JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 11: 1351~1356, Nov. 2019

빅데이터 기반 만성질환자의 삶의 질에 미치는 영향분석

김민경[1] · 조영복[2*]

An Analysis of Impact on the Quality of Life for Chronic Patients based Big Data

Min-kyoung Kim[1] · Young-bok Cho[2*]

[1]Senior Researcher, Department of Research & Development Cent, SONOUM Inc., Chungbuk 28501 Korea
[2*]Assistant Professor, Department of Information Security, Daejeon University, Daejeon 34520 Korea

요 약

본 연구는 빅데이터 플랫폼을 이용해 만성질환자에 따른 개인적 요인과 지역사회요인이 삶의 질에 미치는 영향을 알아보는데 목적이 있다. 연구방법은 2017년 지역사회건강조사 자료와 통계청 시군구별 2차 자료를 사용하였고, EQ-5D 지수와 개인요인 및 지역사회요인을 구분하여 다수준분석을 실시하였다. 연구결과 남자의 경우, 나이가 어릴수록, 학력이 높을수록, 월가구소득이 많을수록, 경제활동을 하는 경우, 스포츠 인프라가 많은 경우 삶의 질이 높았다. 또한 주관적 건강감이 나쁠수록, 스트레스가 많을수록 삶의 질이 낮았다. 향후 의료 빅데이터 분석을 위해 클라우드와 오픈소스를 활용할 수 있는 하드웨어에 독립적인 플랫폼 제공을 위한 연구가 지속되어야 할 것이다.

ABSTRACT

The purpose of this study is to investigate the effect of personal factors and community factors on the quality of life based on the presence of chronic patients based on the Big Data Platform. As a method of study, second data of 2017 community health survey and Statistics Korea by City·Gun·Gu public office were used and a multi-level analysis was conducted after separating EQ-5D index, individual factor and community factor. As a result, men, age, education level, monthly household income, having economic activity, the number of sports infrastructure were positively associated with the quality of life, and subjective health not good, extremely perceived stress were negatively associated with the quality of life. Research will continue to provide a platform independent of hardware that can utilize the cloud and open source for medical big data analysis in the future.

키워드 : 빅데이터, 하둡, 만성질환자, 삶의 질, 다수준회귀분석

Keywords : Big Data, Hadoop, Chronic patients, Quality of Life, Multi-level regression

Received 12 July 2019, **Revised** 26 July 2019, **Accepted** 11 August 2019

*** Corresponding Author** Young-Bok Cho(E-mail:ybcho@dju.ac.kr, Tel:+82-42-280-2406)
Assistant Professor, Department of Information Security, Daejeon University, Daejeon 34520 Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.11.1351 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 4차 산업혁명이 도입되면서 빅데이터를 활용한 융합을 통해 사회·경제·보건 전반에 많은 변화가 나타나고 있다. 빅데이터란 기존 데이터베이스의 데이터 저장·관리·분석 능력을 초과하는 다양한 형식을 가진 대량의 데이터를 의미한다[1]. 다양한 분야에 많은 데이터들이 생산됨에 따라 특히 빅데이터를 활용한 보건의료에 많은 관심을 보이고 있다. 국민의 평균 수명 연장과 만성질환 유병률 증가로 인해 많은 의료비 지출에 대한 문제가 논의되면서 IT와 의료기술을 접목한 U-Health 도입도 추진하고 있다.

우리나라는 급격한 인구의 고령화, 생활습관의 변화, 환경오염 등으로 만성질환이 증가하고 있다. 만성질환은 3개월 이상 지속되는 증상으로 완치가 어려우며 장기간 관리가 필요한 질환이다[2]. 우리나라는 만성질환으로 인한 사망과 질병부담이 높은 상황으로, 만성질환은 전체 사망의 80.8%를 차지하며, 사망원인 10 중 7개가 만성질환이다[3]. 2017년 기준 만성질환자는 약 1,730만명(전체 인구의 33.6%), 만성질환 진료비는 28.2조원(고혈압·당뇨병 5.3조원)으로 전체 진료비(69조원)의 41%이다[4].

OECD(2010) 보고에 의하면, 만성질환은 전 세계적으로 장애와 사망의 주된 원인으로, 전세계인구의 60%가 만성질환으로 사망하고 있는 것으로 추정된다. 이렇게 만성질환자 증가로 인해 사회경제적 부담 또한 증가하고 있다. 건강보험자료에 의하면 고혈압 및 당뇨병으로 인한 진료비는 건강보험 재정의 1, 2위를 차지하고 있다. 고혈압으로 인한 총 진료비는 2008년에 2조 998억원으로 2002년 대비 2.5배 증가하였으며, 당뇨병으로 인한 총 진료비는 동 기간에 2.2배 증가한 1조 1,276억원에 달하였다. 만성질환은 대부분 완치되지 않는 경우가 많기 때문에 삶의 질에 영향을 미칠 것으로 본다.

또한 산업들이 융·복합된 많은 양의 데이터들이 생산되면서 삶의 질에 대한 욕구도 증가하면서 건강관련 삶의 질에 대한 중요성이 부각되기 시작하였다. WHO에서 삶의 질은 한 개인이 살고 있는 문화권과 가치체계의 맥락 안에서 자신의 목표, 기대, 규범, 관심과 관련하여 인생에서 자신이 차지한 상태에 대한 개인적인 지각이라고 정의하였다. 영국의 한 연구에서는 삶의 질은 개인이 가지고 있는 특성과 지역주민이 생활하고 있는 환경적 특성의 상호작용에 의해 영향을 받는다고 하였다[5]. 따라서 삶의 질은 복합적으로 작용하기 때문에 삶의 질을 측정하기 위해서는 객관적, 주관적 지표가 모두 활용하는 것이 정확하다[1]. 기존의 연구들은 특정질환 및 특정 지역에 한하여 개인적 요인 중심으로 연구가 한정되어 왔으며, 빅데이터를 활용한 연구는 아직 미비한 실정이다. 이에 본 연구는 전국을 대표하는 지역사회건강조사 자료와 통계청 자료를 이용하여 만성질환자의 삶의 질에 미치는 요인을 분석하고자 한다.

Ⅱ. 관련연구

2.1. 의료 빅데이터 현황

국내는 보건의료 데이터는 공공 영역과 민간 영역에서 수집되고 있다. 공공 데이터는 보건복지부와 기타 부처가 관할하며 내용과 수집 방식을 고려하여 유전체 데이터, 청구·행정 데이터, 조사데이터로 구분할 수 있다. 민간 영역에서는 의료기관이 환자 진료 과정에서 수집한 임상데이터와 개인의 선택에 의해서 소셜 네트워크 서비스 또는 모바일 장치 등을 통해 수집되는 스트림데이터 등으로 구분될 수 있다. 높은 수준의 IT기술 활용으로 공공과 민간 영역 모두 상당한 수준의 규모와 다양성으로 데이터가 구축되고 있으나 국가적으로 보건의료 빅데이터를 구축하고 활용하는 데는 한계가 있다.

첫째, 기관별로 분산된 보건의료 데이터가 상호연계·통합되어 국가적으로 의미 있는 활용을 유도할 수 있는 법적, 기술적, 정책적 기전이 부족하다. 공공기관 내부의 데이터 통합은 추진되고 있으나 기관 간 데이터 연계는 법적으로 허용된 업무 수행을 위해서만 제한적으로 이뤄지고 있다. 특히 민간영역에서 보유하고 있는 의미있는 임상데이터가 국가적으로 연계되어 활용되는 제도적·물리적 기반이 구축되어 있지 못하다.

둘째, 국가 단위 빅데이터 구축의 한계는 오픈 데이터를 통한 새로운 가치 창출도 제한하고 있다. 새로운 가치를 창출하는 빅데이터는 데이터공개와 접근성 확대를 통해 기대할 수 있다. 개인정보 보호와 데이터 보안이 확인된 데이터는 제한적으로 이용된 일정 기간 이후에 익명화하여 공개함으로써 보건의료시장에서 최종 사용자인 국민의 편익을 높이는 다양한 서비스 상품 개발에 활용되어야 한다.

마지막으로, 유전체 데이터의 활용을 통한 질병 발생 기전 등의 임상지식 창출도 제한적이다. 최근 보건의료 분야에서 빅데이터는 보다 근원적인 질병 발생 기전을 분석하기 위한 유전체 데이터의 활용성을 강조하고 있음에도 유전자 정보의 규모 및 내용과 기타 데이터와 연계하는 정보의 완결성 측면에서 활용 가치를 기대하기 어렵다. 질병관리본부, 국립암센터 등이 한국인 표준 게놈(400명) 및 호발질환 유전자 분석(2만여명) 자료, 암통합 오믹스 자료 등을 구축하고 있지만, 외국에 비해 소규모이고 기타 정보와의 연계 제한으로 정밀의료와 맞춤의료 시대에 대비한 투자와 발전이 필요하다.

2.2. 빅데이터 플랫폼

빅데이터 플랫폼은 빅데이터에서 가치를 추출하기 위한 일련의 과정을 지원하기 위한 프로세스를 규격화한 기술 · 서비스 모임으로 빅데이터의 특성인 Volume과 비정형, 실시간성으로 인해 데이터의 수집과 저장, 분석 등의 모든 영역에서 학문적인 측면보다는 서비스 중심적인 개념으로 표현하는 경향을 보이고 있다. 빅데이터 플랫폼은 데이터에 대해 수집→ 저장 → 처리 → 분석 → 시각화등을 통해 원시데이터(Raw Data)로 부터 Insight 및 가치(Value) 추출하기 위해 분석, 시각화 측면에서는 가치 추출을 위해서 전통적인 통계 분석에서는 인과관계를 최종적인 결과로서 제시한다면, 빅데이터에서는 연관관계·상관관계를 중심으로 시사점을 도출하려는 경향을 보이고 있다[6,7].

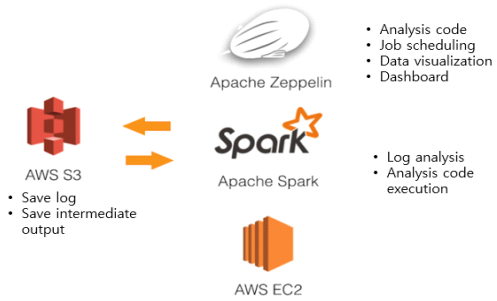

Fig. 1 Big data analysis system architecture

그림 1은 빅데이터 분석 구조를 나타낸 것으로 빅데이터를 다룰 때 가장 많이 쓰는 기술은 Hadoop MapReduce와 연관 기술인 Hive이다. Hadoop은 클러스터 컴퓨팅 프레임웍으로 컴퓨터를 여러 대 연결하면 대수에 따라서 데이터 처리 성능이 스케일되는 기술이다[8]. MapReduce는 슈퍼컴퓨터 없이도 서버를 여러대 연결하여 빅데이터 분석을 가능하게 해준 혁신적인 기술이다. Apache Spark는 Hadoop MapReduce와 비슷한 목적을 해결하기 위한 클러스터 컴퓨팅 프레임웍으로, 메모리를 활용한 아주 빠른 데이터 처리가 특징이다. 또한, 함수형 프로그래밍이 가능한 언어인 Scala를 사용하여 코드가 매우 간단하며, interactive shell을 사용할 수 있다.

Ⅲ. 빅데이터 분석을 통한 삶의 질 측정

3.1. 데이터 분석 환경

만성질환자의 삶의 질에 미치는 요인을 파악하기 위해 2017년 지역사회건강조사자료를 바탕으로 통계청 자료를 시군구 지역으로 매칭하여 데이터베이스를 구축하였다. 지역사회건강조사 자료는 2008년부터 질병관리본부 주관으로 매년 실시하는 전국 표본조사로 254개 보건소 관할 지역에서 수행된다. 통계청 자료는 우리나라를 대표하는 자료로 국민 전체의 지역특성을 파악하는데 대표적인 자료에 해당된다.

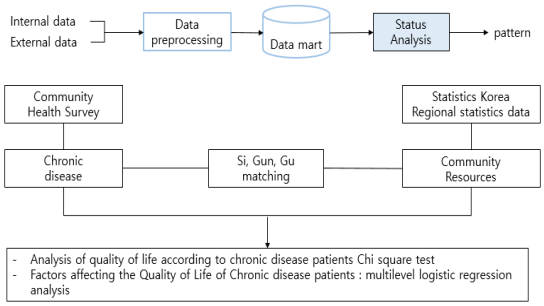

Fig. 2 Select fields for data integration

그림 2는 위에서 설명한 지역사회건강조사자료와 통계청 자료 분석을 위해 통합하고 분석기준을 도식화한 것이다. 그림 3의 Driver Program은 여러 개의 병렬적 작업으로 Worker Node에 있는 Executor에서 실행을 담당하고 SparkCotext는 메인 시작 지점으로 스파크 API를 활용하기 위해 필요하고 클러스터의 연결을 보여주고 RDD를 만드는데 사용된다[6].

Cluster Manager는 Standalone,YARN, Mesos 등 클러스터의 자원을 관리하는 관리자 역할을 수행하고, Worker Node는 하드웨어 즉 서버로 하나의 물리적 장

치에 여러 개 사용이 가능하다. 마지막으로 Executor는 프로세스 즉 하나의 워커 노드로 여러 개 사용이 가능하다. 제안 논문의 스파크를 활용한 데이터의 분산 처리는 빅데이트 분석을 위한 클라우드 그리드 형태의 분산 인덱스 정보를 생성하고 RDD상에 저장한 다음 해당 인덱스 정보를 재사용해 삶의 질에 영향을 미치는 요인 분석을 수행한다.

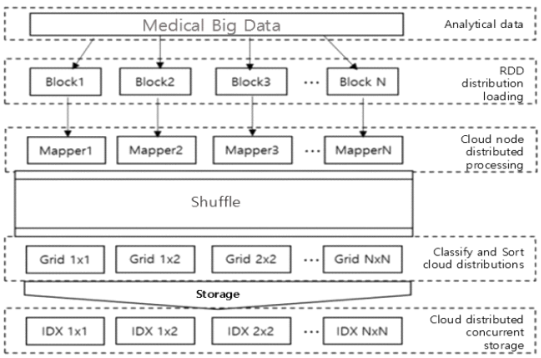

Fig. 3 RDD distribution structure in spark

3.2. 자료 분석 환경

수집된 자료는 빅 데이터 플랫폼을 통해 분류하고 통계적 처리는 SAS 9.4 프로그램을 이용하였다. 삶의 질을 측정하는 방법에는 표준도박법(standard gamlbe:SG), 시간교환법(time trade-off:TTO), 시각아날로그척도(visual analogue scale:VAS), 단축형 건강관련 삶의 질 척도(shortform-12 health survey questionnaire:SF-12)등이 있으며, 본 논문은 EuroQol Group이 개발한 지표로 EQ-5D를 사용하였다. EQ-5D는 광범위한 건강상태 및 치료의 평가에 이용할 수 있어 국가간, 지역간 비교를 위해 널리 사용되고 있다[9]. EQ-5D는 운동능력(mobility), 자기관리(self-care), 일상 활동(usual activity), 통증 및 불편(pain/discomfort), 불안 및 우울(anxiety/depression) 5가지 영역을 EQ-5D index 하나의 값으로 전환하기 위해 한국인을 대상으로 한 질병관리본부의 가중치 (식1)을 이용하였다[10].

Table. 1 Multi-level analysis of chronic patients

Factor		Estimate	SE	Factor		Estimate	SE
Intercept		1.6***	0.53	Economic activity	Yes	3.6***	0.36
Personal Factor				No(ref)			
Gender	Men	1.8***	0.32	Spouse	Yes	1.1	1.35
	Women(ref)			No(ref)			
Age	19-29(ref)			Subjective health	Good(ref)		
	30-39	0.8	1.97	Usually		-2.0***	0.40
	40-49	0.8	1.89	Not good		-13.6***	0.43
	50-59	-0.4	1.77	Perceived stress level	Not at all(ref)		
	60-69	-1.2	1.77	Not very		-1.8***	0.35
	70 +	-4.8***	1.78	Very		-6.7***	0.43
Education level	Ineducation(ref)			Extremely		-16.6***	0.77
	Elementary school	2.5***	0.49	Community Factor			
	Middle school	2.9	0.54	Population density		0.01	0.03
	High school	4.3***	0.54	Elderly's retention rates		-0.03	0.10
	Over college	5.0***	0.65	Doctors engaged in medical institutions		0.04	0.09
Monthly family income (unit 1, 000 won)	Under 999 won(ref)			Financial independence		-0.04	0.02
	1,000-2,999won	2.8***	0.35	Umber of cultural infrastructure		-0.05	0.07
	3,000-4,999won	3.0***	0.58	Umber of sports infrastructure		0.04*	0.02
	Over 5,000won	2.1**	0.86				
AIC				71603.4			
BIC				71668.2			

$$1-(0.05+0.096\times M2+0.418\times M3+0.046\times SC2+0.136\times SC3+0.051\times UA2+0.208\times UA3+0.037\times PD2+0.151\times PD3+0.043\times AD2+0.158\times AD3+0.05\times N3) \quad (1)$$

M2 : 운동능력 수준 2, M3 : 운동능력 수준 3
SC2 : 자기관리 수준 2, SC3 : 자기관리 수준 3
UA2 : 일상활동 수준2, UA3 : 일상활동 수준3
PD2 : 통증/불편감 수준2, PD3 : 통증/불편감 수준3
AD2 : 불안/우울 수준2, AD3 : 불안/우울 수준3
N3 : 수준3이 하나라도 있을 경우

데이터 통합 후 분석을 위한 데이터 분석 하둡 환경은 그림 4에서와 같다. 만성질환자에 따라 인구사회학적 특성 차이를 보기 위하여 카이제곱검정을 실행하였고, 삶의 질과 지역사회요인 파악은 상관분석을 이용하여 통계적으로 유의한 영향을 미치는 요인만을 선별하였다. 최종적으로 만성질환자에 따른 삶의 질에 영향을 미치는 요인을 파악[11]하기 위하여 다수준 회귀분석(multilevel regression analysis)을 실행하였으며, 개인수준의 변수들의 효과는 고정효과로 집단 수준에 의한 변동만이 만성질환자의 삶의 질에 영향을 미치는 것으로 가정하여 모형에 적용하였다.

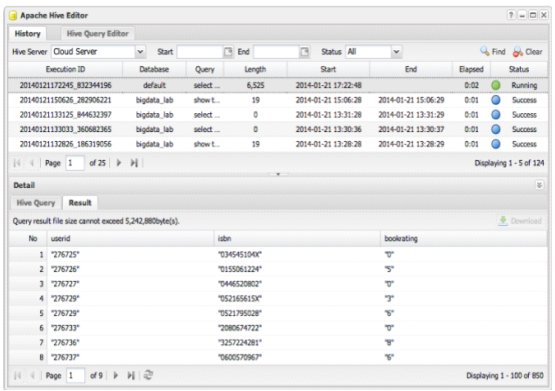

Fig. 4 Data analysis in Hadoop

그림 4는 수집된 자료는 하둡 환경에서 의료데이터를 분석하기 위해 통계 프로그램을 이용하여 분석하는 단계를 나타낸 것이다.

- level 1: 개인적 요인으로 성별, 연령, 교육수준, 월 가구소득, 경제활동여부, 배우지 유무, 주관적 건강, 지각된 스트레스 수준
- level 2: 지역사회요인으로 인구밀도, 고령인구 비율, 의료기관 종사 의사 수, 재정자주도, 문화기반 시설 수, 체육시설 수로 다음 (2)와 같이 정리할 수 있다. (2)의 Y는 종속변수이고 X는 독립변수 그리고 a와b는 회귀계수를 의미한다.

$$Y=a+b_1X_1+b_2X_2+b_3X_3+..+b_nX_n \quad (2)$$

위 표 1은 그림4를 이용해 분석 결과를 나타낸 것이다. 표 1은 집단 내 변동과 더불어 집단 간 변동까지 고려하는 모형으로 다수준분석의 결과이다. AIC, BIC 모두 모형의 적합도를 보는 것으로 좋은 모형으로 나타났다.

Ⅳ. 결 과

이 논문에서는 빅데이터 플랫폼 스파크를 이용해 분산 환경에서 지역사회 건강 조사표와 통계청 자료를 연계한 만성질환자에 따른 삶의 질을 분석하였다. 분석 결과 성별에 따라 남자의 경우 삶의 질이 높고, 나이가 적을수록 삶의 질이 높았고, 학력이 높을수록, 수입이 많을수록, 경제활동을 하는 경우 삶의 질이 높았다. 또한 주관적 건강감이 좋지 않은 경우, 지각하는 스트레스가 매우 많은 경우 삶의 질이 낮았고, 스포츠 인프라가 많을수록 삶의 질이 높았다. 이와 같이 많은 변수적 속성을 지닌 의료 건강데이터를 빅데이터 분석을 위한 클라우드 환경에서 메모리 등 하드웨어의 종속 없이 활용할 수 있음을 증명하였다. 그러나 본 연구의 데이터는 횡단적 단면조사연구이므로 개인 요인과 지역사회 요인에 따른 삶의 질의 인과관계를 정확히 파악하기 어렵고 관련성만 볼 수 있었다. 보건의료 빅데이터와 이를 활용한 다양한 인공지능의 등장으로 데이터에 기반한 임상적 의사결정(Clinical Decision)과 근거중심의학(Evidence-Based Medicine)이 보편화 될 것이며, 더 나아가 인간이 건강을 바라보는 관점과 삶의 방식마저 많은 부분에서 변화시킬 것으로 예상된다. 향후 종단연구를 활용한 의료 빅데이터 분석을 위해 클라우드와 오픈소스를 활용할 수 있는 하드웨어에 독립적인 플랫폼 제공을 위한 연구가 지속되어야 할 것이다.

REFERENCES

[1] M. K. Kim, and Y. B. Cho, “An analysis of Factors Affecting Quality of Life through the analysis of Public Health Big Data,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no.6, pp. 835-841, Jun. 2018.

[2] Korean society and trends : Chronic disease trend and management, pp. 208-215, 2010.

[3] Korea Centers for Disease Control and Prevention [Internet] Available: http://www.cdc.go.kr/CDC/notice/.

[4] Pilot project of primary medical chronic disease management [Internet]. Available: http://www.mohw.go.kr.

[5] A. Bowing, Z. Gabriel, J. Dykes, O. Evans, A. Fleissing, D. Banister, and S. Sutton, “Let’s ask them: A national survey of definition of quality of life and its enhancement among people aged 65 and over,” *International Journal of Aging and Human Development*, vol. 56, no. 4, pp. 269-306, 2003.

[6] M. K. Kim, and Y. B. Cho, “An Analysis of Factors Affecting Quality of Life through the Analysis of Public Health Big Data,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no. 6, pp. 835-841, Jun. 2018.

[7] M. H. Park, Y. B.Cho, S. Y. Kim, J. B. Park, and J. H. Park, “Analysis of Factors for Korean Women's Cancer Screening through Hadoop-Based Public Medical Information Big Data Analysis,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no.10, pp.1277-1286, Oct. 2018.

[8] Y. B. Cho, S. H. Woo, and S. H. Lee, “The Big Data Analysis and Medical Quality Management for Wellness,” *Journal of the Korea Society of Computer & Information*, vol. 19, no. 12, pp. 101-109, Dec. 2014.

[9] N. Anna, and A. Tosteson, “Preference-based health outcome measures on low back pain,” *Journal of the Spine*, vol. 25, no. 24, pp. 3161-3166, Dec. 2000.

[10] South Korean time trade-off values for EQ-5D health states [Internet] Available: http://www.cdc.go.kr/CDC/info/.

[11] Y. B. Cho, S. H. Lee, J. H Park, and M. H Park, “KMMQL-AF-based evaluation of the quality of life for survivors of childhood cancer by age,” *Journal of Convergence for Information Technology*, vol. 6, no. 3, pp. 71-77, Jun. 2016.

김민경(Min-Kyoung Kim)

2006: 연세대학교 보건학 석사
2019: 충북대학교 의학 박사
현 재: ㈜소노엠 기업부설연구소 책임연구원 및 협성대학교 보건관리학과 겸임교수

※ 관심분야: 빅데이터, 건강증진, 원격의료

조영복(Young-Bok Cho)

2005: 충북대학교 전자계산학과 공학석사
2012: 충북대학교 전자계산학과 공학박사
2019: 충북대학교 의학과 의학박사
2012-2018: 충북대학교 소프트웨어학과 초빙교수
현 재 : 대전대학교 정보보안학과 조교수

※관심분야: 의료영상처리, 정보보안, 의료정보보호, 모바일보안