JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 1: 399~406, Mar. 2019

딥러닝과 머신러닝을 통한 당뇨병 데이터 분석

김재호[1] · 김장영[2*]

Diabetes data analysis through deep learning and machine learning

Jae-Ho Kim[1] · Jang-Young Kim[2*]

[1]Graduate Student, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea
[2*]Associate Professor, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea

요 약

당뇨병은 널리 퍼진 만성 질환으로, 세계적으로 영향을 미치고, 경제적으로도 상당한 재정적 부담을 부가하고 있다. 당뇨병은 혈액의 포도당을 조절하는 능력을 저하하고 , 삶의 질과 수명을 감소시킬 수 있는 만성 질환이다. 또한 당뇨병은 인슐린을 생성하지 못하거나, 효과적으로 사용하지 못한다. 당뇨병 문제의 규모는 상대적으로 크지만, 쉽게 인식하지 못한다. 당뇨병을 완치할 수 있는 방법은 없지만 체중 감량, 건강식, 활동적 생활 및 치료를 받는 것과 같은 전략은 많은 환자에서 이 질병의 피해를 완화할 수 있다. 조기 진단은 생활 방식의 변화와 보다 효과적인 치료로 이어진다. 본 연구의 의의는 당뇨병이 있는지 여부에 대한 정확한 예측을 제공하고 당뇨병 위험을 가장 잘 예측하는 위 험 요소는 무엇인지 찾는 것이다. 예측에 있어서 여러 가지 머신러닝 기법과 딥러닝의 CNN과 RNN을 통한 Ensemble Model 사용하고, 평가방법으로 Accuracy와 Recall을사용한다. 이 Ensemble Model은 Transformer구조를 따르고자 했고, 경량화 하였다.

● ABSTRACT

Diabetes is a widespread chronic disease that affects people worldwide and imposes significant financial burdens. Diabetes impairs the ability to regulate blood glucose levels, reducing quality of life and life expectancy. Additionally, diabetes is characterized by either the inability to produce insulin or to use it effectively. Despite its prevalence, diabetes is often underrecognized. While there is no cure for diabetes, strategies such as weight loss, healthy eating, an active lifestyle, and treatment can mitigate the disease's impact in many patients. Early diagnosis leads to lifestyle changes and more effective treatment. This study aims to provide accurate predictions for diabetes and identify the most significant risk factors for its development. The study employs various machine learning techniques, and ensemble models using CNN and RNN, with accuracy and recall as evaluation metrics. This Ensemble Model attempted to follow the Transformer structure and made it lightweight.

키워드 : 당뇨병, 분류, 머신러닝, 딥러닝

Keywords : Diabetes , Classification, Machine Learning, Deep Learning

Received 29 January 2019, Revised 29 March 2019, (출판사에서작성) Accepted 21 April 2019

* Corresponding Author Jang-Young Kim(E-mail:jykim77@suwon.ac.kr, Tel:+82-31-229-8345)
Associate Professor, Department of Computer Science , The University of Suwon, Hwasung, 18323 Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.1.399 pISSN:2234-4772

Ⅰ. 서 론

당뇨병은 지방과 단백질의 대사 장애를 동반한 혈당치의 상승이 특징인 만성 질환이다. 혈당은 췌장에 의한 인슐린 생산의 부족이나 세포가 생산 중인 인슐린을 효과적으로 사용할 수 없기 때문에 세포에서 대사될 수 없기 때문에 상승한다. 당뇨병의 주요 유형은 췌장에서 인슐린이 생성되지 않는 제1형, 체세포가 생성 중인 인슐린의 작용에 저항성을 가지며 시간이 지남에 따라 인슐린의 생산이 점진적으로 감소하는 제2형, 임신 중에 발생하여 합병증을 일으킬 수 있는 임신성 당뇨이다. 임신 중, 그리고 출생 시 산모의 제2형 당뇨병과 자손의 비만의 위험을 증가시킨다. 통제되지 않는 당뇨병은 많은 장기에 합병증을 일으킨다. 크고 작은 혈관과 신경이 손상되면 시력과 신장 기능 상실, 심장마비, 뇌졸중, 하지 절단 등이 발생한다. 당뇨병은 장애를 일으키고 수명을 단축시킨다. 당뇨병은 한국의 사회적 전염병으로 우리에게 다가왔다. 유병률 한국의 당뇨병은 지난 30년 동안 1.5%에서 7~9%로 5~6배 증가했다[1]. 이러한 증가율은 우리나라 등 선진국에 비해 현저히 높다. 또한 미국은 지난 30년 동안 두 배로 늘었다. 세계적으로 당뇨병으로 인한 사망률은 지난 20년 동안 급격히 증가했다. 당뇨병의 직간접적인 비용을 모두 포함한 손실은 총 1조 7천억 달러로, 중저소득 국가들에게는 8천억 달러를 차지할 것으로 추산된다. 의료 시스템과 국가 경제에 대한 경제적 부담 외에도, 당뇨병은 종종 장애와 조기 사망으로 인한 소득 손실과 현금 지급으로 인한 치명적인 개인 지출을 초래한다. 이에 따라 국내에서는 의료비 국민건강보험공단이 보장하는 당뇨병은 3조2000억원, 전체 의료비의 19.2%를 차지했다. 한편, 인지도와 치료율은 1998년에서 2005년으로 당뇨병 환자의 비율이 개선되다. 그러나 적절한 비율 치료받은 당뇨병 환자의 혈당 조절(HbA1c < 7%)은 40%에 불과했다. 공공 접근을 포함한 포괄적이고 통합된 보건 개입이 시급히 필요하다. 당뇨병의 유병률 증가 및 이와 관련된 바람직하지 않은 결과를 제어한다[2]. 본 연구의 의의는 당뇨병이 있는지 여부에 대한 정확한 예측을 제공하고 당뇨병 위험을 가장 잘 예측하는 위 험 요소는 무엇인지 찾는 것이다.

본 연구에서 사용된 데이터 세트는 CDC의 BRFSS2015에서 가져온 데이터로, 253,680개의 설문조사 응답으로 구성되어있으며, 데이터 세트의 목적은 데이터 세트에 포함된 특정 진단 측정값을 기반으로 환자에게 당뇨병이 있는지 여부를 진단적으로 예측하는 것이다. 데이터 세트의 column은 총 22개로 1개는 당뇨병 유무, 나머지 21개는 특성변수로 이루어져 있다. 데이터의 column구성은 Diabetes_binary, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income 으로 총 22개다. Diabetes_binary같은 당뇨병 유무를, High BP는 고혈압 유무를, HighChol은 고콜레스테롤을, CholCheck은 최근 5년간 콜레스테롤 검사 유무를, BMI는 체질량 수를, Smoker은 일생에 담배를 100개 이상 흡연했는지, Stroke는 뇌졸중 유무를, HeartDiseaseorAttack 은 심장질환 및 심근경색증 유무를, PhysActivity 은 지난 30일간 신체활동 유무를, Fruits은 하루 1번이상 과일섭취 유무를, Veggies는 하루 1번이상 야채섭취 유무를, HvyAlcoholConsump은 과음의 유무를, AnyHealthcare은 건강보험을 포함한 모든 종류의 의료보험 유무를, NoDocbcCost은 아팠지만 비용에 의해 병원가지 못한 유무를, GenHlth는 일반적인 건강상태를,, MentHlth는 스트레스, 우울증, 감정문제등을 포함한 정신건강을, PhysHlth은 신체적 질병 및 부상을 포함한 신체적 건강을, DiffWalk는 계단을 오르는데의 어려움을, Sex은 성별을, Age를 13단계의 범주로 나눈것을, Education은 교육수준을, Income은 수입을 뜻한다[3].

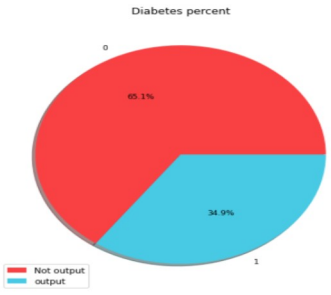

Fig1. Data heart diabetes percentage

위 그림1은 데이터의 당뇨병 퍼센테이지를 보여준다. 설문 응답자 수는 253,680이며, 당뇨병 환자는 86.1%인 데이터다.

Ⅱ. 기존연구

2.1 ADAP학습 알고리즘을 사용한 당뇨병 발병 예측

병렬 처리를 위한 신경망 또는 연결 모델은 새로운 것이 아니다. 그러나 지난 50년 동안 관심의 부활이 일어났다. 부분적으로 이것은 현재 숨겨진 노드라고 하는 것에 대한 더 나은 이해와 관련이 있다. 이러한 알고리즘은 패턴 인식 문제에서 현저한 가치가 있는 것으로 간주된다. 이 때문에 우리는 초기 신경망 모델인 ADAP가 Pima 인디언의 고위험군에서 당뇨병 발병을 예측하는 능력을 테스트했다. 알고리즘의 성능은 민감도, 특이도 및 수신기 작동 특성 곡선과 같은 임상 테스트에 대한 표준 측정값을 사용하여 분석되었다. 민감도와 특이도의 교차점은 0.76이다. 우리는 현재 정확하게 동일한 훈련 및 예측 세트를 사용하여 로지스틱 회귀 및 선형 퍼셉트론 모델에서 얻은 결과와 ADAP결과를 비교하여 이러한 방법을 추가로 조사하고 있다. 알고리즘에 대한 설명이 포함되어 있다[4].

Ⅲ. 분석 알고리즘

3.1 KNN

KNN(K-Nearest neighbor)은 학습에 이용되지 않은 새로운 데이터를 받았을 때 기존 Cluster에서 모든 데이터와 instance거리(데이터와 데이터 사이의 거리)를 측정한 후 가장 많은 속성을 가진 클러스터에 포함하는 지도학습의 Classification이다. 따라서 과거의 데이터를 저장해두고 필요할 때 비교를 수행한다. 이 값은 K의 값에따라 분류의 결과값이 달라진다. 여기서 K개로 분류한다는 뜻을 가진다.

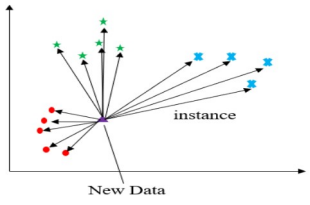

Fig2. KNN Algorithm

위 그림2는 KNN을 그림으로 나타낸 것이다. 새 데이터가 기존 데이터들과 하나씩 거리를 계산하고 거리상 가까운 데이터를 선택하여 할당된다[5].

3.2 DNN(Deep Neural Network)

심층신경망은(DNN) 2개 이상의 은닉층을 포함하는 인공 신경망을 말한다. 스스로 분류하는 labeling을 만들고 많은 학습 데이터와 반복학습, backpropagation등을 통해 머신러닝은 비선형데이터를 다루기 위해 여러 가지 기술들이 도입되었지만, DNN의 경우에는 다수의 은닉층이 있기 때문에 별도의 기술없이 비선형 분류가 가능하다. 은닉층의 경우 입력받은 데이터 x에 대하여 가중치를 부여하고, 활성화 함수를 통해 출력된 값을 채택할지 판단하고 최종적으로 y를 출력한다. 은닉층이 여러 개로 구성되면 보다 정확할수 있지만, Overfitting에 취약하다. 변수 형태에 상관없이 분석이 가능하며, 비선형적 데이터를 학습할수 있고 수 있는 장점이 있지만, 학습에 있어서 연산량이나 Gradient Vanishing과 같은 문제들이 발생한다[6].

3.3 Ensemble Neural Network(RNN + CNN)

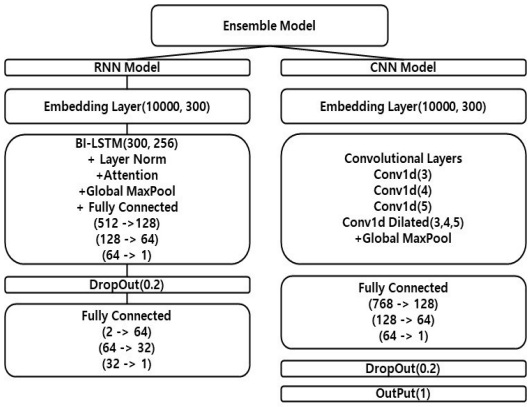

Fig3. Columns Correlation

위 그림은 본 연구에서 제시하는 Ensemble 모델의 구조다. RNN model과 CNN모델로 구성되어 있으며,두 모델 모두 같은 Embedding Layer를 사용한다. 입력 텍스트를 고차원 벡터로 변환한다. RNNModel에서는 LSTM 방식을 이용하며, 두 개의 양방향 LSTM 레이어를 사용하고, dropout이 적용된다. LayerNorm에서 LSTM 출력에 레이어 정규화

가 적용되고 Attention에서는 Multihead Attention이 적용되어 정보가 강조된다. GlobalMaxPool에서는 LSTM의 출력을 최대 풀링하여 벡터 크기를 줄이고, Fully Connected Layers: 최종 출력을 위한 여러 선형 레이어가 존재한다. 마지막으로 과적합 방지를 위한 Dropout으로 구성된다. CNNModel은 Convolutional Layers은 여러 크기의 커널을 가진 컨볼루션 레이어를 포함한다.Dilated Convolutions은 수용 필드 확장을 위한 팽창 컨볼루션을 사용한다. GlobalMaxPool: CNN 출력에 대해 최대 풀링을 적용하고다. Fully Connected Layers: 최종 출력을 위한 여러 선형 레이어가 포함된다. 마지막으로 Dropout: Dropout이 적용되어 있다. 마지막 Ensemble 모델은 Concatenation하여 RNNModel과 CNNModel의 출력을 결합한다. 이 구조는 RNN과 CNN의 장점을 결합하여 더 강력한 성능과 일반화능력, 하이브리드 특징 추출을 통해 더욱더 복합적인 특징을 정교하게 추출하는 것을 목표로 한다[7]. 이 모델은 Transformer구조는 아니지만, 그와 비슷하게 하며, 경량화하기위해 노력하였다.

IV.실험결과 및 분석

본 연구에서 성능이 제일 좋았던 모델 Ensemble Model(RNN+CNN)에서, 기본적으로 Embedding layer를 통과시켰고, RNN에서는 BI-LSTM을 이용했다. LSTM Layer자체는 2개만 쌓았다. Attention으로는 Multihead Attention을 사용했고, features은 고정하였다. 그 이후 Max Pooling을한 후에 Fully Connected Layer를 통과시켰다. CNN에서는 Conv1D를 사용하였고, Sequential Model을 이용해 분류한다. Kernel size는 3, 4 5 로 늘려가며, stride는 1로 layer를 쌓았고, Max Polling 후에 Fully Connected Layer를 통과시켰다. 풀링층에서는 에서는 각 커널 값을 평균화 시켜 중요한 가중치를 갖는 값의 특성이 희미해질수 있기 때문에 max polling을 사용했다. 훈련에 있어서 loss function으로는 binary cross-entropy를 사용했으며, optimizer으로는 Adam을 사용했다. 출력 Activation function으로는 sigmoid 함수를 사용하였다.

4.1 탐험적 데이터분석(exploratory data analysis)

EDA에 있어서 본 데이터에서 상관관계표가 음수인 Income, Education, PhysActivity, HvyAlcoholConsump, Veggies, Fruits 들의 Column들은 제외를 시킨 후 데이터를 학습시켰다.

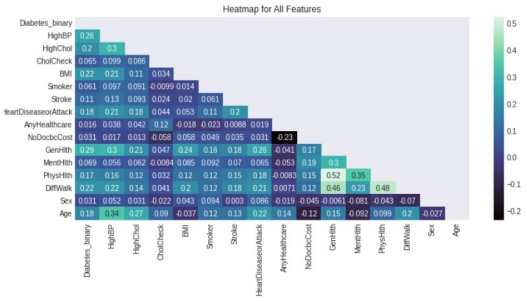

Fig3. Columns Correlation

위 그림4은 서로 다른 column간 상관관계를 보여준다.

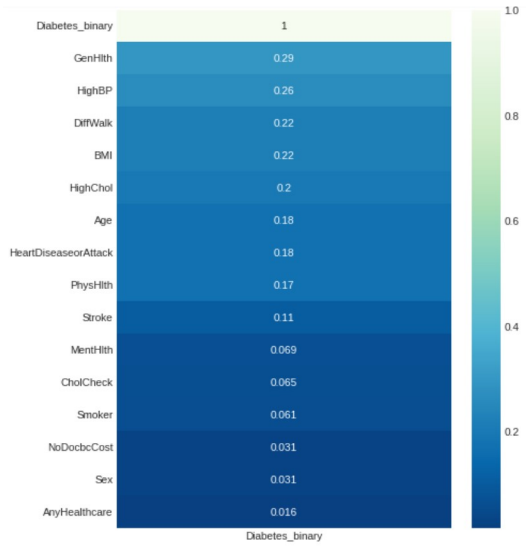

Fig5. Diabetes Correlation

위 그림5는 당뇨병과 다른 column간 상관관계 사진이다. 본 연구에서 Outcome과 높은 상관관계를 가지는 데이터는 0.2 이상들로 GenHlth, HighBP, DiffWalk, BMI, HighChol이 인 것을 확인할 수 있다. 상관관계가 높지 않으므로, 본 연구의 데이터는 다중공선성에 대해 안전하다는 근거가 될 수 있다[8]. 보통, 상관계수가 0.7이상이면 다중공선성이 있다고 판단된다. 여기서, 다중공선성이란, 독립변수간 상관관계가 있는지 나타내는 지표이다.

Table.1 VIF(Variance Inflation Factors)

Features	VIF Factor
const	116.856706
Diabetes_binary	1.193120
HighBP	1.344502
HighChol	1.180932
CholCheck	1.033501
BMI	1.160280
Smoker	1.091872
Stroke	1.081612
HeartDiseaseorAttack	1.175776
PhysActivity	1.157396
Fruits	1.112540
Veggies	1.112397
HvyAlcoholConsump	1.025418
AnyHealthcare	1.113209
NoDocbcCost	1.144200
GenHlth	1.821914
MentHlth	1.239497
PhysHlth	1.623288
DiffWalk	1.536636
Sex	1.075748
Age	1.354954
Education	1.326495
Income	1.505649

위 표 1은 각 컬럼별 VIF값이다. 다중공선성이 적다는 근거로 이용할수 있는 VIF(분산팽창요인) 값이다. const값은 상수항으로, 무시해도 좋다. 학자들마다, VIF값에따른 다중공선성 기준은 학자마다 다르지만, 이 값이 5 미만이면 다중공선성에 크게 문제가 없다고 판단한다. 모두 1점대를 보여주면서 다중공선성이 낮다는걸 보여준다. 이것은, 독립변수의 영향을 더 명확하게 해석할 수 있다[9].

4.2 Accuracy와 Recall

Table.1 Accuracy

Model	Accuracy
KNN	0.875
XGB	0.864
ADA	0.864
SVC	0.863
LR	0.861
Ridge	0.861
RF	0.850
BNB	0.823
DT	0.816
GNB	0.778
CNN	0.865
DNN	0.866
Ensemble	0.892

위 표2은 각각 모델별 Accuracy를 보여준다. 값들은 소수점 세 번째 자리까지만 표현한다.

Table.3 Recall

Model	Recall
GNB	0.687
BNB	0.672
DT	0.596
RF	0.580
KNN	0.554
XGB	0.570
ADA	0.578
LR	0.563
SVC	0.535
Ridge	0.516
CNN	0.67
DNN	0.69
Ensemble	0.79

위 표 3는 각 모델별, Recall값을 나타낸다. 질병에 있어서 Recall은 매우 중요한 지표가 될 수 있다. 왜냐하면 질병에 걸리지 않았는데 질병에 걸렸다고 예측하는 것 보다, 질병에 걸린 환자를 질병에 걸리지 않았다고 놓치는 것은 매우 위험하기 때문이다. 즉, 질병에 있어서는 실제 양성인데 음성으로 예측하는 것은 매우 치명적이라는 것이다. 이것은 과잉진단을 감수하더라도, 모든 잠재적 환자를 식별하는 것이 중요하다. Recall의 수식은 다음과 같다[10].

Ⅴ. 결론 및 향후연구

본 연구에서는 diabates 데이터를 통해서 여러 가

지 머신러닝기법들과 딥러닝의 성능을 분석한다. 이 중, Ensemble 알고리즘이 가장 좋은 결과값을 가진다. 이것은 경량화 모델을 고려할 때 의미가 있다.

본 연구는 CNN + RNN계열이 활용방법을 더 방대하게 만드는데 도움을 줄 수 있다. 이미지 처리 및 분류, 컴퓨터비전, Sequence model, 자연어처리등 비정형 데이터에 주로 사용하는 모델이 정형적인 범주형 데이터를 binary classification에서 사용했을 때, 결과값은 Accuracy는 0.89, Recall은 0.79라는 결과값을 가지므로 충분히 고려할 수 있는 모델임을 증명한다.

마지막으로, 비용적인 부분에서, 학습시간도 짧았다. 가장 큰 의미를 가지는 부분은, Large Model이 주목받는데, 엄청난 비용이 요구되기 때문에, 일반 개인 연구자들은 이러한 작은 Model에 관심을 가질 수 밖에 없다, 가령, 라즈베리파이와 같은 저사양의 장치를 사용할 경우에 큰 모델을 사용하기에는 제약이 생긴다. 그렇기 때문에 파인튜닝을 하는 SLM모델이 많이 사용되는데, 그런 의미에서 본 연구에서 제시된 Ensemble 알고리즘도 예측분야에서는 충분히 고려할 수 있다. 즉, 모델의 크기가 작고 높은 성능을 가지는 것을 요구하는데 이 연구에서는 높은 정확도를 가지는 알고리즘에 대해 탐구하는데 의의가 있다.

REFERENCES

[1] Gojka Roglic, “WHO Global report in diabetes: A summary”, *International Journal of Noncommunicable Diseases*, vol.1, pp. 3-8, Jun. 2016. DOI: 10.4103/2468-8827.184853

[2] S. G. Kim, D. S. Choi, “The present State of Diabetes Mellitus in Korea”, *Journal of the Korean Medical Association*, vol.51, pp. 791-798, Sep. 2008. DOI: https://doi.org/10.5124/jkma.2008.51.9.791

[3] Kaggle Data set, Diabetes Prediction Dataset, 2022. [Internet].Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download

[4] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler , R. S. Johannes, “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus,” *Proceedings Annual Symposium on Computer Application in Medical Care(Proc Annu Symp Comput Appl Med Care)*, pp. 261-265, Nov. 1988.

[5] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, H. Ghayvat, “CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope,” *Electronics*, vol.10, Oct, 2021. DOI:10.3390/electronics10202470

[6] Y. Wang, J. Liu, J. Mišić, V. B. Mišić, S. Lv, X. Chang, “Assessing Optimizer Impact on DNN Model Sensitivity to Adversarial Examples,” *IEEE Access*, vol.7, pp. 152766-152776, 2019. DOI:10.1109/access.2019.2948658

[7] X. Li, G. A. Ng, F. S. Schlindwein, "Convolutional and recurrent neural networks for early detection of sepsis using hourly physiological data from patients in intensive care unit", *2019 Computing in Cardiology (CinC)*, IEEE, 2019. DOI: 10.22489/CinC.2019.054

[8] S. Streukens, S. Leroi Werelds, "Multicollinearity: an overview and introduction of ridge PLS-SEM estimation", *Partial Least Squares Path Modeling: Basic Concepts, Methodological Issues and Applications*, pp.183-207, 2023 DOI:10.1007/978-3-031-37772-3_7

[9] N. Shrestha, “Detecting multicollinearity in regression analysis”, *American Journal of Applied Mathematics and Statistics*, vol.8, no.2, pp.39-42, 2020. DOI:10.12691/ajams-8-2-1

[10] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases", *Diagnostics*, vol.14, 2024. DOI:10.3390/diagnostics14020144

김재호(Jae-Ho Kim)

수원대학교 컴퓨터학부 학사
수원대학교 컴퓨터학부 석사과정
※관심분야 : 인공지능

김장영(Jang-Young Kim)

연세대학교 컴퓨터과학 공학사
Pennsylvania State Univ. 공학석사
State University of New York 공학박사
University of South Carolina 교수
수원대학교 컴퓨터학부 교수
※관심분야 : Big data, AI, Cloud computing, Networks