http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.24, No.1, February 2019, pp.105-120
https://doi.org/10.7838/jsebs.2019.24.1.105

딥러닝 기반 항생제 내성균 감염 예측

Antibiotics-Resistant Bacteria Infection Prediction Based on Deep Learning

오성우(Sung-Woo Oh)*, 이한길(Hankil Lee)**,
신지연(Ji-Yeon Shin)***, 이정훈(Jung-Hoon Lee)****

초 록

세계보건기구(WHO)를 비롯해 세계 각국의 정부기관은 항생제 오남용에 따른 항생제 내성균 감염에 대해 심각하게 경고하며 이를 예방하기 위한 관리와 감시를 강화하고 있다. 하지만 감염을 확인하기 위한 감염균 배양에 수일의 시간이 소요되면서 격리와 접촉주의를 통한 감염 확산 방지 효과가 떨어져 선제적 조치를 위한 신속하고 정확한 예측 및 추정방법이 요구되고 있다. 본 연구는 Electronic Health Records에 포함된 질병 진단내역과 항생제 처방내역을 neural embedding model과 matrix factorization을 통해 embedding 하였고, 이를 활용한 딥러닝 기반 분류 예측 모형을 제안하였다. 항생제 내성균 감염의 주요 원인인 질병과 항생제 정보를 embedding 하여 환자의 기본정보와 병원이용 정보에 추가했을 때 딥러닝 예측 모형의 f1-score는 0.525에서 0.617로 상승하였고, 딥러닝 모형은 Super Learner와 같은 기존 기계학습 모형보다 더 나은 성능을 보여주었다. 항생제 내성균 감염환자의 특성을 분석한 결과, 감염환자는 동일한 질병을 진단받은 비감염환자에 비교해 J01 계열 항생제 사용이 많았고 WHO 권고기준(DDD)을 크게 벗어나는 오남용 청구사례가 6.3배 이상 높게 나타났으며 항생제 오남용과 항생제 내성균 감염 간의 높은 연관성이 발견되었다.

ABSTRACT

The World Health Organization (WHO) and other government agencies aroundthe world have warned against antibiotic-resistant bacteria due to abuse of antibiotics and are strengthening their care and monitoring to prevent infection. However, it is highly necessary to develop an expeditious and accurate prediction and estimating method for preemptive measures. Because it takes several days to cultivate the infecting bacteria to identify the infection, quarantine and contact are not effective to prevent spread of infection. In this study, the disease diagnosis and antibiotic prescriptions included in Electronic Health Records were embedded through neural embedding model and matrix factorization, and deep learning based classification predictive model was proposed. The f1-score of the deep learning model increased from 0.525 to 0.617when embedding information on disease and antibiotics, which are the main causes of antibiotic resistance, added to the patient's basic information and hospital use information. And deep learning model outperformed the traditional machine

* First Author, Graduate School of Information, Yonsei University(supoman@yonsei.ac.kr)
** Co-Author, College of Pharmacy, Yonsei University(hankil.lee@yonsei.ac.kr)
*** Co-Author, Graduate School of Information, Yonsei University(yeond289@gmail.com)
**** Corresponding Author, Graduate School of Information, Yonsei University(jhoonlee@yonsei.ac.kr)
Received: 2019-02-11, Review completed: 2019-02-18, Accepted: 2019-02-22

hospital use information. And deep learning model outperformed the traditional machine learning models.As a result of analyzing the characteristics of antibiotic resistant patients, resistant patients were more likely to use antibiotics in J01 than nonresistant patients who were diagnosed with the same diseases and were prescribed 6.3 times more than DDD.

키워드 : 항생제 내성, 딥러닝, 뉴럴 임베딩 모델, 행렬 분해, 전자의무기록
Antibiotic Resistance, Deep Learning, Neural Embedding Model, Matrix Factorization, Electronic Health Records

1. 서 론

세계보건기구(World Health Organization)에서는 항생제 내성(Antibiotics Resistance)을 세계 공공보건을 위협하는 심각한 이슈로 규정하고 있으며, 전 세계 매년 70만 명 이상이 항생제 내성균에 감염되면서 사망에 이르고 있다고 발표하였다. OECD가 발표한 주요국 항생제 사용 현황에 따르면, 대한민국은 34개 회원국 중 그 수치가 가장 높은 것으로 나타났다. 실제로 항생제 오남용에 따른 국가별 항생제 내성 현황에서도 가장 높은 수치를 기록하였다.

항생제 내성은 박테리아 감염을 예방하고 치료하는데 사용되는 의약품인 항생제 사용에 따라 박테리아가 변하면서 내성이 생기게 된 것을 말한다[20]. 이러한 내성균이 생성됨에 따라 항생제가 제대로 작용하지 못하며, 감염균이 발생할 가능성이 높아지게 된다. 항생제의 치료 효과가 없어지기 때문에 의료비 지출 증가 및 장기 입원을 초래하며 종래에는 사망에 이르게 된다[32].

항생제 내성이 가속화되는 이유는 항생제의 오남용이 심하다는 문제가 있지만, 단기간에 항생제 내성을 증가시킨 주요 원인은 내성균이 빠르게 전파 확산하는 것에 있다. 특히나 병원이나 요양원 같은 의료시설에서 직간접적으로 내성균이 전파 확산하는 경우가 많다[25]. 감염된 환자가 만진 난간이나 휠체어, 바닥에 묻은 감염균을 통해서도 감염이 되기 때문에[21] 병원에서의 감염에 대한 관리가 효과적으로 이루어지지 않으면 항생제 내성균 감염이 빠르게 이루어진다[32]. 이외에도 특정 유전자나 기타 선천적인 요인에 의해서도 감염이 발생하는 것으로 알려져 있다[20].

국내는 감염 예방을 위해 관련 지침 매뉴얼을 지속 개발하였고 항생제 내성균 감염이 의심되는 고위험군 환자에 대해서는 모니터링을 통해 격리 조치가 시행되었다. 환자에 대한 격리 여부를 판단하기 위해서는 항생제 내성균 감염이 있는지 확인해야 하며, 이때 임상검체 배양 및 감수성 테스트를 진행한다[36]. 그 결과가 음성 판정으로 나타나면 감염된 환자를 격리하고 타인과의 접촉이 일어나지 않도록 관리되어야 한다[25].

하지만 배양 결과를 확인하기까지 평균적으로 5.2일이 걸리며 이 기간은 감염 확산을 막기에는 긴 기간이고[21], 결과를 확인한 후에 격리 및 접촉주의를 시행했을 때는 감시의 효과가 떨어지게 된다. 검사 결과가 나오기 이전에 선제적으로 격리하고 접촉주의를 시행하는 것은 윤리적인 문제가 생길 여지가 있고 기타 사회적, 경제적 비용을 발생시키게 된다[30].

따라서 항생제 내성균 감염에 대한 조기 발견과 명확한 진단을 통해 적절한 관리와 감시가 이루어질 수 있도록 신속하고 정확하게 검사 결과를 통보하기 위한 검사시스템 구축이 요구되는 상황이다[25].

의료 분야에서는 오래전부터 통계적 모형을 통해 질병에 대한 특성 분석과 예측 모형(Predictive Modeling)을 개발하여 질병 진단을 확인하는데 활용했다. 컴퓨터 성능이 발달하고 모델을 학습하기 위한 알고리즘들이 크게 개선되고 데이터가 많이 쌓이게 되면서 최근 2년 동안에는 기계학습을 활용한 개인의 질병 유병 혹은 발병에 대해서 예측하는 방법에 대한 연구가 급속하게 늘어났다.

항생제 내성균 감염은 조기진단이 성공적으로 진행되었을 때의 효과가 큰 분야이다. 현재 대부분의 항생제 내성균 감염 예측을 위한 연구는 유전자 데이터를 중심으로 기계학습을 통한 예측을 시도하고 있으며, 이는 전통적인 접근 방법에 비해 더 좋은 예측 결과를 보여준다[23]. 하지만 대부분의 항생제 내성균 감염에 대한 연구 결과를 다른 유전자를 가진 국내의 연구에서 적용하기에는 어려운 부분이 있다.

항생제 내성균 감염의 가장 주된 원인은 항생제의 사용이지만 항생제 처방과 복용에 대한 정보를 사용하여 항생제 내성균 감염을 예측하는 연구는 많지 않다. 기계학습과 딥러닝을 활용한 연구가 다른 의료 영역에서 이미지와 유전자 분석에서 성과가 있기 때문에 비슷한 방법론을 주로 사용하는 상황이다. 하지만 유전자 sequence 분석을 통한 예측은 이미 감염을 통해 유전자 변형이 이루어진 이후에 대한 결과를 예측하는 것이므로 조기 발견과는 거리가 있다. 따라서 직접적으로 박테리아의 변형을 일으키는 항생제 사용이 중점적으로 고려될 필요가 있다.

한편 구글은 2018년 5월 Nature Medicine을 통해 Electronic Health Records(이하 EHR) 데이터를 통해 입원환자의 치료 결과를 예측하는 딥러닝 기반의 예측 모형을 발표하였다[28]. EHR 데이터는 다른 의료 관련 데이터와는 달리 실제적인 환자의 진료기록과 처방 기록 등이 저장되어 환자에 대한 보다 정확한 예후 파악이 가능하다. EHR 데이터를 이용한 연구가 활발해질 것으로 기대되는 가운데 국내에서는 건강보험공단(NHIS)과 심사평가원(HIRA)에서 병원을 이용한 환자들의 질병 진단과 약물 처방에 대한 EHR 데이터를 연구용으로 제공하고 있어 국내에서 EHR 데이터를 이용한 예측 모형 개발이 가능한 상황이다.

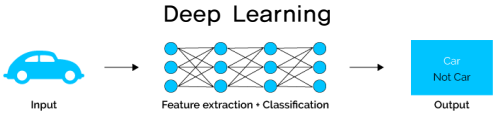
본 연구는 실제 항생제 내성균 감염의 원인이 되는 항생제 처방 정보를 포함하는 EHR 데이터를 활용하여 딥러닝 기반 항생제 내성균 감염을 예측하고자 한다. 그리고 감염 환자에 대한 탐색적 통계분석을 통해 감염 환자의 특성을 분석하고 항생제 오남용 현황을 파악하고자 한다.

2. 이론적 배경

2.1 Prediction of Antibiotic Resistance

딥러닝(Deep learning)이란 기계학습(Machine learning)의 한 분야이며[29], 기존의 사전 정의 과정이 없이 컴퓨터가 스스로 많은 데이터들을 분류하는 방식이다[15]. 딥러닝 기반 예측 모형

은 분류(classification) 예측 작업에서 기존 기계학습 방법보다 크게 향상된 결과를 나타낸다. 딥러닝의 예시로 <Figure 1>과 같이 다량의 자동차 이미지를 입력했을 때 컴퓨터가 스스로 이미지를 분류하여 학습하고 자동차가 아닌 이미지를 입력했을 때 분별해낼 수 있다. 의료 분야에서도 개인화된 맞춤 진단과 처방, 그리고 치료를 제공하기 위한 목적으로 이러한 접근 방법을 활용하기 시작했다[4, 17, 31].



〈Figure 1〉 Example of Deep Learning algorithm〔1〕

기존 연구에서는 폐결핵(tuberculosis) 환자의 항생제 내성균 감염을 예측하기 위해 합성곱 신경망(Convolution Neural Network: CNN) 기반의 딥러닝 모형을 통해서 환자들의 폐를 촬영한 CT 이미지를 분석하였다[10]. 다양한 항생제 내성 종류 별로 분류 예측을 하는 모형을 제안하였고 이미지에서 특징을 검출하는 hand-crafted SIFT 방법과 비교하여 스스로 특징을 찾아내는 딥러닝 모형이 더 나은 성능을 보였음을 밝혔다.

최근 유전자 서열 데이터를 분석하여 다양한 종류의 항생제 내성균 감염을 예측하는 연구가 다수 발표되었다. Virginia Tech의 연구자 Arango-Argoty et al.[2]은 항생제 내성 유전자를 포함하는 DNA 서열을 분석하여 항생제 내성을 분류 예측하는 딥러닝 기반의 AI 솔루션 DeepARG를 개발하였다. 이외에도 유전자 데이터를 활용하여 대장균(Escherichia coli)과 같이 특정 질병과 연관된 항생제 내성을 예측하기 위한 딥러닝 기반 예측 모형이 제안되었다[4].

2.2 Embedding-based Data Representation

의료 분야의 데이터는 환자의 인구통계학적 정보, 질병 진단내역 그리고 약물 처방 및 치료, CT 촬영 이미지, MRI 영상 등의 다양하고 방대한 데이터 안에 포함된 많은 변수를 가지고 있다. 특히 질병코드에 대한 변수나 약물 처방, 치료 등에 대한 변수는 범주형 변수(categorical feature)라고 하며, 데이터 처리 과정에서 변수의 특정 범주에 해당하면, 1 또는 0으로 표현하는 One-hot vector 방식으로 범주를 수치(ex. Aspirin = [0, 0, 1, 0])로 표현하게 되고 이때 범주의 수 k만큼 변수가 생성된다[24]. One-hot vector 방식은 쉽게 사용할 수 있으나 변수의 차원 공간(dimensional space)이 커지면서 많은 문제가 발생하게 된다.

질병이나 약물 변수에 대한 범주는 적게는 수백 개에서 많게는 수만 개이기 때문에 그만큼 변수가 생기며 너무나 높은 다차원 공간(High dimensional feature space)이 생성된다. 오래 전부터 기계학습 분야에서는 이러한 데이터를 “차원의 저주(Curse of Dimensionality)”에 빠져 있다고 이야기한다. 그리고 이러한 다차원 공간에서는 변수가 많아지면서 예측 모형의 성능이 감소하게 된다[35]. 따라서 변수의 수를 줄이거나(feature selection) 특징을 추출(feature extraction)하는 방법들을 통해서 저차원의 공간에서 변수를 매핑하고 의미 있게 표현(representation)하여 모형의 성능을 향상하게 시키면서 연산의 수를 줄이려는 방법들이 제시되고 있다.

이러한 방법들을 기계학습에서 embedding 이라고 부르며 저차원으로 데이터를 표현하면서도 원래의 정보를 포함하도록 한다. 데이터 안에서 서로 간의 의미를 포함할 방법들이 개발되고 있으며 개발된 하나의 embedding 방법은 다른 영역에서도 응용되어 사용되고 있다. 예를 들면, 자연어 처리에서 단어를 저차원의 벡터로 표현하기 위한 word embedding이 음성 인식이나 기계번역과 같은 영역에서 동일하게 사용되고 있다. 좀 더 효율적이고 향상된 embedding 방법들이 고안되면 다른 영역에서도 그러한 방법들을 사용할 수 있다.

2.2.1 Embedding with Matrix Factorization

추천 시스템(Recommender System)은 사용자에게 개인화된 맞춤 서비스 또는 아이템을 추천하기 위해서 사용자의 행동 특징이나 특정 아이템에 대한 선호 정보를 embedding한다. 대표적으로 콘텐츠 기반 필터링(Content-based Filtering)이나 협업 필터링(Collaborative Filtering)에서는 수많은 사용자와 수많은 아이템 간의 관계나 특징 정보를 추출하기 위해 행렬 인수분해 (Matrix Factorization) 방법을 사용하고[16], 사용자와 아이템의 잠재적 특성(Latent feature)을 추출하는 형태로 embedding을 한다[19]. 자연어처리 분야에서는 단어(word) 간의 유사성(similarity)을 추출하기 위해 embedding이 사용되기도 한다[9].

행렬 인수분해는 사용자와 아이템 간의 관계가 표현된 행렬 A를 K 차원의 특성 행렬 U와 I로 분해한다. 사용자의 수를 m, 아이템의 수를 n이라 할 때, 행렬 분해는 행렬 A에 근사하는 행렬을 찾는 것이 목표이며, 그 공식은 $\tilde{A} = UMT$이다.

사용자의 행동 정보는 주로 동시 출현 행렬(Co-occurrence Matrix)의 형태로 사용자의 행동 빈도나 선호 점수 등이 행렬로 표현되며, 단어가 문장 혹은 단락 안에서 몇 번 출현했는지 그 빈도를 단어 별로 동시 출현 행렬의 형태로도 표현한다. 추천 시스템과 자연어 처리에서 동시 출현 행렬을 인수분해하여 아이템 간의 유사성 혹은 단어 간의 유사성을 추출한다[9]. 동시 출현 행렬에 대해서는 특이값 분해(Singular Value Decomposition: SVD) 방식이 많이 사용되고 있다. 기존 연구에서 제안한 Neural Network 기반 embedding 방식이 널리 사용되고 있는 지금에도 행렬 인수분해 방법은 좋은 embedding 성능을 보여주고 있다[8, 18, 19, 22, 38].

특이값 분해와 같은 행렬 인수분해를 통해서 단어(word)나 항목(item), 용어(term)가 가진 잠재 의미 분석(Latent Semantic Analysis: LSA)하는 방법은 유사한 의미가 있는 단어나 항목, 용어는 확률적으로 비슷한 분포 형태로 나타난다는 가정에서 출발한다[7]. 하나의 단락 혹은 문장 안에서 단어가 여러 번 출현하거나 하나의 상품 구매 트랜잭션 안에서 여러 항목이 여러 번 출현하게 되며 이처럼 단어나 항목은 동시출현 행렬(co-occurrence matrix)의 형태로 단어는 행으로 문장 혹은 단락, 트랜잭션은 열이 되어 많은 수의 행과 열을 희소 행렬(sparse matrix)로 표현된다. 특이값 분해는 적은 차원 공간으로 단어 또는 항목을 표현하기 위한 행렬 인수분해 방법의 하나로 차원을 축소해 계산 효율성을 높이면서도 그 안의 잠재적 의미를 끌어내기 위해 사용된다[3].

2.3 Electronic Health Records

EHR 데이터는 환자의 질병 진단내역과 약물

처방내역, 치료 정보 등의 방대한 의료 정보가 디지털 데이터로 저장되어 있다. 과거에는 종이 차트에 적어 보관하였으나 이제는 전자의무기록의 도입이 증가하면서 국내외에서 이를 활용하기가 더욱 용이해졌다. EHR 데이터는 정확한 진료를 할 수 있도록 근거 자료로 사용하여 의사들의 의사결정을 지원하고 개인 맞춤 진단 및 예방 치료를 하는데 활용되고 있다[11, 13].

많은 연구에서 EHR 데이터를 활용하여 예측 모형을 개발하여 헬스케어의 품질을 높이려는 방법들을 제안하고 있다[12, 34]. 이와 같은 연구는 특히나 인플루엔자나 에볼라와 같은 빠르게 전염되고 확산하는 질병에 대해서 EHR 데이터는 의료 종사자들에게 즉각적인 인사이트를 제공하기에 용이하다. 국내에서는 2015년 메르스 사태 이후 감염병 확산에 대한 선제 조치 체계를 국가적 차원에서 마련하기 위해 건강보험심사평가원에서 미래창조과학부, 행정자치부의 지원을 받아 국내 EHR 데이터를 활용하여 '의료 빅데이터 분석을 통한 한국형 질병 감시 예측 시스템'을 구축하였다.

이전에는 많은 양의 EHR 데이터를 활용하여 모형을 만들기 위한 데이터 처리 과정이 중요하였고 도메인 전문가의 사전 지식이 예측 모형의 성능에 큰 영향을 미쳤다. 그리고 처리 과정 중에 너무나 많은 변수로 인해서 전체 데이터보다는 일부 중요하다고 판단되는 데이터만을 사용하는 경우도 많았다. 하지만 최근에는 embedding 방식을 통해 데이터를 저 차원 공간에서 표현하고 특징을 스스로 찾아가는 딥러닝 기법들이 예측 모형에 사용되기 시작하면서 전체 EHR 데이터를 예측 모형 개발에 활용할 수 있게 되었고 더 나은 결과를 보여주고 있다.

3. 연구 모형

3.1 Electronic Health Records Dataset

본 연구에서 사용한 데이터는 건강보험심사평가원(HIRA)에서 건강보험 청구 자료를 기초로 하여 진료개시일을 기준으로 1년간 진료받은 환자를 대상으로 표본 추출한 표본 데이터로 연구 목적으로 제공되는 자료이다. 본 연구는 2016년 건강보험심사평가원 입원환자표본자료(HIRA-NIS-2016-0078)를 활용한 것이며, 연구 결과는 건강보험심사평가원 및 보건복지부와 연관이 없음을 밝힌다. 입원환자표본자료는 국내 전체 입원환자를 대상으로 인구학적 구조인 성별, 연령 변수로 32개 구간을 설정하여 층화추출 방법에 따라 표본이 모집단의 대표성과 범용성을 갖도록 통계적으로 추출되었다[14]. 현재 전 국민의 98%가 국민건강보험에 가입된 청구자료는 전산청구율이 99%에 이르는 만큼 실제 보건의료 환경을 반영하는 데이터로 일반화가 쉽다.

건강보험 자료를 활용한 표본 추출 방식으로 청구자료를 추출하였고 환자 표본자료의 대상 모집단이 되는 건강보험 청구자료는 의료기관에서 건강보험심사평가원에 환자의 진료비용 중에서 국민건강보험이 지급해야 하는 부분을 청구하기 위한 자료이다. 청구자료 모집단에는 입원환자 1년간의 모든 진료, 처치, 처방내역을 포함하며 입원환자의 요양기관에 대한 현황 자료가 함께 포함된다. 의료기관에서는 월별 혹은 일자별로 환자 1명당 한 건의 건강보험 요양급여를 청구한다. 환자를 구분하는 코드를 기준으로 한 환자에게서 발생한 모든 청구 건을 묶는다면 입원한

자 한 단위로 추출할 수 있고 질병 특성이나 조작적 정의에 따라 연구에 활용할 수 있다[14].

입원환자표본자료는 명세서일반내역(general specification)과 진료내역(healthcare services), 질병내역(diagnosis information), 원외처방내역(outpatient prescriptions) 그리고 요양기관 현황정보(Providers) 총 5개의 테이블로 구성된다. 건강보험심사평가원에서 제공하는 입원환자표본자료는 청구를 목적으로 한 자료이기 때문에 실제 환자가 진단받은 질병과 일치하지 않을 수 있으며 비급여 진료와 성형 등의 자료는 포함되지 않는다는 제한점이 있다.

3.2 Embeddings

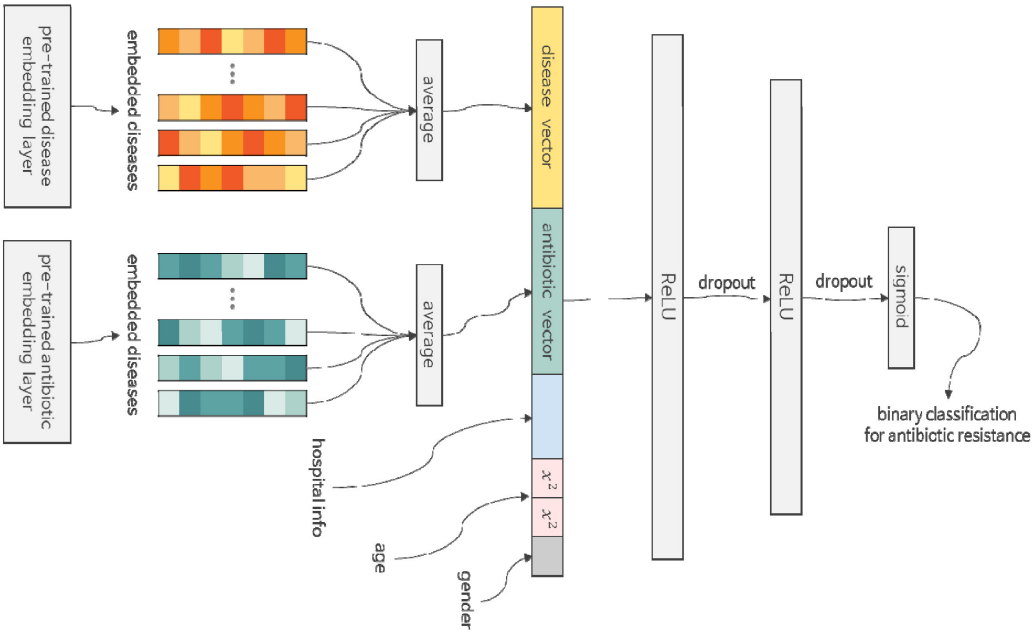
본 연구는 항생제 내성균의 감염을 분류 예측하기 위해서 환자가 진단받았던 질병의 내역과 처방을 받았던 항생제와 기타 약물에 대한 정보를 사용한다. 수천 개의 질병과 수만 개의 항생제를 데이터로 표현(representation)하여 예측 모형의 예측 인자(feature)로 사용하였다. 자연어처리 분야에서 낮은 차원의 실수형 벡터 공간 안에서 문장 속의 단어(word)들을 표현하기 위해 embedding을 한 것처럼[22] 환자의 질병 진단기록과 항생제 처방내역을 각각의 모형을 만들어 embedding space를 생성하였다. 질병과 항생제에 대한 embedding을 통해 낮은 차원으로 질병과 항생제를 표현하면서 그 안에 질병 간, 항생제와 약물 간에 존재하는 의미(semantic)를 내포하게 된다.

3.3 Model Architecture

본 연구는 입원환자에 대한 질병 진단내역과 항생제 및 약물 처방내역 그리고 환자정보 및 병원 이용 기록을 사용하여 항생제 내성균 감염 예측을 위한 딥러닝 기반 분류(binary classification) 예측 모형을 제안한다. 이때 질병 진단내역과 항생제 및 약물 처방내역은 희소 행렬의 형태이므로 사전에 학습한 embedding model을 통해 저차원 벡터로 변환하여 질병 d와 항생제 a를 다른 정보 h와 연결한(concatenate) 입원환자 P를 입력 데이터로 하였다.

이때 k명의 입원환자 i = 1, ⋯, k에 대하여 각 입원환자는 $p_i = (h_1,\ d_i,\ a_i)$이며, 입원환자별 질병 벡터 $d_i \in \mathbb{R}^s$은 s 차원으로, 항생제 벡터 $a_i \in \mathbb{R}^t$은 t차원으로embedding하여 표현하였다. 예측 모형은 입원환자 데이터 P의 입력 정보를 통해서 감염 여부를 분류하기 위해 sigmoid 분류자를 사용하였으며 cross-entropy를 사용하여 예측 분류 값과 실제 분류 값의 손실을 최소화하는 방향으로 모형을 학습시켰다.

환자 단위로 항생제 내성균 감염 여부를 예측하도록 하였으며 모형의 전반적인 아키텍처는 deep neural network를 사용하였다. 청구명세서별로 사전 학습된(pre-trained) embedding layer를 통해 질병 진단내역과 항생제 처방내역을 embedded vector로 변환하고 환자 단위로 평균(average)을 내어 질병 벡터와 항생제 벡터를 표현하였으며 두 벡터를 다른 정보 벡터와 연결(wide concatenate)하여 입력 데이터의 변수로 사용하였다. 연결된 벡터는deep neural network의 input layer에 입력되도록 하며 input layer는 고정된 길이의 dense vector와 fully-connected시키고 dropout을 통해 randomness를 추가하였으며 layer마다 활성화 함수(activation function)로 Rectified Linear Units (ReLU)를 사용하였다. Hidden layer의 깊이



〈Figure 2〉 Model Architecture

(depth)는 dense vector와dropout의 수를 바꿔가며 최적의 성능을 내는 것으로 조절하였다. 다음은 전체 모형에 대한 아키텍처이다.

4. 연구 방법

4.1 연구 대상

연구 대상은 전체 입원환자표본자료에 포함된 입원환자 987,763명은 2016년 전체 입원환자 중 약 13%의 비율로 통계적인 검증을 걸쳐 추출되었다. 항생제 내성균 감염 환자는 전체 입원환자 표본 중 질병 진료내역으로 항생제 내성균 감염과 관련 있는 7차 한국표준질병, 사인분류 기준에 따라 'U82(베타락탐항생제 내성)', 'U83(기타 항생제 내성)', 'U84(기타 항균제 내성)' 질병코드로 진단받은 기록이 있는 환자로 정의하여 추출하였다. 예측 모형 학습을 위해 목표가 되는 감염환자뿐만 아니라 항생제 내성균에 감염되지 않았던 대조군을 추출하여 학습 데이터와 검증 데이터를 구성하였다.

항생제 내성균에 감염되지 않았던 대조군은 감염환자 이외의 환자를 모두 감염 환자가 진단을 받았던 주진단 질병과 부진단 질병 중에서 상위 15개 다빈도 질병을 분석하여 해당 질병들로 주진단과 부진단을 받았지만 항생제 내성균에 감염되지 않았던 대조군 4,369명을 추출하였다. 그리고 감염환자 259명과 유사한 인구통계학적 정보와 병원이용 기록을 가졌으나 항생제 내성균에 감염되지 않은 환자 4,336명을 추가로 추출하여 총 8,705명을 대조군으로 설정하였다. 총 8,964명의 입원환자를 연구 대상으로 한다.

4.2 연구 설계

본 연구는 입원환자가 진단을 받았던 질병과 환자에게 투약이 된 항생제와 약물이 실제 항생제 내성균 감염에 큰 영향을 미쳤다고 가정하여 성별, 연령과 같은 환자에 대한 인구통계학적 정보와 환자의 병원이용 기록을 기준(baseline)으로 하고 질병 진단내역과 항생제와 약물 처방 내역을 추가됐을 때 예측 모형의 성능이 어떻게 변화했는지를 분석하였다.

또한, 항생제 내성균 감염 예측을 위해 새로 제안하는 딥러닝 모형과 기존에 주로 사용되었던 기계학습 모형의 분류 예측의 성능 차이를 비교하였고, 기준 모형은 Logistic regression, Decision tree, Bagging ensemble method, XGBoost, Super learner 총 5개의 알고리즘을 대상으로 하였다. 이 중 Super Learner[26, 33]는 의료 분야에서 예측 모형을 만들 때 많이 사용되었던 알고리즘으로 일부 연구에서는 딥러닝보다 나은 성능을 보였다는 결과가 있다[27, 37]. 분류 알고리즘의 유형 별로 그리고 앙상블 모형의 종류별로 대표적인 예측 모형을 가져와 본 연구에서 제안한 딥러닝 모형과의 비교를 진행하였다.

5. 분석 결과

5.1 자료 수집과 응답자 특성

항생제 내성균 감염 예측 모형을 학습하고 검증하기 위해 입원환자를 대상으로 하였다. 총 4,628명 중 감염환자 259명과 비감염환자 4,369명을 대상으로 검증하였다. 입원 환자의 특성은 <Table 1>과 같다.

예측 모형의 성능을 평가하기 위해서는 f1-score와 예측에 대한 적중률(Hit-ratio)의 결과는 <Table 2>와 같다. 모형의 학습과 검증을 위해서 훈련 데이터와 검증 데이터는 5개의 fold로 나누어 교차 검증을 하였으며, fold별로 생성된 5개의 예측 모형의 성능을 평가하였다. 그

〈Table 1〉 Comparison of Antibiotics-Resistant Bacterial Infected Patients and Noninfected Patients

		Infected patient	Uninfected patient
Total patients		259(0.02%)	4,369(0.44%)
Total Statement		11,507(0.03%)	17,496(0.05%)
Average statement by patient		44	4
Age	~9	9(3.9%)	349(8.0%)
	10~19	1(0.4%)	81(1.9%)
	20~29	5(1.9%)	162(3.7%)
	30~39	5(1.9%)	236(5.4%)
	40~49	11(4.2%)	354(8.1%)
	50~59	35(13.5%)	689(15.8%)
	60~69	38(14.7%)	885(20.2%)
	70~	155(59.8%)	1,613(36.9%)
Gender	Male	111(42.9%)	1,853(42.4%)
	Female	148(57.1%)	2,516(57.6%)

〈Table 1〉 Comparison of Antibiotics-Resistant Bacterial Infected Patients and Noninfected Patients((Continued))

		Infected patient	Uninfected patient
Death in a hospitalization.		47(18.1%)	1(0.0%)
Visiting hospital area	Seoul	138(53.3%)	649(14.8%)
	Busan	28(10.8%)	340(7.8%)
	Incheon	23(8.9%)	428(9.8%)
	Daegu	11(4.2%)	75(1.7%)
	Gwangju	4(1.5%)	149(3.4%)
	Daejeon	12(4.6%)	67(1.5%)
	Ulsan	8(3.1%)	81(1.9%)
	Jeju	3(1.2%)	15(0.3%)
	Sejong	1(0.4%)	1(0.0%)
Patients who are hospitalized to a general hospital more than once		251(96.9%)	39(0.9%)
Main diagnosis related surgical patient		198(76.4%)	24(0.5%)

결과 제안된 딥러닝 모형의 평균 f1-score가 0.617로 가장 높은 성능을 나타났고 선행연구에서 다른 기계학습 알고리즘에 비해 좋은 성능을 보였던 Super learner가 0.563으로 2순위의 결과로 나타났다. 최근 Kaggle과 같은 데이터 과학 경진대회에서 좋은 결과를 보여주고 있는 XGBoost가 0.557로 결과가 나타났다.

〈Table 2〉 Antibiotics-Resistant Bacteria Infection Prediction Results

Model	f1-score	Hit-ratio
Deep learning	**0.617**(±0.08)	0.831
Super learner	0.563(±0.05)	0.881
XGBoost	0.557(±0.05)	0.909
Logistic regression	0.512(±0.10)	0.809
Bagging ensemble	0.466(±0.11)	0.809
Decision tree	0.401(±0.06)	0.378

실제 항생제 내성균 감염과 연관이 있는 질병 진단내역과 항생제 처방내역을 예측 모형의 변수로 사용했을 때 예측 모형의 성능에 어떻게 영향을 주는지를 실험하였으며 입원환자의 기본정보와 병원이용 기록을 기준(baseline)으로 하여 질병과 항생제 정보의 추가에 따른 변화를 비교 분석하였다. 항생제 내성균 감염에 큰 영향을 미치는 병원 이용 정보가 추가되었을 때 f1-score가 0.103에서 0.525로 많이 증가하였으며, 질병과 항생제 정보가 추가되었을 때 0.617로 증가하면서 질병과 항생제의 정보가 항생제 내성균 감염 예측을 위한 주요 변수로 영향을 미치는 것을 파악하였다(Refer <Table 3>).

〈Table 3〉 Changes in Predictive Performance with the Addition of Diseases and Antibiotics Information

	f1-score
Patient Information	0.103
Patient Information + Hospital Usage Information	0525
Patient Information + Hospital Usage Information + Disease & Antibiotics	0.617

6. 결 론

6.1 연구결과 요약

본 연구는 Electronic Health Records 데이터를 사용하여 항생제 내성균 감염 예측을 위한 딥러닝 모형을 제안하였다. 이때 질병 진단내역과 항생제와 약물 처방내역을 embedded vector로 변환하여 예측 모형의 입력변수로 사용하여 감염을 예측할 때 인자로 활용되도록 하여 이에 따른 예측 성능의 변화를 분석하였다. 본 연구의 주요 결과를 요약하면 다음과 같다.

첫째, 입원환자의 과거 질병 진단내역과 항생제 및 약물 투약에 대한 내역을 예측 모형에 반영했을 때 더 나은 결과를 보였다. 선행연구에서는 유전자 서열 분석과 CT 이미지 데이터 분석을 통한 항생제 내성균 감염 예측을 연구하였으나 항생제 내성균에 감염되는 일차적인 원인이 진단된 질병과 그에 따른 항생제 및 약물 사용에 있으므로[20] 이에 대한 고려가 필요했고, 본 연구에서는 이러한 정보를 EHR 데이터에서 얻어 예측 모형에 반영할 수 있었으며 반영한 이후에 더 나은 결과를 보여주었다.

둘째, 딥러닝 모형이 다른 기존 기계학습 모형보다 더 나은 예측 성능을 보여주었다. 오래전부터 MIMIC 데이터를 이용하여 입원환자 사망률과 예상되는 입원 기간을 예측하는 연구들이 진행되었고 최근 딥러닝 기반의 예측 모형들이 기존의 방법들을 뛰어넘는 결과를 보여주고 있어[27] 항생제 내성균 감염 예측에서도 좋은 결과를 기대할 수 있었다.

셋째, Electronic Health Records 데이터를 사용하여 감염환자의 특성을 파악하였으며 국내 항생제 오남용 실태를 확인하였다. 연구에 사용한 데이터는 일부 의료기관의 편향된 데이터가 아닌 국내 전체 입원환자를 대상으로 통계적으로 추출된 대표성과 범용성이 있는 표본자료이다. 보건복지부(2014)에서 보고한 것과 같이 국내 항생제 내성률은 OECD 국가 중에서도 최상위권에 있으며 그 수는 계속해서 증가하고 있다. 본 연구는 정량적으로 이러한 사실을 확인할 수 있었다. 항생제 내성균에 감염된 환자들의 경우 항생제 사용하는 경우가 다른 환자들보다 많았고 J01 항생제 계열과 같이 신체에 많은 영향을 미치는 항생제 사용량도 월등히 많았다.

6.2 연구의 시사점

본 연구는 다음과 같은 시사점을 가지고 있다.

첫째, 본 연구는 다른 항생제 선행연구와는 다르게 내성균 감염 예측을 위해 질병 진단내역과 항생제 및 약물 처방 내역을 사용하였으며 이때 embedding을 기반으로 수많은 질병과 항생제를 적은 차원 공간 안에서 분산된 확률분포 형태로 매핑하여 차원을 축소하면서 동시에 그 안에 존재하는 잠재적 의미와 연관성을 포함할 수 있도록 하였다는 차별점이 있다. 질병을 embedding하기 위해서 자연어처리와 기타 기계학습 영역에서 그 효과를 입증한 neural embedding model을 구축하였으며, 항생제 및 약물 처방을 embedding하기 위해 추천시스템과 자연어처리 분야에서 아이템과 단어 안에 내포된 잠재적 요인을 찾기 위해 많이 사용되는 행렬 인수분해 방법을 사용하였다. 이러한 embedding 과정을 거쳐 축소된 질병 벡터와 항생제 벡터는 그 안에 잠재적 의미가 있는 것뿐만 아니라 연산 효율성을 높여 훨씬 적은 시간

으로도 학습할 수 있었다.

둘째, 항생제 내성균 감염환자의 특성을 분석하고 동일한 질병을 진단받았던 비감염환자와의 비교를 통해 국내 항생제 오남용 실태의 심각성을 파악하였다. 비감염환자와 비교하여 대부분의 항생제 내성균 감염 환자는 높은 빈도로 항생제를 과다 투약받은 것으로 나타났다. 따라서 의사들에게 항생제 처방과 투약에 있어 항생제 내성균 감염의 위험성을 제고하고 세계보건기구에서 권고한 지침(DDD)을 따를 수 있도록 국가적 차원에서 교육과 관리가 필요하다. 게다가 비감염환자에게서도 항생제가 일단 처방이 되면 권고 기준보다 높게 처방이 되는 경향이 있어 국내 항생제 오남용이 빈번하게 이뤄지고 있음을 알 수 있었다.

2016년 8월 보건복지부는 '국가 항생제 내성 관리대책(2016~2020)'을 발표하며 항생제 오남용을 줄이기 위한 여러 실천과제를 추진하였다. 보건복지부는 진료 과정에서 의사의 적정 항생제 처방을 지원하고자 항생제 처방 다빈도 질환에 대한 사용지침을 개발하여 배포하고 있다. 개인의 차원에서는 건강보험심사평가원 홈페이지에서는 항생제 처방률 적정성 평가 등을 해볼 수 있다. 이렇듯 2016년은 항생제 오남용을 줄이기 위한 국가적 차원의 정책을 펼친 첫해였으나 본 연구의 분석 결과를 통해서 관리가 잘 이루어지지 못했다는 것을 확인할 수 있었다.

2016년의 보건복지부의 항생제 오남용 보고서와 Park[25]의 연구에 따르면 사람의 왕래가 잦아 감염 위험이 컸던 종합병원 이외의 지역 의료시설에도 감염이 확산하는 추세라고 보고하였다. 종합병원에서는 메르스 사태 이후 감염 관리가 비교적 잘 이루어지고 있으나 다른 지역 의료시설에는 관리가 어려워 감염에 더 취약한 상황이므로 본 연구에서 제안한 예측 방법을 활용하여 실시간으로 지역 요양기관에서도 감염 예방을 위한 검사와 선제적 격리를 할 수 있도록 국가적 차원에서 정보 시스템 구축을 지원해야 할 것이다.

정보시스템을 통해서 축적된 데이터를 기반으로 새로운 항생제 내성 출현과 국제적 확산을 더욱 효과적으로 감지할 수 있을 것이다. 또한 기존의 '독성 및 내성 등으로 사용이 제한된 항생제'의 재생기술 개발 연구[6] 등을 지원할 수 있을 것으로 기대한다.

References

[1] A Medium Corporation, "I'll tell you why Deep Learning is so popular and in demand," https://medium.com/swlh/ill-tell-you-why-deep-learning-is-so-popular-and-in-demand-5aca72628780, 2019. 02. 22.

[2] Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L., "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data," Microbiome, Vol. 6, No. 1, p. 23, 2018.

[3] Bullinaria, J. A. and Levy, J. P., "Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD," Behavior Research Methods, Vol. 44, No. 3, pp. 890-907, 2012.

[4] Chen, M. L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., and Farhat,

M., "Deep Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome Sequencing Data," BioRxiv, 2018.

[5] Choi, J. W. and Lee, H. J., "An Integrated Perspective of User Evaluating Personalized Recommender Systems: Performance-Driven or User-Centric," The Journal of Society for e-Business Studies, Vol. 17, No. 3, pp. 85-103, 2012.

[6] Chung, J., Bhat, A., Kim, C. J., Yong, D., and Ryu, C. M., "Combination therapy with polymyxin B and netropsin against clinical isolates of multidrug-resistant Acinetobacter baumannii," Scientific Reports, Vol. 6, p. 28168, 2016.

[7] Dumais, S. T., "Latent semantic analysis," Annual Review of Information Science and Technology, Vol. 38, No. 1, pp. 188-230, 2004.

[8] Fonarev, A. Matrix Factorization Methods For Training Embeddings, 2018.

[9] Gamallo, P. and Bordag, S., "Is Singular Value Decomposition Useful for Word Similarity Extraction?," Lang. Resour. Eval., Vol. 45, No. 2, pp. 95-119, 2011.

[10] Gao, X. W. and Qian, Y., "Prediction of Multidrug-Resistant TB from CT Pulmonary Images Based on Deep Learning Techniques," Molecular Pharmaceutics, Vol. 15, No. 10, pp. 4326-4335, 2018.

[11] Ho, J. C., Ghosh, J., and Sun, J., "Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization," In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 115 - 124, New York, NY, USA: ACM, 2014.

[12] Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., and Sun, J., "Limestone: High-throughput candidate phenotype generation via tensor factorization," Journal of Biomedical Informatics, Vol. 52, pp. 199-211, 2014.

[13] Jensen, P. B., Jensen, L. J., and Brunak, S., "Mining electronic health records: towards better research applications and clinical care," Nature Reviews Genetics, Vol. 13, No. 6, p. 395, 2012.

[14] Kim, L., Sakong, J., Kim, Y., Kim, S., Kim, S., Tchoe, B., and Lee, T., "Developing the Inpatient Sample for the National Health Insurance Claims Data," Health Policy and Management, Vol. 23, No. 2, pp. 152-161, 2015.

[15] Kim, Y. H., Shin, G. W., and Lee, Y. H., "The Forecast of Future Technology Based on Deep Learning," Proceedings of KIIT Conference, pp. 219-220, 2015.

[16] Koren, Y., Bell, R., and Volinsky, C., "Matrix Factorization Techniques for Recommender Systems," Computer, Vol. 42, No. 8, pp. 30-37, 2009.

[17] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," In Proceedings of the 25th International Conferenceon Neural Information Processing

Systems, Vol. 1, pp. 1097-1105, USA: Curran Associates Inc, 2012.

[18] Levy, O. and Goldberg, Y., "Neural Word Embedding As Implicit Matrix Factorization," In Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, pp. 2177-2185, Cambridge, MA, USA: MIT Press, 2014.

[19] Liang, D., Altosaar, J., Charlin, L., and Blei, D. M., "Factorization Meets the Item Embedding: Regularizing Matrix Factorization with Item Co-occurrence," In Proceedings of the 10th ACM Conference on Recommender Systems, pp. 59-66, New York, NY, USA: ACM, 2016.

[20] Martinez, J. L., Baquero, F., and Andersson, D. I., "Predicting antibiotic resistance," Nature Reviews Microbiology, Vol. 5, No. 12, pp. 958-956, 2007.

[21] Mcadam, A. J., Hooper, D. C., Demaria, A., Limbago, M. B., O'brien, T. F., and Mccaughey, B., "Antibiotic Resistance: How Serious Is the Problem, and What Can Be Done?," Clinical Chemistry, Vol. 58, No. 8, pp. 1182-1186, 2012.

[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality," In Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3111-3119, USA: Curran Associates Inc, 2013.

[23] Moradigaravand, D., Martin, P., Anne, F., Ville, M., Warringer, J., and Parts, L., "Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data," PLOS Computational Biology, Vol. 14, No. 12, pp. 1-17, 2018.

[24] Omlin, C. W. and Giles, C. L., "Stable Encoding of Large Finite-State Automata in Recurrent Neural Networks with Sigmoid Discriminants," Neural Computation, Vol. 8, No. 4, pp. 675-696, 1996.

[25] Park, S. H., "Management of multi-drug resistant organisms in healthcare settings," Vol. 61, No. 1, pp. 26-35, 2018.

[26] Polley, E. C. and van der Laan, M. J., "Super Learner in Prediction," U.C. Berkeley Division of Biostatistics Working Paper, 1-19, 2010.

[27] Purushotham, S., Meng, C., Che, Z., and Liu, Y., "Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets," CoRR, 2017.

[28] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., and Dean, J., "Scalable and accurate deep learning with electronic health records," Npj Digital Medicine, Vol. 1, No. 1, p. 18, 2018.

[29] Roh, J. H., Kim, H. J., and Chang, J. Y., "Improving Hypertext Classification Systems through WordNet-based Feature Abstraction," The Jounal of Society for e-Business Studies, Vol. 18, No. 2, pp. 95-110, 2013.

[30] Santos, R. P., Mayo, T. W., and Siegel,

J. D., "Active Surveillance Cultures and Contact Precautions for Control of Multidrug-Resistant Organisms-Ethical Considerations," Clinical Infectious Disease, Vol. 47, No. 1, pp. 110-116, 2008.

[31] Socher, R., Lin, C. C. Y., Ng, A. Y., and Manning, C. D., "Parsing Natural Scenes and Natural Language with Recursive Neural Networks," In Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 129-136, USA: Omnipress, 2011.

[32] Song, J. H., "Current status and future strategies of antimicrobial resistance in Korea," The Korean Journal of Medicine, Vol. 77, No. 2, pp. 143-151, 2009.

[33] Van der Laan, M. J., Polley, E. C., and Hubbard, A. E., "Super Learner," Statistical Applications in Genetics and Molecular Biology, Vol. 6, No. 1, 2007.

[34] Xiang, T., Ray, D., Lohrenz, T., Dayan, P., and Montague, P. R., "Computational Phenotyping of Two-Person Interactions Reveals Differential Neural Response to Depth-of-Thought," PLOS Computational Biology, Vol. 8, No. 12, e1002841, 2012.

[35] Xu, X., Liang, T., Zhu, J., Zheng, D., and Sun, T., "Review of classical dimensionality reduction and sample selection methods for large-scale data processing," Neurocomputing, Vol. 328, pp. 5-15, 2019.

[36] Yang, Y., Niehaus, K. E., and Clifton, D. A., "Predicting antibiotic resistance from genomic data," In Machine learning for healthcare technologies, pp. 203-226, IET, 2016.

[37] Young, S., Abdou, T., and Bener, A., "Deep super learner: A deep ensemble for classification problems," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10832 LNAI, pp. 84-95, 2018.

[38] Zhang, M., Hu, B., Shi, C., Wu, B., and Wang, B. (n.d.)., "Matrix Factorization meets Social Network Embedding for Rating Prediction," In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, pp. 121-129, Springer, Cham, 2018.

저 자 소 개

오성우 (E-mail: supoman@yonsei.ac.kr)
2016년 중앙대학교 심리학 (학사)
2017년~현재 연세대학교 정보대학원 비즈니스 빅데이터분석 (석사)
관심분야 Machine Learning, Deep Learning

이한길 (E-mail: hankil.lee@yonsei.ac.kr)
2008년 이화여자대학교 제약학과 (학사)
2011년 이화여자대학교 생명약학부 (약학석사)
2017년 연세대학교 약학과 사회약학 (약학박사)
2017년 연세대학교 약학대학 박사후연구원
2018년~현재 연세대학교 약학대학 연구교수
관심분야 Drug Utilization Research, Pharmacoeconomics, Outcomes Research

신지연 (E-mail: yeond289@yonsei.ac.kr)
2018년 이화여자대학교 의류학/경영학 (학사)
2018년~현재 연세대학교 정보대학원 비즈니스 빅데이터분석 (석사과정)
관심분야 Machine Learning, IT Performance

이정훈 (E-mail: jhoonlee@yonsei.ac.kr)
1995년 University of Manchester(U.M.I.S.T) 전자공학 (학사)
1996년 University of Manchester(U.M.I.S.T) 시스템공학 (석사)
1998년 London School of Economics 경영정보학 (석사)
2003년 University of Cambridge 생산/정보시스템공학 및 경영 (박사)
2004년~현재 연세대학교 정보대학원 교수
관심분야 IT Governance, Performance Measurement in IT, Systems Dynamics, Multi agent system modeling and simulation, Information System Intelligence