JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 6, pp. 1191-1200, Jun. 2020

노인코호트 DB를 이용한 딥러닝 기반의 뇌졸중 질환 예측 모델

유 재 학[1] · 권 순 현[1] · 호치멍 벤자민[2] · 이 경 란[3] · 김 내 수[4] · 표 철 식[4] · 박 세 진[5*]
[1]한국전자통신연구원 KSB융합연구단 선임연구원, [2]한국표준과학연구원 안전융합사업팀 박사후연구원
[3]국민건강보험공단 급여전략실 부장, [4]한국전자통신연구원 KSB융합연구단 책임연구원
[5]한국표준과학연구원 안전융합사업팀 책임연구원

Stroke Disease Prediction based on Deep Learning using the Elderly Cohort DB

Jaehak Yu[1] · Soon-Hyun Kwon[1] · CheeMeng Benjamin Ho[2] · Kyeong-ran Lee[3] · Nae-soo Kim[4] · Cheol-Sig Pyo[4] · SeJin Park[5*]

[1]Senior Researcher, Department of KSB(Knowledge-converged Super Brain) Convergence Research, Electronics and Telecommunications Research Institute (ETRI), Daejeon, 34129, Korea

[2]Postdoctoral Researcher, Research Team for Health & Safety Convergence, Korea Research Institute of Standards and Science (KRISS), Daejeon, 34113, Korea

[3]Manager, Department of benefits strategy, NHIS(National Health Insurance Service), Wonju, 26464, Korea

[4]Principal Researcher, Department of KSBConvergence Research, ETRI, Daejeon, 34129, Korea

[5]Principal Researcher, Research Team for Health & Safety Convergence, KRISS, Daejeon, 34113, Korea

[요 약]

뇌졸중은 전 세계적으로 암과 심장질환 다음으로 발병하는 중요한 사망원인이며, 통계청의 사망 통계분석에 따르면 매일 70여 명의 사망자가 발생하고 있다. 특히, 2030년에는 인구 고령화로 인하여 뇌졸중과 관련된 질환 발생이 3배 이상 급증할 것으로 예상되고 있다. 따라서 뇌졸중 질환으로 사망과 진료비 부담을 줄이고 사회적 손실을 최소화하기 위한 연구가 절실히 요구되고 있다. 본 논문에서는 합성곱신경망(convolution neural network, CNN) 기반의 뇌졸중 질환 예측을 가능케 하는 새로운 모델을 설계 및 구현하였다. 본 논문에서는 국민건강보험공단에서 공개한 60세 이상의 고령자 코호트 558,147명의 데이터를 이용하여 뇌졸중 질환 예측 모델을 검증하였다. 실험을 통하여 한국인 고령자를 위한 뇌졸중 질환 예측 결과와 모델의 정확성을 확인하였다.

[Abstract]

Stroke is the leading cause of death worldwide after cancer and heart disease. According to a statistical analysis of deaths by Statistics Korea, about 70 deaths occur every day. By 2030, the incidence of stroke disease is expected to surge more than three times due to an aging population. Therefore, research is required to reduce the burden of death and medical expenses and minimize social loss due to stroke disease. In this paper, we designed and implemented a new model that enables CNN-based stroke disease prediction. The model for predicting stroke disease was verified by using data from 558,147 elderly over the age of 60 published by the NHIS(national health insurance service). Through this experiment, we confirmed the accuracy of the model and the prediction of stroke disease for the Korean elderly.

색인어 : 뇌졸중, 질환 예측, 딥러닝, CNN, 뇌졸중 심층 분석
Key word : Stroke, Disease Prediction, Deep Learning, Convolution Neural Network, Stroke In-depth Analysis

http://dx.doi.org/10.9728/dcs.2020.21.6.1191

Received 12 May 2020; **Revised** 15 June 2020
Accepted 25 June 2020

***Corresponding Author; SeJin Park**

Tel: +82-42-868-5450
E-mail: sjpark@kriss.re.kr

Ⅰ. 서 론

2018년 통계청이 사망원인 통계를 살펴보면 총 사망자 수는 298,820명이며, 남성은 161,187명, 여성은 137,633명으로 보고되었다[1]. 사망원인별로는 악성신생물(암)이 79,153명과 심장질환 32,004명, 폐렴으로 23,280명, 뇌혈관 질환으로 22,940명으로 보고되었다. 특히, 한국인의 경우 뇌혈관 질환으로 인한 사망률은 10만명 당 남성은 42.7명, 여성은 46.7명으로 높은 사망률을 보이고 있다[1]. 이러한 뇌혈관 질환으로 인한 사망자 수는 2005년 이후로 감소추세를 보이지만, 악성신생물을 제외한 단일 질환으로 사망원인 3위에 해당되는 고위험 질환이다. 특히, 60세 이상의 고령자는 순환계통 질환으로 대표되는 심장질환과 뇌혈관 질환으로 인한 사망률이 점차 증가하고 있다.

뇌졸중은 성인 및 고령자에게 기능 장애를 일으키는 가장 중요한 질환 중 하나로 장애 정도에 따라 사회적 또는 경제적 활동에 어려움이 생기는 치명적 질환 중 하나이다[2], [3], [4], [5]. 특히, 급성 뇌졸중 질환 발생 시 중추신경계와 자율신경계의 장애가 유발되어 심전도의 부정맥과 같은 장애가 발생될 수 있다. 이러한 뇌졸중은 장애 정도에 따라 사회적 또는 경제적 활동에 어려움이 생기는 치명적 질환이다[2], [3], [4]. 뇌졸중은 환자의 장애 양상 또는 동반되는 질환에 따라 다양하게 나타날 수 있다. 뇌졸중 환자는 개인별로 현재의 장애 수준을 정확히 평가하고 재활치료 또는 의료기관 방문을 유도해야 한다[4], [5]. 이러한 뇌졸중은 그 증상과 분류가 다양하므로, 뇌졸중으로 인한 장애 및 동반되는 신경학적 손상을 신뢰성 있게 평가하기가 어려운 실정이다. 특히, 뇌졸중 과거력이 없는 사람의 빠른 발견도 중요하지만, 과거력이 있는 사람의 경우 재발 확률이 9배가량 높다. 따라서 이들을 지속적으로 추적 관찰하여 빨리 의료기관 방문과 의료진에게 진단 및 치료를 받을 수 있도록 지원할 수 있는 기술이 절실히 요구된다.

최근의 연구문헌 조사에 의하면, 뇌졸중 환자의 재발을 예방하고 초기 장해를 평가하기 위해서는 환자의 상태를 추적 관찰하고 뇌졸중의 주요 위험요인을 발견하기 위한 연구가 성공적으로 진행 중이다[6], [7], [8]. 뇌졸중 환자의 장애를 평가하는 방법으로는, 1972년 Mathew scale이 처음 발표된 이후, 유럽 뇌졸중 척도, 캐나다 신경학적 척도, 미국 국립 보건원의 뇌졸중 척도(national institutes of health stroke scale, NIHSS) 등이 발표되었다[7], [8], [9]. 이 중에서, Brott 등에 의해 개발된 미국 국립보건원 뇌졸중 척도는 뇌졸중 발병 이후의 장애에 정량적 측정에 널리 쓰이는 도구이다. NIHSS는 의식 수준, 응시, 시각, 안면마비, 상지와 하지 운동, 사지 운동실조, 감각, 언어능력, 구음장애, 무시, 원위부 운동을 측정하는 총 14개의 항목으로 구성되며, 환자 한 명당 수행시간이 6.6분정도 소요된다. 그리고 환자의 입원 초기에 비교적 측정이 쉽고 간단하게 실시할 수 있는 평가 척도로 사용되고 있다. 이러한 NIHSS는 검사와 재검사 간 신뢰도 및 타당도가 다양한 연구를 통해서 이미 검증된 도구로 전 세계적으로 널리 쓰이고 있다. 하지만, NIHSS은 각 뇌졸중 환자의 장해를 총괄적으로 평가하는데 쓰이는 도구지만, 무엇보다 초기 장해를 평가하기 위한 정확한 예측정보 결과를 제공하지 못한다는 단점을 가지고 있다.

또 다른 연구방법으로는, 뇌졸중 질환의 주요 위험 요인으로 밝혀진 흡연, 고혈압, 당뇨병 등의 정보로 향후 10년 후의 뇌졸중 발생 가능성을 예측하지만, 다양한 요인들이 상호 작용하여 발생될 가능성을 배제하고 있다. 이러한 위험 요인 정보를 기반으로 통계적 방법과 기계학습 방법을 적용한 연구도 진행되고 있다. Kannel 등[10]의 로지스틱 모형, Cox의 비례위험모형(proportional hazards model)[11], [12], [13]이나, Weibull 모형[14]이 보고되었지만, 이러한 선행 연구들은 한국인의 뇌졸중 질환 발생에 대한 위험도를 예측하는데 적당하지 않다.

본 논문에서는 딥러닝의 대표적인 분류 및 예측 모델인 CNN을 기반으로 60세 이상 고령자의 뇌졸중 질환을 조기발견 및 예측할 수 있는 시스템을 제안한다. 시스템의 뇌졸중 질환 예측 정확도와 성능 검증을 위해, 국민건강보험공단의 노인코호트 DB(ver 1.0)를 사용하였다[15]. 노인코호트 DB는 2002년부터 2013년까지 총 558,147명의 코호트 데이터로 구성되며, 인구 사회학적 특성, 의료이용현황, 노인 장기요양 정보 등을 포함하고 있다. 본 논문에서는 한국표준 질병분류코드[16]에서의 주요상병 중 뇌졸중 질환 명세와 일치하는 I60~I69까지의 주상병 코드를 뇌졸중 질환으로 정의하였고, 이에 해당되지 않는 병명을 뇌졸중이 아닌 기타 질환의 일반 고령자로 정의하여 실험하였다. 결과적으로, 뇌졸중 질환 발병 후의 중증도 측정을 통한 초기장애 진단, 향후 10년 이후의 질환 발생 확률 값으로 제공하는 기존 선행연구의 한계점을 벗어나, 코호트 DB의 정보만으로 과학적이고 정확한 예측이 가능함을 검증하였다. 또한, 이러한 예측 모델은 고령자를 대상으로 의료진 또는 병원에서의 뇌졸중 질환에 대한 조기발견 및 예측함으로써, 오진률 수치를 줄이고 선제적으로 치료 및 대응이 가능하다는 추가적인 장점을 갖는다.

본 논문의 구성은 다음과 같다. 2장에서는 뇌졸중 질환의 정의 및 연구방법을 서술하고, 3장에서는 본 논문에서 제안하는 CNN 기반의 뇌졸중 질환 예측 및 분석 모델에 대해 기술한다. 4장에서는 실험결과 및 성능 분석을 상세히 설명하고, 마지막으로 5장에서는 결론 및 향후 연구과제에 대해 논한다.

Ⅱ. 관련 연구

2-1 고령자의 뇌졸중 질환

뇌줄중은 뇌혈관이 막히거나 터져서 갑작스럽게 운동장애와 감각장애, 발음장애, 의식장애나 사지 마비와 같은 뇌 기능에 이상이 발생하는 질환이다[2], [3], [4], [5]. 이러한 뇌졸중은 혈관이 막혀서 발생하는 뇌경색(허혈성)과 혈관이 터져서 발생하는 뇌출혈(출혈성)로 나눌 수 있다. 뇌경색은 동맥경화증으로 손상된 뇌혈관에 혈전이 생겨 혈관이 좁아져 막히는 뇌혈전

증과, 심장이나 경동맥과 같은 큰 동맥에서 생긴 혈전이 혈액을 타고 다니다 뇌혈관을 막아 생기는 뇌색전증 등으로 구분된다. 뇌출혈은 뇌실질내 뇌출혈과 지주막하 출혈이 대표되는데, 뇌실질내 뇌출혈은 외부의 충격 없이 자발적으로 뇌에 출혈이 일어나는 것으로 고혈압이 주요 원인으로 보고되었다. 지주막하 출혈은 탄력이 약해진 혈관 벽 일부가 파리모양으로 부푼 뇌동맥류가 파열되어 뇌를 싸고 있는 지주막 아래로 피가 새어나와 발생되는 질환이다. 지주막하출혈은 환자의 1/3이 병원에 도착하기 전에 사망에 이를 만큼 치명적인 질환으로 보고되고 있다. 한국의 통계청[1] 및 건강보험심사평가원[17]에 따르면, 뇌졸중은 단일 질환으로는 심장 질환 및 폐렴 다음으로 높은 사망률을 보이고 있다. 특히 대한민국은 인구의 고령화와 만성질환자 증가로 뇌졸중으로 인한 환자 수가 매년 증가하고 있다. 이러한 뇌졸중은 한번 발생하면 사망 혹은 반신마비와 같은 심한 후유증을 남길 수 있기 때문에, 초기 발생 시 빠른 발견과 처치가 무엇보다 중요하다.

2-2 선행연구 및 주제탐구

한국인을 대상으로 뇌줄중 발생 예측모형을 구축한 연구로는, Jee 등의 연구가 있다[18]. Jee 등은 한국 국민건강보험공단의 건강검진 자료를 토대로 뇌졸중 위험요인들 중에서 연령, 당뇨병, 수축기혈압, 흡연, 총콜레스트롤, 운동, 체질량지수, 음주량 등을 이용해 10년 이내에 뇌졸중이 발생할 평균 위험에 대한 예측 모형을 개발하였다. 하지만, 이 연구는 framingham heart study [11], [12]의 뇌졸중 위험도 예측 모형 구축방법과 동일한 방식을 사용하고 있다. 또한, 뇌졸중의 주요 위험 요인들 중 일부를 고려하지 않았으며 뇌졸중 이외의 원인으로 인한 사망 가능성, 즉 경쟁위험(competing risk)을 고려하지 않았다는 한계점을 가지고 있다. 이러한 뇌졸중은 발병 후 3시간 이내에 어느 종류의 뇌졸중(뇌경색, 뇌출혈)인지, 뇌의 어느 부분이 얼마나 손상된 것인지 정확히 파악해야 한다[11], [19], [20]. 또한, 늦어도 3시간 이내에 전문적인 치료를 받을 수 있는 병원에 후송하는 것이 중요하다.

또 다른 방법에서는, 선행연구와 임상시험을 통해 뇌졸중의 주요 위험 요인들을 발견하였으며, 이러한 위험 요인으로는 흡연, 고혈압, 당뇨병, 비만 등이 있음을 보고하였다[11], [12]. 즉, 뇌졸중은 어느 한 가지 요인에 의해서 발생되기 보다는 다양한 위험 요인들이 상호 작용하여 발생될 가능성이 크다. 따라서 각 개인별 뇌졸중의 위험 요인을 평가하고 질환을 예측 또는 발병 초기에 발견할 수 있는 새로운 방법론이 대두되고 있다. 이러한 위험 요인을 기반으로 다양한 통계적 방법과 기계학습 방법 등을 이용한 뇌졸중 질환 예측에 대한 연구가 진행되어 왔다. Kannel 등[10]이 로지스틱 모형을 이용한 연구를 시작으로, Cox의 비례위험모형[11], [12], [13]이나, Weibull 모형[14]에 기반한 연구가 보고되어 왔다. 하지만, 이러한 위험 요인 기반의 선행 연구들은 한국인의 뇌졸중 질환 발생에 대한 위험도 예측에 적용하기 어려운 실정이다. 특히 한국인 고령자의 위험도 예측 모형을 새롭게 모색해야 하는 필요성이 대두되고 있다.

뇌졸중 질환은 언제 발병될지 명확하지 않으며, 특히 뇌졸중 과거력이 있는 사람이 없는 사람에 비해 재발될 확률이 9배 이상 높은 것으로 보고되었다. 또한 뇌졸중 재발률은 그 종류와 인종, 위험요인 등에 따라 다양하지만 일반적으로 1년 이내 재발률이 10~15%라는 임상 연구가 보고되었다. 따라서 뇌졸중 환자의 초기 발병을 빠르게 탐지하고 예측하는 것도 중요하지만, 과거에 뇌졸중 병력이 있는 사람의 재발 발견 또한 중요한 연구 이슈 중 하나이다.

2-3 데이터 설명

본 논문에서는 국민건강보험공단에서 제공한 노인코호트 DB(ver 1.0)를 기반으로 실험하였으며, 노인성 질환의 위험요인과 예후 분석 등 노인을 대상으로 하는 연구지원을 위해 구축된 연구용 데이터베이스이다[15]. 노인코호트 DB의 모집단은 2002년부터 2013년까지 건강보험 및 의료급여 자격을 유지하고 있는 만 60세 이상의 노인 대상자 550만 명을 대상으로 하였으며, 모집단의 10%를 단순 무작위 추출하였다. 최종적으로 추출된 표본 대상사의 수는 558,147명이며, 데이터는 코호트 형식으로 구축된 자료이다. 구축된 데이터의 상세 내용으로는 사회·경제적 정보 및 장애, 사망정보 등을 포함하는 자격정보, 진료 및 건강검진 등을 포함하는 의료이용정보, 요양기관 현황정보, 노인 장기요양 서비스 신청 및 이용정보 등을 포함하고 있다. 연구용 자료로 공개된 노인코호트 DB는 총 27억여개의 튜플과 210GB 크기의 용량을 가지며, 세부적으로는 18개의 항목 테이블로 구성되어 있다[15].

일반적인 기계학습 및 딥러닝 기반의 분류 또는 예측을 수행하기 위해서는, 원시 데이터(raw data) 자체가 완전하지 않고 일관성이 없는 중복된 데이터를 다수 포함할 가능성이 크다[21], [22], [23]. 따라서 본 논문에서는 노인 질환 중 뇌졸중을 조기발견 및 예측에 필요한 데이터 처리 작업 및 예측 모델의 성능 개선을 위하여 노인코호트 DB 전처리 작업을 수행하였다. 먼저, 각 테이블에서의 중복 및 노이즈 튜플을 제거하였으며, 다음으로 분석을 위해 지정된 데이터 형식으로 변환(normalization) 및 일반화(generalization) 하였다. 그리고 본 논문의 실험 목적인 뇌졸중 질환 분석을 위해 주상병 코드가 I60~I69인 데이터만 추출하기 위해 테이블 간의 조인키를 활용하여 하나의 데이터 마트(data mart)를 구축하였다. 또한 데이터의 속성을 파악하여 병원, 약국 등의 기관에 대한 정보를 세부 DB화 하였으며, 건강항목과 처방전 등과 같은 건강관련 데이터를 분리하였다. 아래의 표 1은 뇌졸중 질환과 관련된 데이터 추출 및 전처리를 반영한 노인코호트 DB의 주요 테이블명과 튜플 수를 나타내고 있다.

표 1. 노인코호트 DB의 주요 테이블 및 튜플 수
Table 1. Number of main tables and tuples in the elderly cohort DB

Table Name	Number of Tuples
JK_DB Table	2,105,304
GY_DB Table	67,219,744
GJ_DB Table	502,830
T01_ DB Table	401,251
YK_DB Table	196,884

※ JK_DB: Qualification DB, GY_DB: Diagnosis DB, GJ_DB: Medical Checkup DB, T01_DB: Long-Term Care DB, YK_DB: Medical Institutions DB

다음의 표 2는 표 1로부터 실험 및 검증을 위해 진료명세 테이블의 주상병 코드가 I60~I69로 뇌졸중 질환 고령자 및 일반 고령자의 검사항목과 건강관련 데이터를 추출하기 위해 구축한 데이터 마트와 튜플 수를 나타내고 있다.

표 2. 데이터 마트 구축 및 튜플 수
Table 2. Number of data mart deployments and tuples

Table Name	Number of Tuples
DS_DB ⋈ GJ DB Table	14,334,844
JK_DB ⋈ GJ DB Table	502,830
JK_DB ⋈ DS_DB Table	67,219,744
DS_DB ⋈ PRE_DB Table	170,024,778
DS_DB ⋈ GY_DB Table	199,092,962
DS_DB ⋈ CH_DB Table	209,245,603

※ ⋈: Join Symbol Between DB, DS: Diagnosis Specification, PRE: Prescription, CH: Corporal History

Ⅲ. CNN 기반의 고령자 뇌졸중 질환 예측 시스템

CNN(convolution neural network)은 1989년 LeCun이 소개하였으며, 필기체 인식을 위한 프로젝트를 통해 개발되었다[24]. 이후 2003년 Behnke[25]와 Simard[26]에 의해 일반화되고 단순화 되면서 이론적 개념정립과 다양한 분야에서 활발하게 연구 및 적용되는 기초가 되었다. 이러한 CNN은 특징 추출 기능과 다중 분류(multi classification) 등의 예측을 실행할 수 있는 기능이 있다. 먼저 특징 추출 단계에서는 컨볼루션 레이어(convolution layer)와 서브샘플링 레이어(sub-sampling layer)를 통과한다. 동일한 계수를 갖는 필터를 반복적으로 적용하여 변수의 수를 획기적으로 줄여줌으로써, 토폴로지 변화에 무관한 항상성(invariance)을 보장 받는다. 일반적으로 1개의 특징 맵(feature map)에 대해 1개의 서브샘플링 연산을 수행한다. 이러한 과정을 거치면서 지역적 특징(local feature)으로부터 전역적 특징(global feature)을 얻어내며, 여러 단계의 컨볼루션 레이어와 서브샘플링 레이어를 거치면서 특징 맵의 크기는 작아지고 전체 데이터를 표현할 수 있는 대표적 특징들만 남게 된다. 최종 분류 단계에서는 은닉 레이어(hidden layer) 간의 노드들이 모두 연결된 완전 연결 구조(fully connected)의 입력으로 연결되어 작동된다. 특히, CNN 모델의 중요 파라미터인 가중치(weight)는 back-propagation 알고리즘을 이용하여 예측 정확도를 높이도록 조정할 수 있다. 또한, 학습률(learning rate)과 훈련 횟수 등의 하이퍼 파리미터 튜닝(hyperparameter tuning)을 조정하며 최적의 뇌졸중 질환 예측모델을 찾도록 학습시켰다. 그리고 batch normalization 레이어를 추가함으로써, 파라미터를 결정할 때 발생될 수 있는 기울기 사라짐(gradient vanishing)문제를 방지하도록 하였다[27]. 더욱이 CNN 모델에서 빠른 학습을 위해 학습률 값을 크게 설정하는 경우에도 batch normalization 레이어에서 기울임 사라짐뿐만 아니라 지역 최적해(local minima)에 빠지는 문제를 최소화할 수 있다[28].

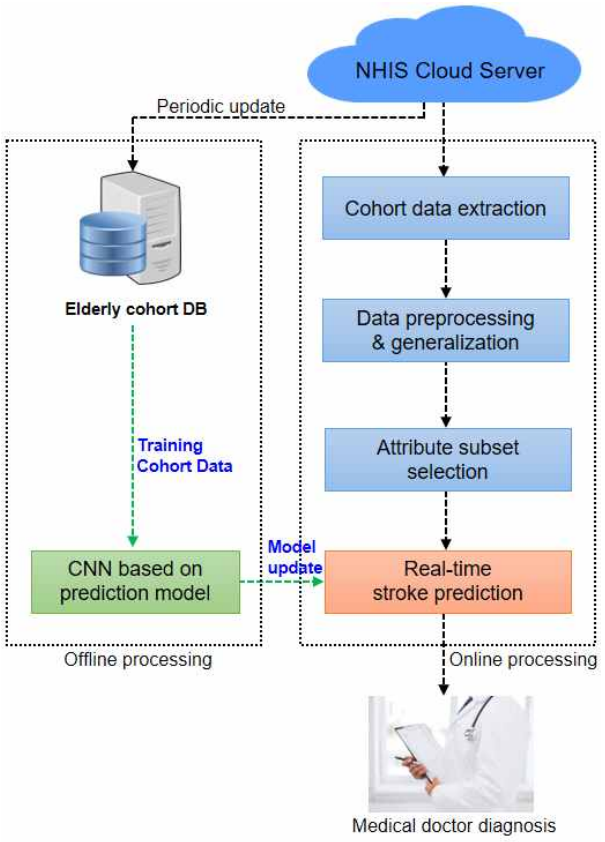


그림 1. 제안하는 시스템의 전체 구조도
Fig. 1. The overall architecture of our proposed system

본 논문에서 제안하는 CNN 기반의 뇌졸중 질환 예측 시스템은 총 5개의 모듈로 구성된다. 세부적으로는 1개의 오프라인

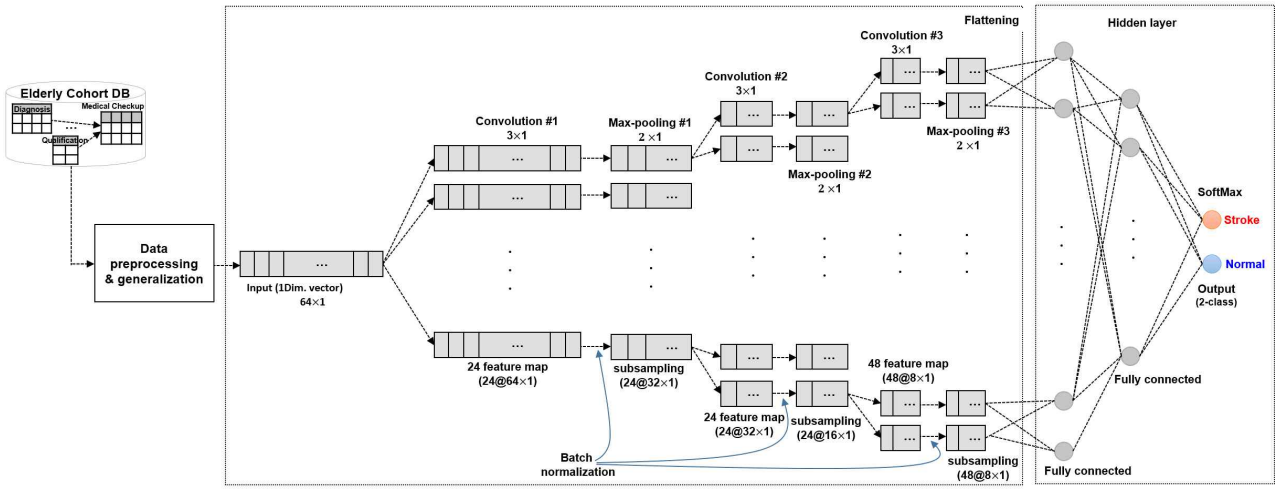


그림 2. 노인코호트 DB를 이용한 Stroke-CNN의 구조
Fig. 2. The architecture of Stroke-CNN(convolution neural network) using elderly cohort DB

처리(offline processing)인 CNN based on prediction model 모듈과, 4개의 온라인 처리(online processing) 모듈인 Cohort data extraction, data preprocessing & generalization, attribute subset selection, real-time stroke prediction 모듈로 구성된다(그림 1 참조). 1) 오프라인 처리 모듈에서는 국민건강보험공단(NHIS)의 클라우드 서버에서 제공하는 노인코호트 데이터를 시스템에서 설정한 주기대로 업데이트 한다. 다음으로 수집한 노인코호트 데이터를 기반으로 CNN based on prediction model에서 최적의 학습 및 학습모델을 저장하여 온라인 처리의 real-time stroke prediction 모듈로 전달한다. 2) Cohort data extraction에서는 시스템 요청 시, 국민건강보험공단의 클라우드 서버로부터 뇌졸중 질환 예측에 대한 원시 데이터를 실시간 수집한다. 3) Data preprocessing & generalization 모듈은 데이터 분석과 CNN 기반 뇌졸중 예측 모델 탑재를 위해 지정된 데이터 형식으로 변환 및 일반화를 실행시킨다. 4) Attribute subset selection 모듈은 CNN 기반 뇌졸중 예측 모델의 최적 성능을 보장하기 위해, 코호트 속성 집합을 선택함으로써 시스템의 예측 정확도와 분석 속도를 향상시킨다. 5) Real-time stroke prediction 모듈에서는 기 학습해놓은 CNN 기반의 뇌졸중 예측 모델을 탑재하고, 선택된 코호트 속성 집합을 이용하여 실시간으로 질환의 위험도 정도를 분석한다. 분석한 정보는 의료진 또는 병원에 전달되어 뇌졸중의 위험 상황 정도를 판단한다.

Real-time stroke prediction에서는 CNN based on prediction model을 생성하는 모듈에서 그 성능이 검증된 1차원(1-Dimension, 1-D) 형태의 CNN 구조[29], [30]로, 국민건강보험공단의 클라우드 서버로부터 코호트 데이터를 수집하고 전처리 및 최적 속성 집합을 이용하여 뇌졸중 질환 예측 및 위험 정도를 판단한다(그림 2 참조). CNN based on prediction model 모듈에서 학습하고 real-time stroke prediction 모듈에 탑재되는 노인코호트 DB를 이용한 1-D CNN 구조의 설명한다. 본 논문에서 새롭게 제안하는 1-D CNN 기반의 뇌졸중 질환 예측 모델을 Stroke-CNN으로 명명하며, 3개의 컨볼루션과 서브샘플링 레이어 및 2개의 완전 연결 구조의 은닉 레이어로 구성된다. 각각의 컨볼루션 레이어 이후에는 ReLU(rectified linear unit) 활성화 함수(activation function)를 사용하였다. ReLU 활성화 함수는 0보다 작은 값이 나올 경우 0으로 반환하고, 0보다 큰 값은 그대로 반환하는 함수이다. 다음으로 서브샘플링 레이어 이전에 batch normalization을 추가함으로써 파라미터들을 결정할 때 발생될 수 있는 기울기 사라짐 문제는 방지하도록 구조를 설계하였다. 마지막으로 은닉 레이어에서는 노드들이 모두 연결된 2개의 완전 연결 구조 레이어를 두도록 설계하였다. 일반적인 CNN 모델에서는 완전 연결 구조를 여러 개 레이어를 쌓아 모델링하지만, 본 논문에서 제안하는 Stroke-CNN에서는 2개의 레이어만 사용하여 뇌졸중 질환 예측 모델을 구성하였다. 은닉 레이어의 최종 레이어는 뇌졸중 노인과 정상 노인을 분류하는 계층으로, 본 논문에서는 각 클래스의 확률 값을 평가하기 위하여 softmax 레이어를 최종 계층에 두었다. 최종 할당되는 클래스에서는 평가한 확률 값이 큰 값을 갖는 쪽으로 결정된다.

Ⅳ. 실험 결과 및 분석

본 장에서는 노인 코호트 DB를 이용하여 뇌졸중 질환을 예측하는 Stroke-CNN을 이용한 실험결과를 살펴본다. 본 실험에서 새롭게 제안하는 Stroke-CNN은 딥러닝의 파라미터인 학습률(learning rate)과 훈련 횟수, 하이퍼 파리미터 값의 설정에 따른 실험 결과를 살펴본다. 또한, Stroke-CNN 구조에서의 컨볼루션과 서브샘플링 레이어의 개수, batch normalization 적용여부 등을 고려하여 실험을 수행하였다. 특히, 초기 학습률이 0.01로 시작할 경우, 전체 훈련 횟수의 50%에서 0.01에서 0.001로 1/10로 줄인 다음, 다시 75%의 훈련 횟수에서 0.0001로 다시 1/10으로 줄이고, 최종 훈련 횟수에서 학습을 종료하였다[31].

4-1 실험 데이터 구성

본 절에서는 제안하는 Stroke-CNN 기반의 뇌졸중 질환 예측 및 시스템 검증을 위하여, 노인 코호트 DB의 고령자별로 총 64개의 속성을 추출하였다. data preprocessing & generalization 모듈과 attribute subset selection 모듈을 통해서 결측치 및 널(null) 값들을 삭제 및 보정 처리하였으며, 추출한 속성들은 아래의 식 (1)에서와 같이 Z-score 정규화(normalization)를 수행 하였다. 정규화 적용의 예로, 혈청크레아티닌수치 값이 최소 값과 최대 값의 범주와 크기가 다양하기 때문에 측정단위에 종속되는 문제가 있으므로 이를 방지할 필요가 있다. 이러한 정규화 과정은 해당 데이터가 0.0에서 1.0과 같은 작은 범위 내에 위치하도록 변환함으로써, 모든 속성별로 동일한 가중치가 적용된다.

$$\vec{x}_i = \frac{x_i - \mu}{\sigma} \times \alpha \tag{1}$$

식 (1)에서 σ 와 μ 는 속성 x 의 표준편차와 평균을 의미하고, α 는 가중치 값으로, 본 논문의 실험에서는 1.0으로 설정하였다.

본 실험에서 추출한 속성에는 체질량 지수, 요단백, 총콜레스테롤수치, 혈청크레아티닌, 감마지피티수치 등과 연속형 속성 값과 일일 음주량, 흡연여부, B형간염 항원보유, 고강도 신체활동 등의 이산형 속성 값을 포함하도록 구성하였다. 속성 부분집합의 선택은 중복된 속성이나 무관한 속성을 제거하여 데이터 집합을 축소시키며, 목표에 적합한 최적 속성집합을 찾아 데이터 분류를 수행한 결과 확률분포(probability distribution)가 모든 속성을 이용하여 얻어낸 확률분포와 최대한 같도록 하는 것이다. 본 실험에서는 추출한 속성들을 모두 사용하는 것이 아니라, 속성 부분집합의 선택 방법 중 그 성능이 검증된 Hall[32]의 방법을 사용하였다. Hall의 방법은 속성 값과 최적우선탐색(best first search) 값인 Y에 대한 엔트로피(entropy), 목표 클래스와 속성들 간의 피어슨 상관계수(Pearson's correlation coefficient)를 이용한 조건부 확률을 계산한다. 먼저 각 속성들에 대한 정보 이익을 얻기 위해 임의의 속성 Y에 대한 엔트로피를 식 (2)와 같이 계산한다[21], [32], [33].

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 (p(y)) \tag{2}$$

최종적으로 각각의 부분집합 $F_S \subset F$ 가 전체 속성들을 얼마나 효율적으로 표현하는지를 평가하기 위하여 메리트 함수(merit function)(식 (3))을 사용한다. 메리트 함수의 값이 가장 큰 부분집합이 전체 속성들을 최적으로 표현할 수 있는 부분집합임을 의미한다[32], [33].

$$Merit(F_S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{3}$$

여기서, k는 부분집합 F_S에서의 속성들의 개수를 의미하며, $\overline{r_{cf}}$는 F_S에 포함된 속성의 평균 분포, $\overline{r_{ff}}$는 속성의 평균 상관관계 값을 의미한다.

4-2 Stroke-CNN 기반 실험 결과 및 분석

본 실험에서는 뇌졸중 고령자 38,669명과 랜덤하게 추출한 일반 고령자 38,669명의 코호트 데이터를 사용하였으며, 총 77,338명의 데이터 셋으로 실험하였다. 그리고 뇌졸중 질환 예측의 학습의 랜덤성을 보완하고자 10-fold 교차검증(cross validation)을 적용하였다. 이러한 교차검증을 통해서 보다 정확한 뇌졸중 질환 예측 모델의 품질 척도를 제공하고 일반화하고자 하였다. 이후의 모든 실험에서는 10-fold 교차검증으로 Stroke-CNN 모델에 동일하게 적용하였다. Optimizer은 Adam으로 고정하였고, 학습률과 훈련 횟수 등의 하이퍼 파리미터 튜닝은 아래의 실험과정에서 보다 상세하게 설명하였다.

첫 번째 실험에서는 Stroke-CNN의 구조를 다르게 정의하여 실험한 결과와 분석을 수행하였다. 본 실험에서는 컨볼루션 및 은닉 레이어의 개수 달리 설정하고 서브샘플링의 방법과 batch normalization 사용 여부를 변경하며 실험하였다. 초기의 학습률은 0.01을 시작으로 15,000번째 훈련 횟수(iteration)에서는 1/10으로 줄인 0.001, 22,500번째 훈련 횟수에서는 0.0001 값으로 설정하고 30,000번째 훈련에서 학습을 종료시켰다. 다음의 표 3은 컨볼루션 레이어를 2~4개, 은닉 레이어의 완결연결 구조는 1~3개 및 서브 샘플링과 batch normalization 사용여부에 따른 실험결과를 보이고 있다. 실험결과, 컨볼루션과 완전연결 구조는 각각 3개와 2개를 사용하고, 서브샘플링은 max pooling과 batch normalization을 사용했을 때 고령자의 뇌졸중 질환 예측 정확도가 높은 것으로 나타났다.

표 3. Stroke-CNN 구조별 뇌졸중 예측 정확도 분석
Table 3. Stroke prediction accuracy analysis by Stroke-CNN structure

Conv. #	BN	S-S	FC #	Accuracy (%)
2	○	average	1	74.14
2	×	max	3	73.57
2	○	max	2	78.69
3	○	average	2	84.81
3	×	max	3	82.98
3	○	max	2	**85.64**
4	○	average	1	82.33
4	×	max	3	83.63
4	○	max	2	84.88

※ Conv. #: number of convolution layer, BN: batch normalization, S-S: sub-sampling, FC #: number of fully connected layer

다음 실험에서는 컨볼루션과 완전연결 구조는 각각 3개와 2개 고정하고 서브 샘플링과 batch normalization 사용하였으며, 학습률과 훈련 횟수와 같은 하이퍼 파라미터를 튜닝하며 최적의 뇌졸중 질환 예측 모델을 탐색 및 분석하였다. 본 실험을 통해 학습률은 0.001, 훈련 횟수는 40,000번 이상일 때 전반적으로 안정적인 예측 성능을 보였다. 다음은 표 4는 본 논문에서 제안하는 노인 코호트 DB 기반의 Stroke-CNN 모델에서 89.47%의 만족스러운 성능을 보여주고 있음을 확인하였다.

표 4. Stroke-CNN에서의 파라미터 튜닝을 통한 최적의 뇌졸중 예측 모델 분석

Table 4. Analysis of optimal stroke prediction model with parameter tuning in Stroke-CNN

Learning rate	Max iteration	Test iteration	Weight decay	Momentum	Accuracy (%)
0.01	30,000	50	0.01	0.5	84.80
0.001	30,000	100	0.001	0.9	85.79
0.0001	30,000	100	0.0001	1.0	85.71
0.01	40,000	50	0.001	0.9	87.69
0.001	40,000	50	0.001	0.9	**89.47**
0.0001	40,000	100	0.0001	0.9	88.96
0.01	50,000	50	0.01	0.9	86.10
0.001	50,000	100	0.001	0.9	89.35
0.0001	50,000	50	0.001	0.5	88.87

Ⅴ. 결 론

본 논문에서는 기존의 뇌졸중 초기 장애를 평가하거나 향후 10년 후의 발병 가능성을 예측한다는 한계와 단점을 보완하는 차원에서 Stroke-CNN을 기반으로 한 새로운 뇌졸중 질환 예측 시스템을 제안하였다. 제안된 시스템은 실시간으로 질환 예측을 위하여 국민건강보험공단에서 공개한 노인 코호트 DB에서 뇌졸중과 관련된 코호트 속성을 전처리 및 최적 부분집합을 선택하여 빠르고 정확한 뇌졸중 예측을 실시하였다. 평가항목인 예측 정확도에서 만족스러운 수치 등을 실험 결과와 분석을 통해 확인함으로써 제안된 시스템의 성능을 검증하였다. 추가적으로, 제안된 딥러닝 CNN 기반의 뇌졸중 예측 모델은 다른 질환에 대한 조기발견 및 예측으로도 확장이 가능하다는 추가적인 장점을 갖는다.

향후 연구로는 코호트 DB뿐만 아니라 실시간 생체신호 정보와 결합한 종합적인 질환 예측시스템을 연구 및 개발하고자 한다.

감사의 글

이 논문은 2015년 정부(미래창조과학부)의 재원으로 국가과학기술연구회 융합연구단 사업(No. CRC-15-05-ETRI)의 지원을 받아 수행된 연구이며, 고령자 코호트 DB는 국민건강보험공단의 지원을 받았음.

참고문헌

[1] Statistics Korea, The cause of death statistics in Koreans. Available: http://kostat.go.kr/portal/korea/kor_nw/1/6/2/index.board.

[2] J. M. Racosta, F. D. Guglielmo, F. R. Klein, P. M. Riccio, F. M Giacomelli, M. E. G. Toledo, F. P. Cassara, A. Tamargo, M. Delfitto, and L. A. Sposato, "Stroke severity score based on six signs and symptoms the 6 score - a simple tool for assessing stroke severity and in-hospital mortality," *Journal of Stroke*, Vol. 16, No. 4, pp. 178-183, September 2014.

[3] B. C. Meyer and P. D. Lyden, "The modified national institutes of health stroke scale: its time has come," *Journal of Stroke*, Vol. 4, No. 4, pp. 267-273, 2009.

[4] Q. Zhang, Y. Xie, P. Ye, and C. Pang, "Acute ischaemic stroke prediction from physiological time series patterns," *Journal of Australasian Medical*, Vol. 6, No. 5, pp. 280-286, 2013.

[5] H. J. Lee, J. S. Lee, J. C. Choi, Y. Cho, B. J. Kim, H. Bae, D. Kim, W. Ryu, J. Cha, D. H. Kim, H. Nah, K. Choi, J. Kim, M. Park, J. Hong, S. I. Sohn, K. Kang, J. Park, W. Kim, J. Lee, D. Shin, M. Yeo, K. B. Lee, J. G. Kim, S. J. Lee, B. Lee, M. S. Oh, K. Yu, T. H. Park, J. Lee, and K. Hong, "Simple estimates of symptomatic intracranial hemorrhage risk and outcome after intravenous thrombolysis using age and stroke severity," J*ournal of Stroke*, Vol. 19, No. 2, pp. 229-231, May 2017.

[6] K. M. Lee, Y. H. Jang, Y. H. Kim, S. K. Moon, J. H. Park, S. W. Park, H. J. Yu, S. G. Lee, M. H. Chun, and T. R. Han, "Reliability and validity of Korean version of national institutes of health stroke scale," *Annals of Rehabilitation Medicine*, Vol. 28, No. 5, pp. 422-435, 2004.

[7] C. D. Bushnell, D.C. Johnston, and L. B. Goldstein, "Retrospective assessment of initial stroke severity: comparison of the NIH stroke scale and the Canadian neurological scale," *Journal of Stroke*, Vol. 32, pp. 656-660, 2001.

[8] S. M. Lai, P. W. Duncan, and J. Keighley, "Prediction of functional outcome after stroke: comparison of the orpington prognostic scale and the NIH stroke scale," *American Heart Association*, Vol. 29. No. 9, pp. 1838-1842, 1998.

[9] D. W. Powers, "Assessment of the stroke patient using the NIH stroke scale," *Emergency Medical Services*, Vol. 30, No. 6, pp. 52-56, 2001.

[10] W. B. Kannel, D. L. McGee, and W. P. Castelli, "Latest perspectives on cigarette smoking and cardiovascular disease," *Journal of Cardiac Rehabilitation*, Vol. 4, No. 7, pp. 267-277, 1984.

[11] P. A. Wolf, R. B. D'Agostino, A. J. Belanger, and W. B. Kannel, "Probability of stroke: a risk profile from the framingham Study," *American Heart Association*, Vol. 22, pp. 312-318. 1991.

[12] R. B. D'Agostino, P. A. Wolf, A. J. Belanger, and W. B. Kannel, "Stroke risk profile: adjustment for antihypertensive medication: the framingham study," *American Heart Association*, Vol. 25, pp. 40-43, 1994.

[13] T. J. Wang, J. M. Massaro, D. Levy, R. S. Vasan, P. A. Wolf, R. B. D'Agostino, M.G. Larson, W. B. Kannel, and E. J. Benjamin, "A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the framingham heart study," *The JAMA Network*, Vol. 290, No. 8, pp. 1049-1056, 2003.

[14] K. J. Carroll, "On the use and utility of the Weibull model in the analysis of survival data," *Journal of Controlled Clinical Trials,* Vol. 24, No. 6, pp. 682-701, 2003.

[15] National Health Insurance Service (NHIS). Available: https://www.nhis.or.kr/static/html/wbd/g/a/wbdga0101.html.

[16] Korea Informative Classification of Diseases (KOICD), Available: http://www.koicd.kr/.

[17] Health Insurance Review & Assessment Service (HIRA), Available: http://www.hira.or.kr/eng/main.do.

[18] J. S. Lee, J. M. Park, T. H. Park, K. B. Lee, S. J. Lee, Y. J. Cho, M. K. Han, H. J. Bae, and J. Y. Lee, "Development of a stroke prediction model for Korean," *Korean Neurological Association*, Vol. 28, No. 1, pp. 13-21, 2010.

[19] H. Ay, K. L. Furie, A. Singhal, and W. S. Smith, "An evidence-based causative classification system for acute ischemic stroke," *American Neurological Association*, Vol. 58, pp. 688-697, October 2005.

[20] A. Trialists' Collaboration, "Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients," *British Medical Journal (BMJ)*, Vol. 324, pp. 71-86, 2002.

[21] J. Yu, H. Kang D. Park, H. Bang, and D. Kang, "An In-depth analysis on traffic flooding attacks detection and system using data mining techniques," *Journal of Systems Architecture*, Vol. 59, No. 10, pp. 1005-1012, November 2013.

[22] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3th ed. Morgan Kaufmann Pub., ch. 3, 2012.

[23] J. Yu, D. Kim, H. Park, S. Chon, K. Cho, S. Kim, S. Yu, S. Parn, and S. Hong, "Semantic Analysis of NIH stroke scale using machine learning techniques," in *Proceeding of 2019 International Conference on Platform Technology and Service (PlatCon)*, Jeju, Korea, pp. 82-86, 28-30 Jan. 2019.

[24] Y. LeCun, B. Boser, J. S. Denker, and D. Henderson, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, Vol. 1, No. 4, pp. 541-551, 1989.

[25] S. Behnke, Hierarchical neural networks for image interpretation, LNCS 2766, Springer, 2003.

[26] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *In Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Vol. 2, pp. 958-962, August 2003.

[27] G. Ahn, J. Yoo, S. Lee, and S. Kim, "Explainable convolutional neural networks for multi-sensor data," *Journal of the Korean Institute of Industrial engineers*, Vol. 45, No. 2, pp. 146-153, April 2019.

[28] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

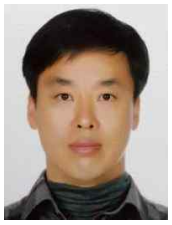
[29] B. L. P. Cheung, "Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks," *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 222-225, 4-7 March 2018.

[30] C. Luciano, P. Veltri, V. Eugenio, and Z. Ester, "Deep learning techniques for electronic health record analysis," *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 23-25 July 2018.

[31] J. Bae, J. Yim, J. Yu, K. Kim, and J. Kim, "Performance analysis of hint-kd training approach for the teacher-student framework using deep residual networks," *Journal of The Institute of Electronics and Information Engineers*, Vol. 54, No. 5, pp. 35-41, May 2017.

[32] Hall, M., Correlation-based feature selection for machine learning, PhD dissertation, Department of Computer Science, Waikato University, hamilton, NZ, 1998.

[33] J. Yu, B. Lee, and D. Park, "Real-time cooling load forecasting using a hierarchical multi-class SVDD," *Multimedia Tools and Applications*, Vol. 71, pp. 293-307, 2014.



유재학(Jaehak Yu)

2001년 : 건국대학교 (이학사-전산과학)
2003년 : 고려대학교 (이학석사-전산학)
2010년 : 고려대학교 (이학박사-전산학)

2006년~2008년 : 고려대학교 컴퓨터정보학과 초빙전임강사
2010년~현재 : 한국전자통신연구원 선임연구원
※관심분야 : 데이터마이닝, 기계학습, 딥러닝, 헬스케어분석, 침입탐지, 지능형 데이터베이스

권순현(Soon-Hyun Kwon)

1998년 : 인천대학교 (이학사-전산과학)
2003년 : 숭실대학교 (이학석사-컴퓨터학)
2008년 : 숭실대학교 (이학박사 수료-컴퓨터학)

2000년~2005년 : ㈜현대정보기술 현대상선 IT실 근무
2013년~현재 : 한국전자통신연구원 선임연구원
※관심분야 : 온톨로지, 온톨로지 추론, 시맨틱웹, 지식임베딩, 시맨틱 IoT/IoE 플랫폼, 지능형 데이터베이스

호치멍 벤자민(Chee Meng Benjamin Ho)

2012년 : 퀸즐랜드대학교 (공학사-생명공학)
2018년 : 난양공업대학교 (공학박사 수료-메치컬공학)

2019.09~현재 : 한국표준과학연구원 의료계측량센터, 연구원
2019.09~현재 : 한국전자통신연구원 연구원
※관심분야 : 헬스 관련 빅데이터 수집 및 분석, 지능형 건강 모니터링 시스템, 웨어러블 센서

이경란(Kyeong-ran Lee)

1995년 : 원광대학교 (전문학사-국어국문학)
1999년 : 방송통신대학교 (학사-국문학)
2006년 : 서강대학교 (석사-언론대학원)

2014년~2018년 : 국민건강보험공단 빅데이터실 차장·부장
2019년~2019년 : 국민건강보험공단 기획조정실 미래전략부장
2020년~현재 : 국민건강보험공단 급여전략실 급여분석부장
※관심분야 : 건강보험, 빅데이터, 헬스케어, 출판

김내수(Nae-soo Kim)

1985년 2월 : 한남대학교 (이학사-수학)
1989년 2월 : 한남대학교 (이학석사-수학)
2001년 2월 : 한남대학교 (공학박사-컴퓨터공학)

1976년~1990년: 국방과학연구소 근무
1990년~현재: 한국전자통신연구원 (실장/책임연구원)
※관심분야 : 인공지능(AI), 사물인터넷(IoT), 위성통신

표철식(Cheol-Sig Pyo)

1991년 2월 : 연세대학교 (공학사-전자공학)
1999년 2월 : 한국과학기술원 (공학석사-전기 및 전자공학)

1991년 1월~현재 : 한국전자통신연구원 (책임연구원)
2014년 8월~2015년 8월 : 미국 조지아공대 방문연구원
2015년 12월~현재 : 한국전자통신연구원 KSB융합연구단장
※관심분야 : IoT, 기계학습, 딥러닝, 지식융합 지능서비스

박세진(Se Jin Park)

1983년 : 고려대학교 (공학사-산업공학)
1985년 : 고려대학교 (공학석사-산업공학)
1994년 : 고려대학교 (공학박사-산업공학)

1988 - 현재 : 한국표준과학연구원 책임연구원
2017 - 현재 : 과학기술연합대학원 대학교 의학물리학과 교수
※관심분야 : 헬스 관련 빅데이터 수집 및 분석, 지능형 건강 모니터링 시스템, 감성공학, HCI