Journal of KIIT. Vol. 19, No. 6, pp. 1-7, Jun. 30, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.6.1

지역 특성 데이터에 기반한 코로나19 사망 위험 요인 분석

김 건*, 배지훈**, 이종혁***

Analysis of Risk Factors for COVID-19 Death based on Regional Characteristics Data

Geon Kim*, Ji-Hoon Bae**, and JongHyuk Lee***

요 약

세계보건기구(WHO)가 코로나바이러스 감염증(COVID-19)에 대한 팬데믹을 선언함과 동시에 전 세계에서 유례없는 사망자가 발생하고 있다. 그러나 COVID-19 예방과 확진자 분류에 대해 중점을 두고 정작 사망률에 관한 연구는 아직 미미한 실정이다. 본 논문에서는 COVID-19의 사망률을 감소시키기 위해 사망 위험에 영향을 미치는 주요 요인을 데이터 파이프라인을 통해 도출하고 이를 선형회귀분석에 적용하여 사망자 수에 관한 모형을 생성한다. 이 모형에 대한 검정통계량에서 결정계수는 0.749로 모형이 적합하다고 볼 수 있다.

Abstract

At the same time as the World Health Organization (WHO) declared a pandemic for the coronavirus (COVID-19), an unprecedented number of deaths are occurring around the world. However, there are still few studies on the mortality rate, focusing on the prevention of COVID-19 and the classification of confirmed cases. In this paper, to reduce the death rate of COVID-19, the main factors affecting the risk of death are derived through a data pipeline and applied to linear regression to create a model for the number of deaths. In the test statistic for this model, the R-squared is 0.749, indicating that the model is suitable.

Keywords
COVID-19, risk of death, linear regression, data analysis

* 대구가톨릭대 인공지능·빅데이터공학과 학사과정
- ORCID: http://orcid.org/0000-0002-4806-2644
** 대구가톨릭대 인공지능·빅데이터공학과 조교수
- ORCID: http://orcid.org/0000-0002-0035-5261
*** 대구가톨릭대 인공지능·빅데이터공학과 조교수(교신저자)
- ORCID: http://orcid.org/0000-0002-8163-9388

· Received: Feb. 26, 2021, Revised: Jun. 23, 2021, Accepted: Jun. 26, 2021
· Corresponding Author: JongHyuk Lee
Dept. of Artificial Intelligence and Big Data Engineering, 13-13 Hayang-ro, Hayang-eup, Gyeongsan-si, Gyeongbuk, Korea,
Tel.: +82-53-850-2882, Email: jonghyuk@cu.ac.kr

Ⅰ. 서 론

세계보건기구에서 2020년 3월 11일 코로나 감염증-19(COVID-19, 이하 코로나19)를 전염병의 세계적 대유행인 글로벌 팬데믹으로 공식 선언하였다[1]. 2021년 5월 21일 현재 전 세계 220개국으로 확산되어 확진자 165,940,194명, 사망자 수는 3,446,534명을 기록하고 있어 사망률은 약 2.08%이다[2]. 코로나19 감염자가 사망에 이르기까지 위험 요인으로 고령, 면역 저하, 비만, 만성 신장 질환, 심장질환, 폐 질환 등의 개인에 국한된 사망 위험 요인이 발표[3]된 바 있으나 지역 특성에 따른 사망 위험 요인은 아직 구체적으로 명시된 바가 없다.

본 논문은 코로나19로 인한 사망 위험에 영향을 미치는 지역 특성 데이터를 분석한다. 이를 위해 시군구별 보건 분야 예산액, 인구밀도, 의료시설 현황 등에 관한 데이터를 수집하고 이를 선형회귀분석하여 시군구별 사망자 수를 예측하는 모형을 제시한다. 이를 통해 지역 특성에 관한 연구 사례가 미비한 점을 보완하여 코로나19로 인한 사망 위험을 낮추기 위한 정책 수립에 활용하고자 한다.

본 논문의 2장에서는 코로나19 발생 및 사망 관련 요인에 관한 관련 연구와 본 논문에서 사용한 선형회귀분석 방법을 설명한다. 3장에서는 본 논문에서 수행한 데이터 수집, 저장, 처리, 분석 과정을 설명한다. 4장에서는 분석 결과를 보인다. 마지막으로 5장에서는 본 논문의 결론과 향후 연구에 관하여 기술한다.

Ⅱ. 관련 연구

2.1 코로나19 발생 및 사망 관련 요인

통계청에서 발간한 2020년 고령자 통계에 의하면 경북지역은 65세 이상 고령인구 비중이 20.7%로 2020년 4월 30일 0시 기준으로 경상북도 코로나19 사망률(3.8명)은 전국 사망률(2.3명)보다 높은 수준이며, 연령별 코로나19 사망률은 80세 이상이 58.6%로 절반 이상을 차지하고 있다[4]. 이에 지역에 기반한 코로나19 발생률과 사망률 관련 요인에 대한 연구가 진행되고 있다.

“경상북도 지역의 코로나19 발생률 및 사망률 관련 요인”에 관한 연구[5]는 사망에 영향을 미치는 요인을 확인하기 위해 종속 변수로 확진 판정 후 사망 여부를 사용하였고 독립 변수로 연령, 성별, 기저질환 유무와 같은 개인 특성과 거주 형태와 같은 지역 특성을 사용하였다. 그리고 각 요인별 사망률의 차이를 보이기 위해 Pearson 카이제곱검정(x^2)을 시행하고, 사망에 영향을 미치는 요인을 확인하기 위해 로지스틱 회귀분석을 이용하여 다변량 분석을 수행하였다. 분석 결과 경북지역 확진자들의 사망 관련 요인은 연령, 기저질환 유무, 거주 형태에서 유의한 차이가 있었다.

“질병관리청의 데이터를 이용한 COVID-19 확진자의 사망에 대한 영향요인과 혈액변수의 타당도”에 관한 연구[6]는 코로나19 확진자에 대하여 임상소견뿐만 아니라 일반혈액검사를 시행하여 환자의 중증도 분류 기준을 변수로 포함하였다. 그리고 코로나19 확진자의 사망에 대한 위험 요인을 분석한 결과 치매, 열, 암, 호흡곤란, 만성폐쇄성 폐질환, 의식변화, 심장질환이 있거나 혈소판, 림프구 수치가 높거나, 이완기혈압이 비정상이거나, 나이가 많을수록 사망 위험이 커짐을 확인하였다. 또한 코로나19 확진자의 사망에 대한 혈액검사의 타당도를 확인한 결과, 림프구, 헤모글로빈, 헤마토크리트, 혈소판, 백혈구 순으로 사망에 대한 정확도가 높았다.

“COVID-19로 인한 사망자 수의 결정 요인: 대유행 초기 단계에서 저소득 국가와 고소득 국가 간의 차이”에 관한 연구[7]는 코로나19 사망자 수에 영향을 미치는 요인을 저소득 및 고소득 국가에 중점을 두었다. 다중 선형회귀분석을 통해 두 국가 그룹에서 코로나19 사망을 결정하는 사회-경제적 요인을 평가했다. 최종적으로 교통 인프라 시스템, 기대 수명 및 보건 관련 지출 비율과 같은 요소가 고소득 및 저소득 국가의 사망자 수 차이에 영향을 미치는 주요 요인으로 도출되었다.

이처럼 기존의 연구들은 개인 특성에 중점을 두거나 특정 지역과 국가에서의 사망 위험에 영향을 미치는 요인을 분석하였다. 이와 달리 본 논문은 지역 특성에 중점을 두고 분석 범위를 우리나라 전체

226개의 시군구로 확장하여 코로나19 사망 관련 요인의 적용 범위를 더 넓힌다.

2.2 선형회귀분석

선형회귀분석[8]은 변수 간의 상관관계를 분석하기 위해 사용되는 통계 방법이다. 선형회귀분석은 독립 변수 X와 종속 변수 Y, 상수항 b 사이의 관계를 모델링하는 방법으로 단순히 두 변수 사이의 관계를 분석할 경우 단순 선형회귀라고 하며, 여러 개의 독립 변수와 종속 변수 사이의 관계를 분석할 경우 다중 선형회귀라고 한다[9]. 본 논문은 코로나19 사망 예측모형을 위해 다중 선형회귀 모형을 기반으로 한다.

III. 데이터 분석

3.1 데이터 파이프라인

본 논문에서 만들고자 하는 데이터 파이프라인은 아래의 그림 1과 같다. 선형회귀분석에 사용되는 데이터를 수집하여 주기적으로 최신화시켜 준다. 수집한 데이터는 오브젝트 스토리지 또는 데이테베이스에 저장하며, 분산병렬처리 플랫폼을 통해 데이터를 처리 및 분석한다. 다음으로 시군구별로 예측된 코로나19 사망자 추이를 시각화한다. 본 논문에서 사용한 분석 도구는 파이썬(Python) 언어 및 관련 라이브러리이다.

3.2 데이터 수집

본 논문에서는 KOSIS 국가통계포털[10]의 지역적 특성과 관련된 다양한 데이터 세트를 활용한다. 지역적 특성 데이터는 시군구별 분류가 가능하며, 특정 시군구 측면의 세부화를 통해 사망 위험 요인에 대한 폭넓은 기초 자료로 활용될 수 있는 장점이 있다.

데이터 세트는 시군구별 보건소 운영 예산액, 인구밀도, 흡연율, 음주율 등 약 1만 개의 통계수치가 포함되며 모두 2020년의 데이터를 사용한다.

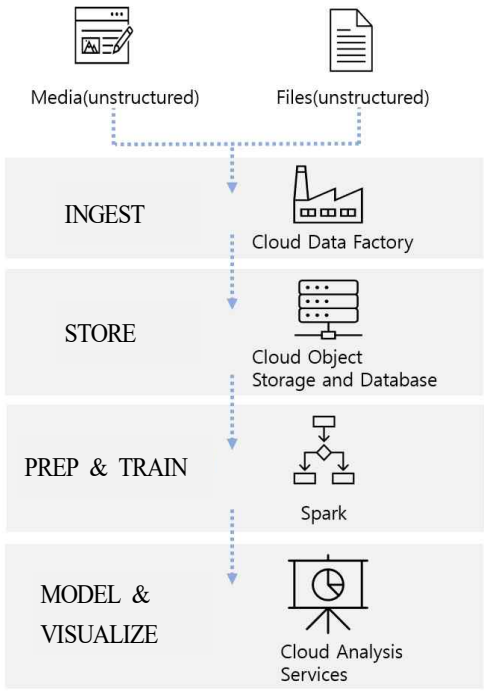

그림 1. 데이터 파이프라인
Fig 1. Data pipeline

사망자 수 데이터는 2021년 3월을 기준으로 하고, 사망자 수 데이터 중 수집할 수 없었던 시군구는 결측치로 두고 시도별 사망자 수 데이터에서 각 시군구 평균값으로 대체한다.

표 1. 실험 데이터 세트
Table 1. Experimental data sets

Var.	Description	Source
Y_1	No. of COVID-19 deaths	Data portal (2020)[11]
X_1	Health sector budget per person (KRW)	Local finance integrated open system (2020)[12]
X_2	Age	MOIS (2020)[13]
X_3	Population density(person/㎢)	e-Narajipyo (2020)[14]
X_4	Aging rate(%)	KOSIS (2020)[10]
X_5	Smoking rate(%)	KOSIS (2020)[10]
X_6	Drinking rate(%)	KOSIS (2020)[10]
X_7	Medical facilities per person	KOSIS (2020)[10]
X_8	Medical staff per person	KOSIS (2020)[10]

표 1의 데이터 중 보건 분야 예산액, 의료시설 현황, 의료 인력 데이터는 시군구별 인구수로 나누어 각각 시군구 별 1인당 보건 분야 예산액, 1인당 의료시설 현황, 1인당 의료 인력 현황 데이터로 변환시켜준다. 표 1의 데이터는 우리나라의 행정구역 시군구의 개수에 맞게 총 226개의 행을 가진다.

3.3 데이터 처리

본 논문의 분석에 사용할 데이터 중 극단값이 있으면 선형회귀분석의 결과가 왜곡될 수 있다. 따라서 상자 그림(Box plot)을 통해 각 변수에 극단값이 있는지 확인한다.

인구밀도 변수를 상자 그림으로 표현하면 그림 2와 같은데, 시군구별 인구밀도 값 사이에 많은 극단값들이 출력되는 것을 관찰할 수 있다. 따라서 log 변환을 수행하여 한쪽으로 치우친 데이터를 그림 3과 같이 정규분포가 되도록 변환한다.

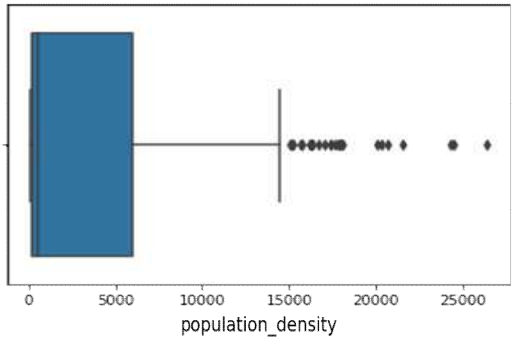

그림 2. 로그 변환 전 인구밀도 상자 그림
Fig. 2. Box plot of population density before log

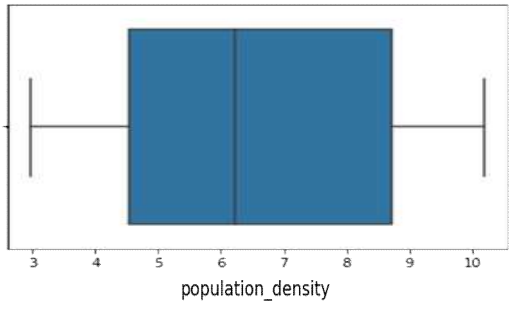

그림 3. 로그 변환 후 인구밀도 상자 그림
Fig. 3. Box plot of population density after log

흡연율 변수를 상자 그림으로 표현하면 그림 4와 같으며 양옆의 극단값을 제거한다. 그 외 보건분야 예산액, 나이, 고령화 비율, 음주율도 동일하게 극단값을 제거해주며, 의료시설 현황, 의료진 현황, 사망자 수는 log 변환을 수행하여 정규분포가 되도록 변환한다. 이때, 사용한 변수들의 왜도는 표 2와 같다.

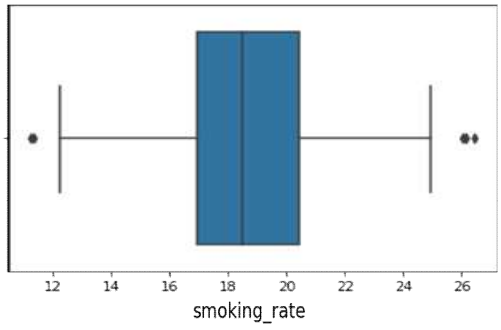

그림 4. 흡연율 상자 그림
Fig 4. Box plot of smoking rate

표 2. 데이터 왜도 분포
Table 2. Data skewness distribution

Var.	Skewness
Health sector budget per person (KRW)	1.79102
Age	0.96817
Population density(person/㎢)	-0.21782
Aging rate(%)	1.10219
Smoking rate(%)	0.29893
Drinking rate(%)	-0.60558
Medical facilities per person	-0.28141
Medical staff per person	-0.22477
No. of COVID-19 deaths	0.05508

3.4 데이터 분석

Pearson의 상관분석 결과, 보건분야 예산액, 의료시설 현황은 코로나 사망자 수 간에 음의 상관관계를 보이며 인구밀도, 음주율은 양의 상관관계를 보인다.

먼저, 코로나19 사망자 수 예측모형을 개발하기 전 모든 독립변수를 대상으로 회귀분석을 수행한다. 분석결과, 회귀계수로 R-squared는 0.324, Adj. R-squared는 0.278, F-statistic은 7.071로 도출되며 그 외 분석 결과는 표 3과 같다.

표 3. 코로나 사망자 1차 회귀분석 결과
Table 3. Results of the first regression analysis of COVID-19 deaths

Var.	Coef.	P > |t|
Intercept	-11.9730	0.031
Health sector budget per person (KRW)	5.946e-06	0.121
Age	0.4660	0.004
Population density(person/㎢)	0.1974	0.031
Aging rate(%)	-0.3460	0.001
Smoking rate(%)	-0.0716	0.080
Drinking rate(%)	-0.0494	0.093
Medical facilities per person	-0.6099	0.046
Medical staff per person	0.3684	0.195

다음으로 독립변수 간의 다중공선성이 존재하는지 진단하기 위해 분산팽창계수(VIF, Variance Inflation Factor)를 구한다. 표 4의 독립변수들의 VIF 결과를 살펴보면, 나이와 고령화 비율 변수가 서로 관련성이 있다는 것을 알 수 있다. 하지만 일반적으로 나이, 고령화 비율 모두 p-값이 유의미하기 때문에 VIF가 크더라도 특별히 대처할 필요가 없다.

표 4. 다중공선성 진단
Table 4. Multicollinearity diagnostics

Var.	VIF
Health sector budget per person (KRW)	6.792189
Age	52.159869
Population density(person/㎢)	5.256171
Aging rate(%)	64.265378
Smoking rate(%)	1.333425
Drinking rate(%)	3.127795
Medical facilities per person	2.161445
Medical staff per person	2.499414

Ⅳ. 분석 결과

본 장에서는 코로나 사망과 관련된 다양한 독립변수(보건분야 예산액, 나이, 인구밀도 등)들을 사용하여 시군구별 코로나19 사망자 수 예측모형을 개발하는 것으로써 여러 독립변수를 사용하고자 다중회귀모형을 적용한다. 독립변수의 추가와 제거를 거쳐 최적의 회귀모형을 도출하는 단계별 회귀 방법을 이용하여 변수를 선택한다.

첫 번째 회귀분석 결과를 보면 Intercept의 계수가 -11.9730가 산출된다. 코로나19 사망자의 수는 음수가 될 수 없기에 절편을 0으로 고정해준다. 그리고 두 번째 회귀분석 수행 결과, 결정계수(R-Squared)는 0.757, 조정된 결정계수(Adjusted R-Squared)는 0.741, F-통계량(F-Statistic)은 46.31로 도출되었으며 그 외 분석결과는 표 5와 같다.

표 5. 코로나 사망자 2차 회귀분석 결과
Table 5. Results of the second regression analysis of COVID-19 deaths

Var.	Coef.	P > |t|
Health sector budget per person (KRW)	4..938e-06	0.196
Age	0.1608	0.032
Population density(person/㎢)	0.2072	0.025
Aging rate(%)	-0.1731	0.01
Smoking rate(%)	-0.0595	0.146
Drinking rate(%)	-0.0653	0.024
Medical facilities per person	-0.5886	0.058
Medical staff per person	0.5174	0.066

다음으로 p-값이 0.1 이상의 변수들을 제거해준다. 따라서 보건분야 예산액, 흡연율 변수를 제거해준 뒤 남은 독립변수들로 세 번째 회귀분석을 수행한다. 분석결과, 결정계수는 0.749, 조정된 결정계수는 0.737, F-통계량은 60.34로 도출되었으며 그 외 분석 결과는 표 6과 같다.

표 6. 코로나 사망자 3차 회귀분석 결과
Table 6. Results of the third regression analysis of COVID-19 deaths

Var.	Coef.	P > |t|
Age	0.1174	0.093
Population density(person/㎢)	0.1875	0.012
Aging rate(%)	-0.1157	0.054
Drinking rate(%)	-0.0658	0.024
Medical facilities per person	-0.7130	0.016
Medical staff per person	0.6019	0.031

독립변수들의 p-값이 모두 0.1 미만으로 90% 신뢰 검정 결과 유의미하게 나온 것을 알 수 있다. 하지만 독립변수 중 고령화 비율, 음주율, 의료진은 계수부호가 의미상으로 맞지 않아 추가 분석이 필요하여 제외하기로 하였다. 따라서 본 논문의 분석 결과를 종합하면 나이가 많을수록, 인구밀도가 높을수록, 의료시설이 적을수록 시군구별 사망자 수가 높다는 것을 보여주고 있다.

상기에서 도출한 회귀분석 결과들로부터 종합적으로 검토한 결과, 시군구별 코로나19 사망자 수 예측 모형을 식 (1)과 같이 유도하였다.

$$\ln(Y) = 0.1174(X_1) + 0.1875(X_2) - 0.713(X_3) \quad (1)$$

주 : Y=시군구별 코로나19 사망자(명)
X_1=나이
X_2=인구밀도(명/㎢)
X_3=1인 대비 의료시설 현황

이 모형의 검정통계량에서 결정계수는 0.749, 조정된 결정계수는 0.737로 모델이 적합하다고 볼 수 있으며, F-통계량은 60.34로 높게 나왔으며 모형이 통계적으로 적절하였음을 알 수 있다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 코로나19 사망 위험에 영향을 미치는 지역 특성 요인으로 나이, 인구밀도, 의료시설을 도출하였고 이를 다중 선형회귀분석 기법을 사용하여 시군구별 사망자 수 예측을 위한 모형을 생성하였다. 이 모형을 검정한 결과 결정계수는 0.749였으며 모형이 적합하다고 판단된다.

본 논문의 분석 결과가 지역별 의료 자원 배분의 효율성 증진과 정책 활용에 도움이 될 것으로 기대한다. 본 논문은 계속해서 추가적인 개인 특성과 지역 특성을 함께 고려하여 코로나19 사망 요인에 대해 종합적으로 분석할 계획이다.

Acknowledgement

본 논문은 2021년 한국정보기술학회 하계종합학술대회 및 대학생논문경진대회에서 “시군구별 코로나19 사망자 수 예측을 위한 선형회귀분석[15]”로 발표된 논문을 확장한 것이다.

References

[1] J. S. Yoo, "Review and comparison of national responses and health policies for the COVID-19 Pandemic", PhD Thesis, Dec. 2020.

[2] "COVID-19 Real-Time Situation Plate", https://coronaboard.kr/ [accessed: May 21, 2021]

[3] C. J. Han, "Determinants of COVID-19 Preventive Behaviour among University students in the Seoul Metropolitan Area", Master’s Thesis, Dec. 2020.

[4] "Korea Centers for Disease Control and Prevention. Cases in Korea", http://ncov.mohw.go.kr/bdBoardList_Real.do?brdId=1&brdGubun=11&ncvContSeq=&contSeq=&board_id=gubun= [accessed: May 21, 2021]

[5 D. H. Kim and S. J. Park, "Factors related to COVID-19 Incidence and Mortality rate in Gyeongsangbuk-do, Korea", Korean Society for Agricultural Medicine and Community Health, Vol. 45, No. 4, pp. 235-244, Dec. 2020.

[6] Y. R. Kim, S. H. Nam, and S. R. Kim, "Impact Factors and Validity of Blood Variables on Death in COVID-19 patient: Using Data of Korea Disease Control and Prevention Agency", Journal of the Korea Society of Computer and Information, Vol. 25, No. 11, pp. 179-185, Nov. 2020.

[7] M. Valero and J. N. Valero-Gil, "Determinants of the number of deaths from COVID-19: differences between low-income and high-income countries in the initial stages of the pandemic", SSRN Electronic Journal, Jun. 2020.

[8] C. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regressionanalysis and reporting", The Journal of Educational Research, Vol. 96, No. 1. pp. 1-15, 2002.

[9] "Simple and Multiple Linear Regression in Python", https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9 [accessed: May 21, 2021]

[10] "COVID-19 regional characteristics", https://kosis.kr/index/index.do [accessed: Mar. 2, 2021]

[11] "COVID-19 outbreak status", https://www.data.go.kr/ [accessed: Mar. 2, 2021]

[12] "Appropriation Budget by Function by Structure",

https://lofin.mois.go.kr/portal/service/openInfPage.do?infId=RL5CZ30XLWDXZAKXS5LP134169 [accessed: Mar. 2, 2021]

[13] "Resident registration population and other status", https://jumin.mois.go.kr/etcStatAvgAge.do [accessed: Mar. 2, 2021]

[14] "Regional Population and Population Density", https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1007 [accessed: Mar. 2, 2021]

[15] G Kim, J. H. Bae, and J. H. Lee, "Linear regression analysis for predicting the number of deaths due to COVID-19 by city, province, and district", Proceedings of The 2021 Summer Conference of the KIIT, Jeju, Korea, Jun. 2021.

저자소개

김 건 (Geon Kim)

2019년 3월 ~ 현재 : 대구가톨릭대학교 인공지능·빅데이터공학과 학사과정

관심분야 : 빅데이터 분석, 인공지능

배 지 훈 (Ji-Hoon Bae)

2000년 2월 : 경북대학교 전자전기공학부(공학사)

2002년 2월 : 포항공과대학교 전자·컴퓨터공학부(공학석사)

2016년 2월 : 포항공과대학교 전자·전기공학과(공학박사)

2002년 2월 ~ 2019년 8월 : 한국전자통신연구원 책임연구원

2019년 9월 ~ 현재 : 대구가톨릭대학교 인공지능·빅데이터공학과 조교수

관심분야 : 인공지능, 딥러닝/머신러닝, 레이다 영상 및 신호처리, 최적화 기법

이 종 혁 (JongHyuk Lee)

2004년 2월 : 고려대학교 컴퓨터교육과(이학사)

2006년 2월 : 고려대학교 컴퓨터교육학과(이학석사)

2011년 2월 : 고려대학교 컴퓨터교육학과(이학박사)

2011년 3월 ~ 2011년 10월 : 고려대학교 정보창의교육연구소 연구교수

2011년 11월 ~ 2012년 11월 : University of Houston Post-Doc. 연구원

2012년 12월 ~ 2017년 8월 : 삼성전자 책임연구원

2017년 9월 ~ 현재 : 대구가톨릭대학교 인공지능·빅데이터공학과 조교수

관심분야 : 클라우드 컴퓨팅, 빅데이터, 인공지능