱 회귀, 서포트벡터머신, 랜덤 포레스트, 아다부스트 알고리즘을 이용하였으며 선행연구를 참고한 변수 선택방법과 주성분 분석 및 오토인코더를 사용한 차원 축소를 통해 모델의 성능을 개선하고자 하였다. 기계학습 알고리즘의 비교 분석을 위해 혼동행렬을 기반으로 모델을 평가하였다. 정확도가 가장 높은 모델은 로지스틱회귀분석에 변수선택을 적용한 모델이 0.59로 가장 높았으며, 정밀도가 가장 높은 모델은 에이다부스트에 변수선택을 적용한 모델이 0.63으로 가장 높았다. 재현율의 경우 로지스틱회귀모델에 오토인코더를 적용한 모델이 0.97로 가장 높았고, F1 score의 경우 에이다부스트에 변수 선택을 적용한 모델이 0.82로 가장 높았다.

향후 연구에서는 국민건강영양조사 자료를 이용하여 연령별 다양한 특성에 따른 치주질환 유무와 관련된 요인을 추출하고 예측모델을 구축할 것이다.


## 5. Acknowledgements

이 성과는 정부(교육부, 과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1I1A1A01065199, No. 2019R1F1A1051569).


## 6. 참고 문헌


[1] M. A. Jeong, J. H. Kim. "Association between cardiovascular disease and periodontal disease prevalence." *Journal of the Korea Convergence Society*, Vol 2, No. 4, p. 47-52, 2011.

[2] M. E. Nunn. "Understanding the etiology of periodontitis: an overview of periodontal risk factors." *Periodontol*, Vol 2003, No. p. 11-23, 2000.

[3] J. Yu, S. Hwang. "A convergence study on the effect of periodontal disease on health-related quality of life in adults over 40s." *Journal of the Korea Convergence Society*, Vol 12, No. 6, p. 49–56, 2021.

[4] J. O. Jung, G. J. Oh. "A study of the relationship between socioeconomic status, oral health behaviors and periodontitis in the elderly Korean population." *J Korean Acad Oral Health*, Vol 35, p. 57-66, 2011.

[5] H. K. Kim, K. H. Choi, S. W. Lim, H. S Rhee. "Development of prediction model for prevalence of metabolic syndrome using data mining: Korea National Health and Nutrition Examination Study." *Journal of Digital Convergence*, Vol 14, No. 2, p. 325–332, 2016.

[6] J. Y. Choi. "A study on the relationship between dietary intake and cataract in Koreans over 60 years of age & development of cataract prediction model based on artificial intelligence(AI) – The Korea National Health and Nutrition Examination Survey 2015-2017." Kyungnam University, 2021.

[7] J. H. Lee. "The relationship between metabolic syndrome components and the number of remaining teeth in Korean adults." *Journal of Korean Academy of Oral Health*, Vol 44, No. 3, p. 130-137, 2020.