



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공 학 석 사 학 위 논 문

소변 스펙트럼 분석을 위한
기계학습 모델



2020년 8월

부 경 대 학 교 대 학 원

컴 퓨 터 공 학 과

최 재 혁

공 학 석 사 학 위 논 문

소변 스펙트럼 분석을 위한 기계학습 모델

지도교수 정 목 동

이 논문을 석사 학위논문으로 제출함.

2020년 8 월

부 경 대 학 교 대 학 원

컴 퓨 터 공 학 과

최 재 혁

최재혁의 공학석사 학위논문을 인준함.

2020년 8월 28일

위원장	공학박사	한영선 (인)
위원	공학박사	장원두 (인)
위원	공학박사	정목동 (인)



목 차

목차.....	i
요약.....	v
Abstract.....	vi
I. 서론.....	1
1.1 연구의 개요.....	1
1.2 연구의 필요성 및 목적.....	3
1.3 연구의 방법 및 범위.....	5
II. 관련연구.....	6
2.1 소변검사.....	6
2.2 스펙트럼.....	8
2.3 기계학습.....	9
III. 소변 스펙트럼 분석.....	11
3.1 의료 데이터와 혼합모델의 이용 사례.....	11
3.1.1 기계학습 기반 체외충격파 요로결석 쇄속술 예측 모델.....	11
3.1.2 효과적인 뇌전증 발작 감지를 위한 스태킹 앙상블 기반 심층 신경망 모델링.....	12
3.2 데이터 수집 및 가공.....	13
3.2.1 소변검사 결과 수집.....	13
3.2.2 스펙트럼 변환.....	14
3.3 시스템 구조.....	18

IV. 실험 및 평가.....	21
V. 결론 및 향후연구.....	25
참고문헌.....	27



그림 목 차

[그림 1] 소변 검사 구조.....	6
[그림 2] 배킹과 부스팅 흐름도	10
[그림 3] 뇌전증 발작 감지를 위한 스테킹 앙상블 심층신경망 모델 실험 결과....	12
[그림 4] 소변 스펙트럼 분석 순서	13
[그림 5] 소변 스펙트럼 변환.....	15
[그림 6] 스펙트럼 A ~ W 18개 채널.....	16
[그림 7] 소변 스펙트럼 데이터.....	17
[그림 8] 각 모델의 역할.....	18
[그림 9] 제안 시스템 군집화 성공 데이터를 이용한 학습 흐름도.....	17
[그림10] 제안 시스템 군집화 실패 데이터를 이용한 학습 흐름도.....	18
[그림11] 제안 시스템 예측 흐름도.....	18

표 목 차

[표 1] 병원의 소변검사 문제점.....	3
[표 2] 소변검사 10가지 성분 명.....	5
[표 3] 소변 성분별 질병.....	7
[표 4] 스펙트럼 가시광선, 적외선 세부 영역.....	8
[표 5] 체외충격파 요로결석 쇄석술 예측 모델 실험결과.....	11
[표 6] 병원에서 제공받은 소변검사 성분 명 및 결과.....	14
[표 7] 측정 도표.....	22
[표 8] 각 성분별 통계.....	23

소변 분석을 위한 혼합모델

최 재 혁

부 경 대 학 교 대 학 원 컴 퓨 터 공 학 과

요 약

병원에서 수술 전/후 환자, 신장질병 환자, 중환자, 요양병원 환자의 신장 및 순환기계를 확인하기 위해 소변 상태를 간호사가 주관적으로 판단을 내리고 있다. 이를 개선하기 위해 본 논문에서는 소변 스펙트럼 분석 시스템을 제안하고, 스펙트럼 데이터와 혼합 모델의 첫 부분인 비지도학습을 이용한 소변 검사 결과를 군집화 하고 지도학습을 사용하여 분류한다. 혼합 모델을 이용한 소변 스펙트럼 분석 결과를 지도학습 결과와 평가 한다. 본 논문은 병원의 간호사 부족 및 검사 시간 단축, 주관적인 평가를 해결하여 환자에게 기존 의료 서비스 보다 좋은 서비스를 제공 할 수 있을 것으로 기대된다.

Hybrid Machine Learning Model for Urine Spectrum Analysis

Jae Hyeok Choi

Department of Computer Engineering, Graduate School
Pukyong National University

Abstract

Nurses make subjective judgments on urine conditions to check the kidneys and circulatory systems of pre- and post-operative patients, kidney disease patients, intensive patients, and nursing hospital patients in hospitals. To improve this, this paper proposes the Urine Spectrum Analysis System, and clusters the results of urine tests using the Unsupervised Learning technique, the first part of the Spectrum data and the system, and evaluates the results of the classification guidance and learning using the Supervised Learning technique. It is expected that the hospital will be able to provide better services to patients than conventional medical services by reducing the hospital's lack of nurses, reducing test time and resolving subjective assessments.

I. 서론

서론에서는 본 논문에 대한 전체적인 흐름에 대한 설명과 연구의 필요성 및 목적, 연구의 범위에 대하여 설명하고자 한다.

1.1 연구의 개요

최근 개인의 의료정보를 바탕으로 분석 및 예측하여 각종 의료 서비스를 제공하는 연구가 활발하게 진행되고 있다.[1] 본 연구에서는 이러한 개인 의료정보 중 병실에서 수기로 측정하는 소변 상태 결과를 머신러닝의 혼합모델을 이용하여 예측하고자 한다.

소변 상태 검사는 중환자의 경우는 첫 24시간 소변량이 사망률과의 높은 연관성이 있어 중요한 검사이다. [2] 그래서 수술 전/후 환자, 신장 질병 환자, 중환자, 요양병원에서 신장과 순환기계 확인을 위해 유린백(Urine Bag)를 사용하여 측정을 하고 있다. 측정 방식은 간호사가 시간당 소변량, 색상 탁도, 성분을 소변 비색표를 보고 주관적으로 결정을 하고 있으며 최근 병원에서는 간호사 부족으로 인해 대 소변량 측정, 맥박, 혈압측정 등을 간호보조 인력에게 위임하고 있다.[3] 또한 소변을 측정 후 전자의무기록(EMR, Electronic Medical Record)을 통해 의료진과 공유하는 시간까지 1~3시간 소요가 되어 환자에 대한 빠른 대응이 불가하다.

그래서 본 연구는 기존 간호사 소변 검사와 관련한 물리적 자원소비를 최소화 하는데 목적이 있다. 따라서 소변을 가시광선 적외선 영역의 스펙트럼으로 변환한 18개 채널의 데이터를 비지도 학습과 지도학습을 혼합한 혼합모델을 이용한 시스템을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2 장 관련 연구에서는 소변해석, 스펙트럼, 머신러닝 기술들을 살펴보고, 3 장에서는 의료 데이터와 혼합모델의 이용 사례와 제안하는 소변 스펙트럼 분석 구성을 살펴본다. 4 장에서는 제안시스템을 평가하고, 마지막으로 5 장에서는 결론 및 향후 연구 방향을 제시한다.



1.2 연구의 필요성 및 목적

소변은 소변 혈액이 신장을 통과하면서 혈액 내의 노폐물이 걸러진 후 방광에 모였다가 배출되는 것이다. 이러한 소변 생성 과정을 통해 신장을 체내 수분량을 조절하게 되며 체내에서 재사용할 수 있는 단백, 전해질 및 기타 화합물을 보분하는 노폐물을 배출시키는 역할을 하고 있다.

소변 검사란 소변의 색, 혼탁도 등 물리적 성상을 검사하고 소변으로 배출되는 여러 종류의 노폐물을 반 정량적으로 검출하는 검사이며 요로 감염 등의 신장 및 비뇨기계 질환, 내분비 질환, 대사성 질환 및 전해질 이상을 비롯한 각종 질환의 선별검사로 이어지기 때문에 매우 중요한 검사이다. 소변은 채취가 용이하여 환자에게 주는 부담이 적어 건강검진, 임신 중, 입원 혹은 수술 전 검사 목적으로 모든 환자에게 일차적으로 시행하는 검사이다.

표 1 병원의 소변검사 문제점

정보 공유시간 지연
간호사가 소변을 측정 후 EMR(전자의무기록, EElectronic Medical Record) 또는 EHR(전자건강기록, Electronic Health Record)을 통해 의료진과 공유하는 시간까지 약 1~3시간 소요
의료 인력 부족
소변 상태를 측정에는 많은 시간이 소요되는데 간호사 인력 부족으로 소변 상태 측정을 간호보조 인력에 위임(대, 소변량측정, 맥박, 혈압측정, 흡인, 위관급식 등)
간호사 주관적 소변검사 결정
환자의 소변 이상 유무 1차적 판단이 간호사의 주관에 결정

표 1은 현재 병원에서 소변 검사를 할 때 정보 공유시간 지연, 의료 인력 부족을 문제점으로 나타내고 있다. 이러한 문제점을 개선하기 위해 소변 스펙트럼 분석을 제한한다. 또한 기존 지도학습으로 의료 데이터를 학습하였다면 본 논문에서는 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning) 즉 혼합모델(Hybrid Model)을 제안하여 기존 지도학습 보다 좋은 결과를 얻을 수 있는 방향을 제시한다.



1.3 연구의 방법 및 범위

본 연구에서는 서부산 C병원에서 약 7만개 소변검사 데이터를 제공받아 연구를 진행하였고 부경대학교 기관생명윤리위원회에서 승인 받은 연구이다.(승인번호 : 1041386-202003-HR-15-02) 아래 표 2는 제공받은 소변검사 데이터의 10가지 성분 명을 나타내고 있다.

본 논문의 연구 방법은 비지도학습과 지도학습을 혼합하여 사용을 하며 비지도학습으로 군집화 하고 지도학습을 사용하여 분류한다.

연구 범위는 본 논문에서 10가지 성분 중 검출이 되면 위험한 4가지 성분을 바탕으로 하여 성분의 미검출, 검출을 분류를 기준으로 한다.

표 2 소변검사 10가지 성분 명

번호	표기	성분 명(한글)	성분 명(영어)
1	BLD	혈뇨	Blood
2	BIL	요빌리루빈	Billirubin
3	URO	유로빌리노겐	Urobilinogen
4	KET	요케톤체	Ketone Body
5	PRO	요단백	Protein
6	NIT	요아질산염	Nitrite
7	GLU	포도당	Glucose
8	P.H	산도	PH
9	S.G	요비중	Specific Gravity
10	LEU	류코사이트	Leukocyte

II. 관련 연구

2.1 소변검사

소변검사는 3가지 방법이 있으며 물리적 성상 검사, 화학적 검사, 요침사 검사가 있다. 물리적 성상 검사는 소변의 육안적 색조 및 혼탁도, 냄새를 검사하고 화학적 검사는 요시험지봉을 이용하여 10가지 항목을 반정량적으로 검출하는 검사이다 마지막으로 요침사 검사는 현미경을 이용하여 적혈구, 백혈구, 세균 및 각종 결정을 관찰하는 검사이다. 보통 병원에서 소변검사 라고 한다면 물리적 성상 검사와 화학적 검사를 의미하며 요침사 검사는 소변검사 결과에 따라 추가적인 검사로 진행을 한다.

그림 1은 소변검사 구조이며 보통 물리적 성상 검사와 화학적 검사는 같이 진행을 한다. 간호사가 유린 백에서 소변의 색, 혼탁도, 냄새를 검사 후 요시험지봉과 비색표를 가지고 육안으로 확인 후 추가적인 검사가 필요할 경우 임상병리실의 임상병리사가 요시험지봉을 기계로 다시 한번 검사를 한다. [4]

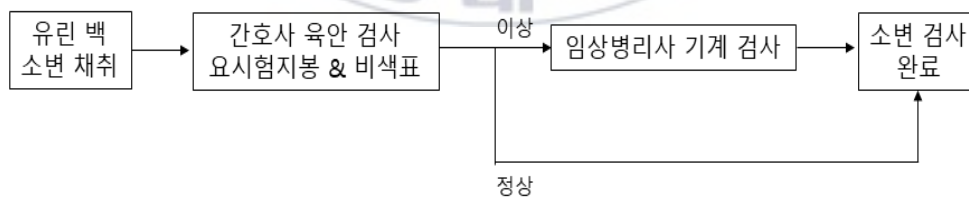


그림 1. 소변 검사 구조

간호사마다 지정된 병실의 환자를 검사 후 전자의무기록에 올려 다른 의료진과 공유하는 시간이 1시간 ~ 3시간 소요가 된다.

표 3 소변 성분별 질병

번호	성분 명	병명
1	혈뇨	요로결석, 전립선 비대, 신장염
2	요빌리루빈	간염, 폐쇄성 황달
3	유로빌리노겐	간질환
4	요케톤체	대사성질환, 당뇨병
5	요단백	신장염
6	요아질산염	요로감염
7	포도당	당뇨병
8	산도	신지대사 기능
9	요비중	신장 기능
10	류코사이트	백혈구 염증 수치

보통 소변 검사를 하면 신장기능 확인, 요로결석, 전립선 비대, 신장염 위험 가능성, 당뇨, 방광염, 신우신염 등을 검사를 한다. 위 표 3 은 소변 성분별 질병 항목 이다.

2.2 스펙트럼

스펙트럼은 파장 또는 주파수에 따라 측정을 하여 감마선($\sim 0.01\text{nm}$), 엑스선($0.01\text{nm} \sim 10\text{nm}$), 자외선($10\text{nm} \sim 400\text{nm}$), 가시광선($400 \text{ nm} \sim 700\text{nm}$), 적외선($700\text{nm} \sim 1000\text{nm}$), 마이크로파($1000\text{nm} \sim 1\text{m}$), 전파($1\text{m} \sim$) 총 7가지 영역으로 구분한다.

본 논문에서는 아래의 표 4 처럼 가시광선, 적외선 2가지 영역을 사용하여 실험을 진행하였다.[5]

표 4 스펙트럼의 가시광선, 적외선 세부 영역

가시광선							적외선
보라	남색	파랑	초록	노랑	주황	빨강	
380~ 425	425~ 450	450~ 495	495~ 570	570~ 590	590~ 620	620~ 780	780~ 1000

단위 : nm(nano meter))

2.3 기계학습(지도학습, 비지도학습, 혼합모델)

기계학습은 컴퓨터에게 사람이 직접 지시하지 않아도 데이터를 통해 컴퓨터가 학습을 하고 그것을 사용해 컴퓨터가 자동으로 문제를 해결할 수 있도록 하는 것을 의미한다.

기계학습은 크게 3가지로 구별할 수 있는데 지도학습, 비지도학습, 강화학습으로 구분이 되며 본 논문에서는 지도학습과 비지도학습 그리고 두가지 경우를 혼합하여 사용하는 혼합모델만 사용한다.

지도학습은 주어진 훈련 데이터의 입력 데이터와 목표 데이터로 이루어지는 경우이다. 즉 훈련 데이터가 사진1, 사진2, 사진3 경우 각각 훈련 데이터에 고양이, 강아지, 새로 지정하여 학습을 진행한다.

비지도학습은 데이터 내에서 비슷한 특징을 집단화, 군집화 하는 방법을 말하며 훈련 데이터가 목표 데이터 없이 오직 입력 데이터로만 학습하는 경우이다. 예를 들어 지도학습과 다르게 문장1, 문장2, 문장3 데이터를 가지고 학습을 진행한다.

혼합모델은 2가지 이상의 모델을 사용하는 학습 기법이며 본 논문에서는 비지도학습과 지도학습 방법을 상호 보완적으로 사용하지만 이 경우 비지도학습에서 지도학습으로 넘어갈 때 비지도학습의 결과인 군집화에서 지도학습 데이터로 만들어야 되는 문제점이 있는데 본 논문에서는 전문가 확인 시스템 방법으로 해결을 한다.[6]

본 논문에서 사용하는 비지도학습 모델은 K-means 알고리즘이며 지도학습으로 사용된 모델은 XGBoost(Extreme Gradient Boosting)이다. K-means 알고리즘은 주어진 데이터를 k개의 군집화 하는 알고리즘으로 각 군집화의 거리 차이의 분산을 최소화하는 방식으로 동작한다. XGBoost는 앙상블 모델의 부스팅(Boosting)기법을 이용한 모델이다. 앙상블 모델은 여러 모델을 이용하여 학습을 하여 모델의 예측결과를 평균하여 예측을 하는 방법이며 다수의 모델을 사용을 하여 혼합 모델이라고도 한다. 앙상블 모델에서 그림 2 처럼 배깅(Bagging) 기법과 부스팅 기법으로 나뉘며 배깅 기법은 다수의 모델을 병렬 형식으로 학습 시키는 기법이며 각 모델은 독립적으로 학습이 진행이 하며 최종 결과는 각 모델 결과를 합하여 구한다. 대표적인 배깅은 랜덤 포레스트(Random Forest)가 있다. 부스팅 기법은 약한 분류가 가능한 다수의 모델을 결합하여 강한 모델을 만드는 기법이다 주어진 데이터를 약한 분류기를 통해서 학습한 후 학습된 결과에서 나타나는 오차를 또 다른 약한 분류기에서 학습시켜 오차를 줄여 나간다. 첫 번째 학습을 통해 생성된 모델에서 오류를 발생 시키는 데이터를 여러 모델을 순차적으로 걸쳐 오류를 줄여나가며 최종 결과는 배깅 기법과 같이 각 모델 결과를 합하여 구한다.[7] 대표적인 부스팅은 AdaBoost, XGBoost가 있다. 본 논문에서 사용하는 기법은 부스팅 기법이며 그중에 XGBoost를 2진 분류 지도학습으로 사용을 한다.

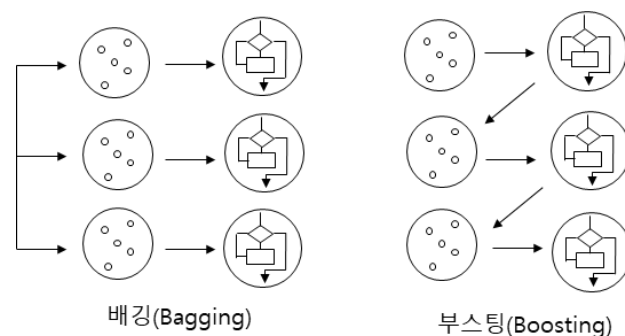


그림 2. 배깅과 부스팅 흐름도

Ⅲ. 소변 스펙트럼 분석

3.1 의료 데이터와 혼합모델의 이용 사례

3.1.1 기계학습 기반 체외충격파 요로결석 쇄석술 예측 모델

요로 결석을 제거하는 방식 결석의 위치에 따라 경피신석제거술, 역행성 신내 수술, 요관경하결석제거술, 방광경하결석제거술, 체외충격파쇄설술로 나뉘며 국가수리과학연구소에서 진행한 이 연구는 체외충격파쇄석술을 이용하기 위한 기계학습 모델을 XGBoost로 제안 하여 아래 표 5의 결과를 보면 SVM 및 RF 보다 뛰어난 성능을 보여주고 있다.[8]

표 5 체외충격파 요로결석 쇄석술 예측 모델 실험 결과

	Stone free rate	One session success	Colic pain
SVM(Support Vector Machine)			
Training Accuracy	73.23	59.50	48.96
Test Accuracy	69.67	59.74	53.17
RF(Random Forest)			
Training Accuracy	85.49	75.39	74.60
Test Accuracy	84.82	79.14	76.32
XGBoost(Extreme Gradient Boosting Trees)			
Training Accuracy	88.78	75.71	77.64
Test Accuracy	87.07	80.17	76.74

3.1.2 효과적인 뇌전증(간질) 발작 감지를 위한 스택킹 앙상블 기반 심층신경망 모델링

뇌의 전기적 활동에서 얻은 뇌파 신호는 신경성 질환의 진단에 흔히 사용이 된다. 이 뇌파신호는 뇌의 전기적 활동을 나타내며 뇌에 대한 정보를 포함하고 있다. Kemal Akyol 연구자는 이러한 뇌파의 정보를 이용하여 뇌전증을 다수의 DNN 모델을 이용하여 분류할 것을 제안을 한다. 아래 그림의 Sensitivity는 뇌전증이 있는 환자를 검사하여 뇌전증이 있다고 나오는 의미이며, Specificity는 뇌전증이 없는 환자를 검사하여 뇌전증이 없다고 나오는 의미이다. 아래의 그림 3을 보면 Specificity는 좋은 성능을 보여주고 있다.[9]

Comparison of epileptic seizures detection results of the proposed study and previous studies.

Studies	Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)
Guo et al. (2010)	ANN	95.50	99.00	98.27
Peker et al. (2016)	Complex-valued neural networks	100.0	98.01	99.33
Fu et al. (2015)	SVM	-	-	98.80
Jaiswal and Banka (2017)	ANN	98.30	98.82	98.72
Wang et al. (2017)	SVM	97.98	99.56	99.25
Thara et al. (2019)	DNN	98.59	91.47	97.21
Gao et al. (2020)	CNN	-	-	99.26
Ullah et al. (2018)	CNN	98	98	98.1
Proposed study	SEA-based DNN	93.11	98.18	97.17

그림 3. 뇌전증 발작 감지를 위한 스택킹 앙상블 심층신경망 모델 실험 결과

3.2 소변 데이터 수집 및 가공

본 논문의 전체적인 소변 스펙트럼 분석 구조는 그림 4 의 그림 처럼 소변검사, 스펙트럼 변환, 비지도학습 : K-means 군집화, 영역 전문가 확인 시스템, 지도학습 : XGBoost 로 구성이 되어있다.

소변 검사에서 BLD, PRO, GLU, LEU의 데이터를 스펙트럼으로 변화 후 혼합모델의 비지도 학습을 이용해 군집화를 후에 영역 전문가 확인 시스템에서 영역 전문가-의사의 확인을 통해 데이터를 가공 후 지도학습으로 최종 결과를 예측한다.

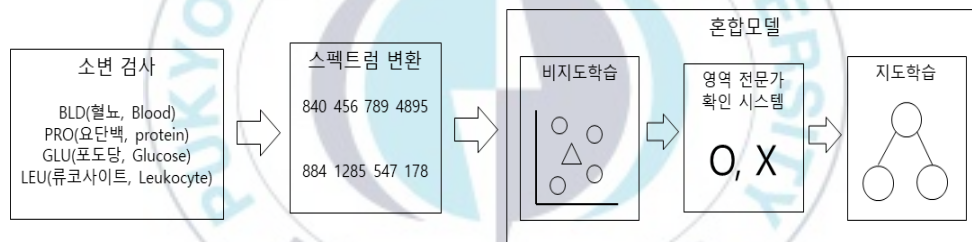


그림 4. 소변 스펙트럼 분석 순서

3.2.1 소변검사 결과 수집

아래 표 6 는 병원으로부터 소변검사 결과 데이터 수집 항목이며 BLD, PRO, GLU 총 4개의 성분의 결과를 약 7만개 데이터를 받았다. 각 성분의 검사결과가 뜻하는 의미는 neg(Negative)의미는 미검출을 의미하고 숫자 결과는 낮은 수치부터 큰 수치까지 반정량적으로 측정이 된다.

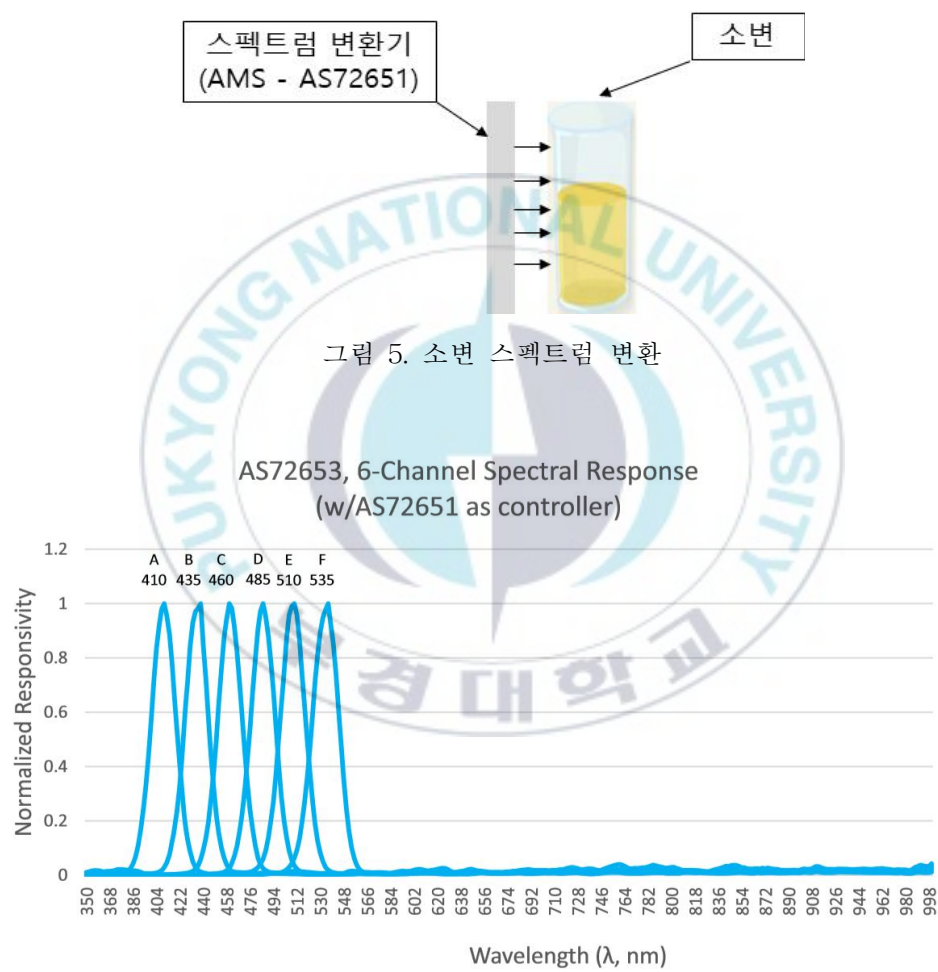
표 6 병원에서 제공받은 소변검사 성분 명 및 결과

번호	표기	검사 결과
1	BLD	neg, 5, 10, 50, 250
2	PRO	neg, 10, 30, 100
3	GLU	neg, 1000, 2000
4	LEU	neg, 10, 25, 75

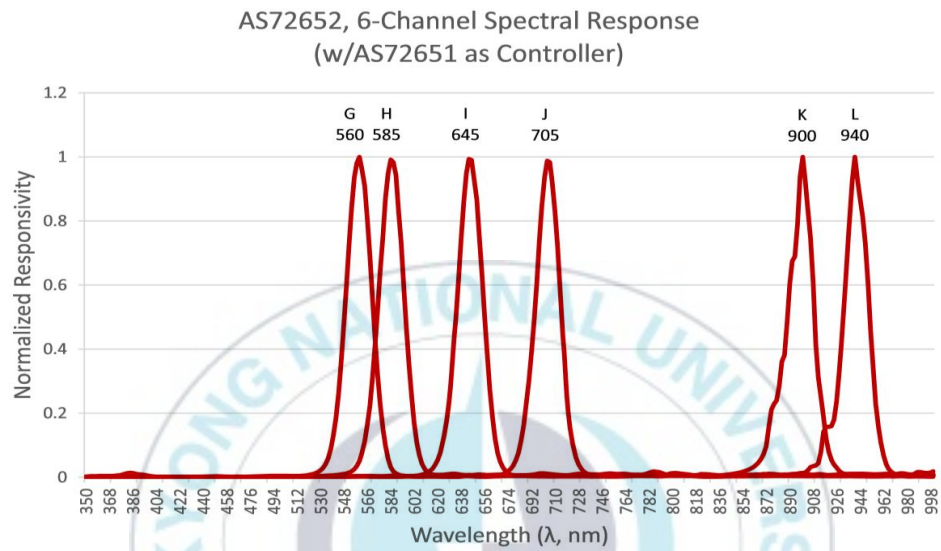
3.2.2 스펙트럼 변환

스펙트럼 데이터 변환하기 위해서 서부산 C병원에서 환자의 소변을 받아 아래 그림 5 처럼 AMS회사의 AS72651 기계를 이용하여 소변을 스펙트럼 데이터로 변환하였고 2장 관련 연구 스펙트럼에서 얘기 하였던 스펙트럼 영역 중 가시광선, 적외선이며 넓은 스펙트럼 영역을 얻기 위해 3개의 기계 그림 6 의 (a),(b),(c)를 사용하여 각 기계마다 6개의 총 18 채널을 설정하여 변환을 하였다. (a)는 보라, 남색, 파랑, 초록을 검출 하였으며 (b)는 초록, 노랑, 주황, 빨강, 적외선 영역 2개를 검출 하고 마지막으로 (c)는 주황, 빨강 영역 3개, 적외선 영역 2개를 검출 하였다. 그리하여 최종적으로 그림 6 의 A ~ W까지의 스펙트럼 데이터로 변환 하였다.

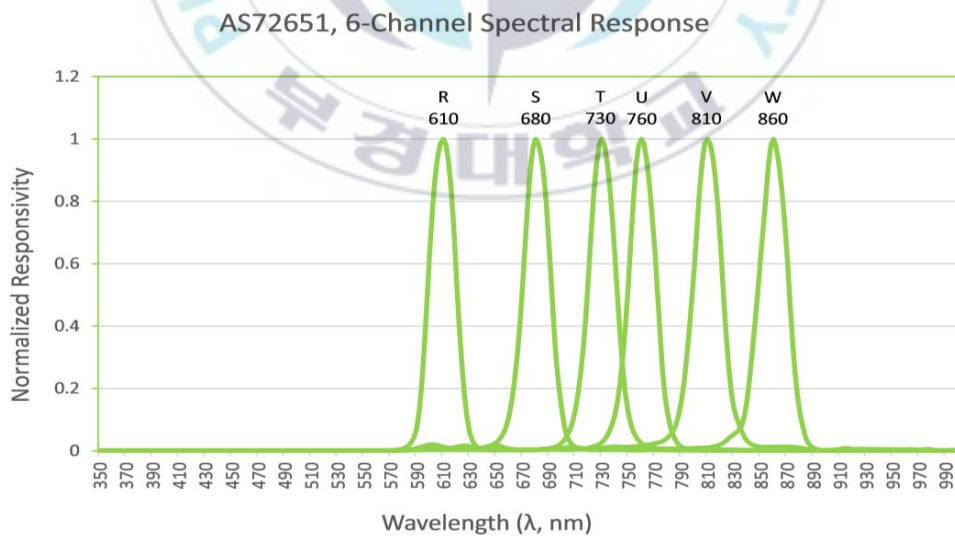
소변을 스펙트럼 변환 방법은 아래 그림 5 처럼 소변을 스펙트럼 변환기기를 이용하여 연구에서 필요한 18 채널 스펙트럼 데이터를 수집한다.



(a)



(b)



(C)

그림 6. (a),(b),(c) 스펙트럼 A ~ W 18개 채널 [5]

A	B	C	D	E	F	G	H	I	J	K	L	R	S	T	U	V	W
830	1562	2794	1618	2598	2669	3409	4047	3627	1376	121	76	10	47	21	66	1301	8234
826	1559	2797	1610	2596	2670	3427	4066	3638	1379	121	77	10	47	21	66	1295	7840
818	1550	2786	1609	2593	2665	3415	4054	3632	1378	121	77	10	47	21	66	1318	8230
827	1554	2788	1615	2594	2666	3437	4078	3641	1384	122	77	10	47	21	66	1315	8147
824	1558	2786	1613	2598	2666	3405	4047	3629	1374	121	77	10	47	21	66	1307	8171
820	1541	2746	1595	2569	2630	3431	4085	3651	1382	122	77	10	47	21	66	1324	8264
826	1557	2781	1611	2589	2656	3408	4060	3637	1377	121	77	10	47	21	66	1312	8121
806	1521	2728	1584	2558	2634	3424	4065	3634	1379	121	77	10	47	21	66	1300	8016
821	1554	2777	1603	2583	2655	3420	4057	3625	1377	121	77	10	46	21	66	1321	8204
826	1559	2788	1612	2593	2668	3398	4045	3631	1369	121	77	10	47	21	66	1309	8227
828	1564	2797	1618	2601	2671	3413	4058	3637	1376	121	77	10	47	21	66	1306	8149
821	1557	2787	1603	2589	2668	3388	4031	3609	1367	120	76	10	47	21	66	1326	8239

그림 7. 소변 스펙트럼 데이터



3.3 시스템 구조

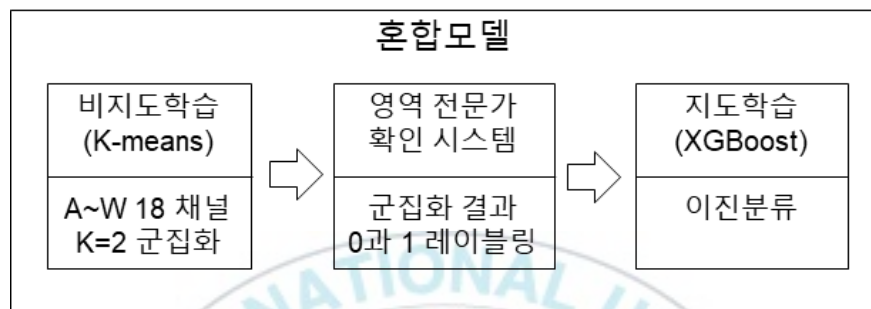


그림 8. 각 모델의 역할

일반적으로 사용하는 혼합모델, 앙상블 모델은 다수의 지도 학습 모델을 가지고 학습을 한다. 하지만 비지도 학습과 지도 학습을 같이 사용하는 경우는 다수의 지도 학습과 다른 방법이 필요하다. 비지도 학습과 지도 학습의 차이점은 학습할 데이터에 있는데 비지도 학습은 학습 데이터의 정답을 알 수 없고, 지도 학습은 정답을 알고 있다 즉 비지도 학습의 결과는 데이터의 특징에 따라 군집화 결과가 나오며 지도 학습의 결과는 정답이 출력이 된다. 그래서 비지도 학습 결과를 지도 학습 데이터로 사용하기 위해서는 위 그림 8처럼 중간에서 비지도 학습 군집화 결과에 레이블링을 해주기 위한 방법이 필요 하다. 본 논문에서는 비지도 학습과 지도 학습을 혼합모델을 만들기 위한 방법으로 영역 전문가 확인 시스템을 제안한다.

그림 8의 혼합모델의 구성은 BLD, PRO, GLU, LEU 4가지의 소변 검사 성분을 수집하여 스펙트럼 데이터를 얻은 후에 혼합 모델의 비지도 학습 K-means, 영역 전문가 확인 시스템, 지도 학습으로 구성이 된다.

소변 검사 결과 데이터 수집 하여 A~W의 18 개 채널 스펙트럼 데이터를 얻어 혼합모델의 첫 번째 부분인 비지도 학습은 한 데이터가 18개이며 있으며 K-means 군집화를 사용 하여 K=2, 즉 2가지로 군집화를 하여 다음 혼합모델인 영역 전문가 확인 시스템으로 결과 값을 보낸다.

영역 전문가 확인 시스템에서는 비지도 학습 K-means 군집화 결과를 영역 전문가-의사가 확인하여 BLD, PRO, GLU, LEU의 각 성분의 미검출과 검출로 구분을 한다. 그리고 구분한 데이터를 군집화가 성공한 데이터 와 군집화가 실패한 데이터를 실험하기 위한 2가지 형식의 데이터로 분류를 한다. 아래 그림 9 처럼 군집화가 성공한 데이터의 경우는 전문가 확인 시스템에서 군집화 결과가 정답인 데이터만 지도 학습으로 전달 후 진행을 한다.

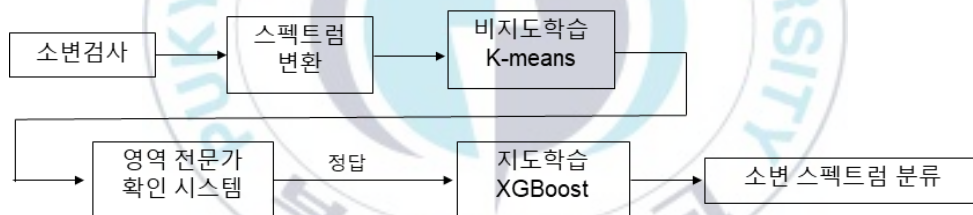


그림 9. 제안시스템 - 군집화 성공 데이터를 이용한 학습 흐름도

군집화가 실패한 데이터의 경우에는 아래 그림 10 처럼 영역 전문가 확인 시스템에서 오류 부분 찾아 미검출을 검출로 군집한 결과와 검출을 미검출로 군집한 결과를 수정 하여 다음 지도 학습으로 전달을 한다.

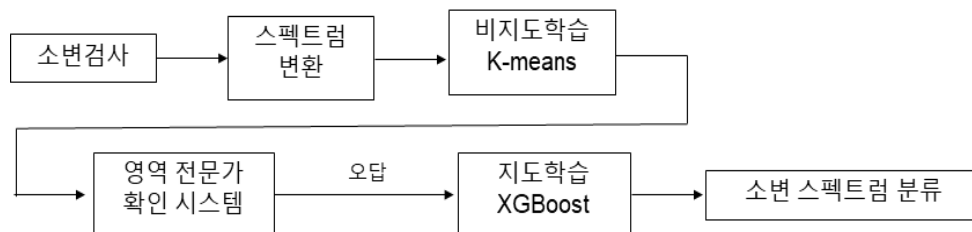


그림 10. 제안시스템 - 군집화 실패 데이터를 이용한 학습 흐름도

마지막 지도학습에서 군집화 결과 값을 앙상블 모델의 XGBoost 이용하여 0,1 이진 분류를 하여 최종적으로 소변의 각 성분이 검출 또는 미검출 소변 스펙트럼 분류 결과를 얻는다.

아래 그림 11 은 제안시스템의 예측 모델 흐름도이며 학습 모델에서 비지도 학습 후 영역 전문가 확인 시스템에서 선정된 데이터를 지도학습을 이용하여 얻은 가중치를 기반으로 예측을 한다.

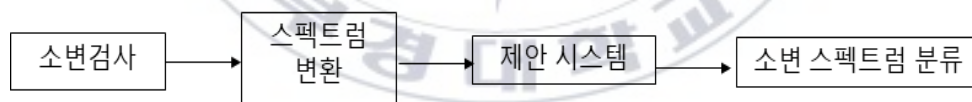


그림 11. 제안시스템 예측 흐름도

IV. 실험 및 평가

실험 환경은 CPU i7-8700, 메모리 16GB 사용하였으며 운영체제는 윈도우 10 아나콘다의 컴퓨터 노트북에서 텐서플로우 2.0 버전을 사용 하여 실험을 하였다.

표 8 은 각 성분별 마다 비교 평가를 위해 표 7 의 측정 도표 기반으로 평가를 한다. 평가 방식은 True Positive(TP) : 실제 검출인 정답을 검출이라고 예측, False Positive(FP) : 실제 미검출인 정답을 검출 이라고 예측, False Negative(FN) : 실제 검출인 정답을 미검출 이라고 예측, True Negative(TN) 실제 미검출인 정답을 미검출 이라고 예측. 이렇게 총 4가지로 평가를 하였으며 True Positive를 Type 1, False Positive를 Type 2, False Negative를 Type 3, True Negative를 Type 4로 작성하여 평가표 표 7을 설명한다.

가장 먼저 이진분류 DNN, XGboost, 2가지의 제안시스템을 비교 평가를 진행한다. 제안시스템의 경우는 비지도 학습에서 정상적으로 군집화된 데이터는 출력이 되기 때문에 다른 평가 모델과 전체 데이터수가 달라 병원에서 중요하게 생각하는 오진을 각 모델마다 추가로 평가를 하였다.

각 성분의 DNN, XGBoost, 제안시스템을 비교하면 전체적으로 좋은 결과를 보여주고 있다 하지만 XGBoost와 군집화 성공 과 군집화 실패 제안시스템의 GLU 성분을 제외하고는 낮은 오진율을 보여주고 있으며 제안시스템이 XGBoost 보다 낮은 오진율을 보여주고 있다.

마지막으로 군집화 성공 과 군집화 실패 2가지 제안시스템은 BLD를 제외한 나머지 성분에서 성공한 군집화 데이터를 사용한 경우가 실패한 군집화 데이터 보다 좋은 성능을 보여주고 있다.

이로 인해 지도학습만 사용하는 것 보다 비지도 학습과 지도학습을 같이 사용한 모델이 더 좋은 성능을 보여주는 것을 알 수 있다. 그리고 제안시스템에서 2가지 방법으로 연구를 하였으며 군집화 성공 데이터를 사용한 제안시스템은 일반적인 비지도 학습 사용법인 군집화가 되지 못한 노이즈 데이터는 학습에 필요가 없는 데이터이기 때문에 제외를 하고 학습을 하는 방법이다.

다른 하나의 제안 모델은 군집화가 실패한 데이터를 사용하여 군집화 하지 못하는 데이터를 전문가 확인 통하여 사용 가능한 데이터로 만들어 실험을 하였다.

이 실험 결과로 오진율을 최소화 할 수 있는 방법은 군집화 실패 데이터를 사용 하는 것 보다 군집화 성공 데이터를 사용하는 것이 좋은 성능을 보여 줄 수 있다.

표 7 측정 도표

		실제 정답	
		검출(1)	미검출(0)
분류 결과	검출(1)	TP : Type 1	FP : Type 2
	미검출(0)	FN : Type 3	TN : Type 4

Type 1-True Positive(TP) : 실제 검출(1)인 정답을 검출(1)라고 예측 (정답)
 Type 2-False Positive(FP) : 실제 미검출(0)인 정답을 검출(1)라고 예측 (오답)
 Type 3-False Negative(FN) : 실제 검출(1)인 정답을 미검출(0)라고 예측 (오답)
 Type 4-True Negative(TN) : 실제 미검출(0)인 정답을 미검출(0)라고 예측 (정답)

표 8 각 성분별 통계

BLD(혈뇨, Blood)

	DNN	XGboost	제안시스템 1 (군집화 성공)	제안시스템 2 (군집화 실패)
Accuracy	97%	98%	100%	100%
오진율	2.3%	1.7%	0%	0%
Type 1	1812	1655	946	898
Type 2	304	6	0	0
Type 3	9	226	0	0
Type 4	11737	11975	7241	4778

PRO(요단백, Protein)

	DNN	XGboost	제안시스템 1 (군집화 성공)	제안시스템 2 (군집화 실패)
Accuracy	98%	99%	100%	99%
오진율	1.2%	0.2%	0%	0.1%
Type 1	800	819	466	403
Type 2	84	1	0	0
Type 3	78	21	0	1
Type 4	12900	13021	7780	5212

GLU(포도당, Glucose)

	DNN	XGboost	제안시스템 1 (군집화 성공)	제안시스템 2 (군집화 실패)
Accuracy	99%	99%	100%	99%
오진율	0.6%	0.1%	0%	0.1%
Type 1	382	393	266	148
Type 2	85	0	0	1
Type 3	4	1	0	0
Type 4	13391	13468	8019	5429

LEU(류코사이트, Leukocyte)

	DNN	XGboost	제안시스템 1 (군집화 성공)	제안시스템 2 (군집화 실패)
Accuracy	98%	99%	100%	99%
오진율	1.7%	0.7%	0%	0.1%
Type 1	938	818	442	533
Type 2	237	0	0	1
Type 3	2	99	0	1
Type 4	12685	12945	7651	5234

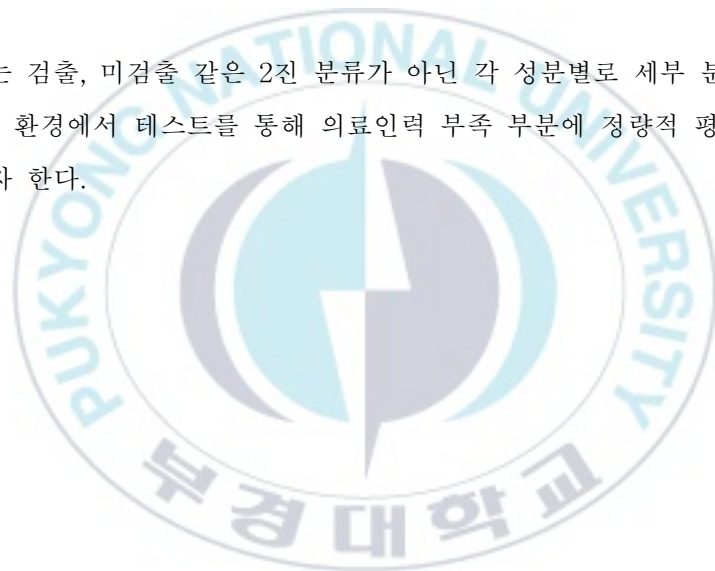
이 실험 결과로 의료인력 부족으로 인한 기존 소변검사를 간략하게 할 수 있을 것이며 또한 영역 전문가 확인 시스템으로 간호사의 주간적인 검사를 최소화 하고 장시간이 걸렸던 전자의무기록 공유시간도 단축 할 수 있다. 마지막으로 제안시스템, 혼합모델을 사용하여 낮은 오진을 보여 현재보다 높은 의료 서비스를 제공 할 수 있을 것이다.

V. 결론 및 향후연구

본 논문에서는 병원의 소변검사 문제점으로 정보 공유 시간 지연, 의료인력 부족, 간호사의 주관적 소변검사 결정을 해결하기 위해 소변 스펙트럼 분석 시스템을 제안 하였다. 소변 스펙트럼 분석 시스템은 소변 검사 데이터를 가지고 스펙트럼으로 변환 ,비지도학습 : K-means, 영역 전문가 확인 시스템 그리고 지도학습 : XGBoost으로 구성을 하고있는 혼합모델을 사용하였다. 지도 학습만을 사용하는 혼합 모델이 아니라 비지도 학습과 지도 학습을 사용하여 지도 학습 보다 좋은 결과를 보여주었다 또한 비지도 학습의 군집화 결과를 지도 학습을 위한 데이터로 변환 하는 영역 전문가 시스템을 제안 하여 비지도 학습과 지도 학습이 같이 사용하는 혼합모델을 제시 하였다. 추가적으로 영역 전문가 확인 시스템에서 군집화 성공한 데이터와 군집화 실패한 데이터를 각각 실험을 하였고 군집화가 실패한 데이터의 경우에는 전문가 영역 확인 시스템에서 수정을 하고 지도 학습으로 보내며 군집화가 성공한 데이터의 경우 전문가 확인 시스템에서 정답인 데이터만 지도 학습으로 보낸다. 군집화는 데이터에서 필요 없는 노이즈를 찾아 노이즈 제거 하여 데이터 효율을 향상 시키는 방향이 있다. 군집화가 실패한 데이터, 노이즈를 이용한 방식은 병원 환경에서 검사 결과의 군집화가 실패한 경우는 오진이다. 그래서 실패한 데이터를 다시 학습을 하여 비교 평가하였다. 실험 결과는 군집화 성공 데이터가 군집화 실패 데이터 보다 낮은 오진율을 보여주어 실 환경 병원에서는 군집화 성공 데이터를 사용하는 것이 기존 보다 더 좋은 의료 서비스를 제공할 수 있을 것이다.

또한 간호사가 직접 병실을 돌아다니면서 소변검사를 함으로서 생긴 정보 공유시간 지연 문제와 의료 인력 부족 문제를 기계학습을 이용하여 소변 검사를 간략하게 만들었고 간호사가 주관적으로 소변검사 결정을 하는 부분을 제안시스템의 영역 전문가 확인 시스템을 통하여 해결을 하였다. 또한 기계학습 적으로 소변검사 데이터가 지도학습으로만 예측하는 것 보다 혼합모델을 사용할 경우 진료 성공을 높일 수 있음을 연구를 통해 알 수 있었다

향후에는 검출, 미검출 같은 2진 분류가 아닌 각 성분별로 세부 분류 하여 실제 의료 환경에서 테스트를 통해 의료인력 부족 부분에 정량적 평가를 수행하고자 한다.



참고문헌

- [1] Minghao Piao, 변정용, “SVM 을 이용한 건강검진정보 기반 진료과목 예측”, 한국정보처리학회 제 6 권, 5 호, pp. 303-308, 2017
- [2] 김완철, 김호철, 서지영 외 “F-133 중환자실 입원 첫 24 시간 소변량이 예후에 미치는 영향”, 대한결핵및호흡기학회 추계학술대회초록집, 제 116 차, pp. 173, 2013.
- [3] 이태화, 강경화, 고유경, 조성현, 김은영, “국내외 간호사 인력정책의 현황과 과제”간호행정학회지, 제 20 권, 1 호, pp. 106-116, 2014
- [4] 질병관리본부 국가건강정보포털 KCDC, 건강 질병 검색 “소변검사“;
<https://health.cdc.go.kr/health/HealthInfoArea/HealthInfo/View.do?idx=850>
- [5] ams company AS7265x : “Smart 18-Channel VIS+NIR Spectral_ID Sensor with Electronic Shutter datasheet”; <https://ams.com/ko/spectral-sensing>, 2018
- [6] CHRISTOPHER M. BISHOP Pattern Recognition and Machine Learning, pp. 1-4
- [7] Chen, Tianqi, and Carlos Guestrin. “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
- [8] 국가수리과학연구소 현윤경 박사 “2019 의료수학 문제해결 사례 공유 세미나-기계학습 기반의 의료데이터 분석과 예측” 2019,
- [9] Akyol, Kemal. “Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection.” Expert Systems with Applications 148 ; 2020

감사의 글

어느덧 2년의 석사과정을 마치고 학위 논문을 제출하게 되었습니다. 지난 시간을 돌이켜보니 많은 아쉬움과 후회가 남습니다. 학업적 성취에 있어서의 아쉬움만이 아닌, 고마운 많은 분께 감사의 마음을 제대로 전하지 못했기에 더욱 그러한 것 같습니다. 그동안 힘이 되고 방향을 잡아주셨던 많은 분께 이 자리를 빌려 감사의 말씀을 전하고자 합니다.

가장 먼저 석사과정 시작부터 지금까지 세심한 지도와 많은 격려로 이끌어 주신 정목동 교수님께 진심으로 감사드립니다. 석사과정 중에 어려움이 있었고 이러한 부분이 있을 때 현명하게 지도하여 주신 정목동 교수님께 어떻게 감사의 말씀을 드려야 될지 어려울 정도입니다. 교수님, 감사합니다.

연구의 의료적 부분을 도와주신 서부산 C병원의 센터장님에게 깊은 감사를 드립니다. 기계학습 모델의 조언을 해주신 닥터스랩의 직원분들에게 감사를 표하고 싶습니다. Lab. 식구들에게도 감사를 표하고 싶습니다. 자주 만나지는 못하지만, 항상 격려와 웃음으로 응원을 주었던 친구들에게도 감사의 마음을 전합니다.

마지막으로 항상 사랑으로 키워주시고 부족한 자식을 믿어주신 부모님께 감사의 말씀을 드립니다.

2020년 7월 20일

최재혁 드림