Journal of the Korea Industrial Information Systems Research Vol.25 No.6, Dec. 2020 :33-45
http://dx.doi.org/10.9723/jksiis.2020.25.6.033 ISSN:1229-3741

사용자 참여형 웨어러블 디바이스 데이터 전송 연계 및 딥러닝 대사증후군 예측 모델+

(Deep Learning Algorithm and Prediction Model Associated with Data Transmission of User-Participating Wearable Devices)

이 현 식[1], 이 웅 재[2], 정 태 경[3]*
(Hyunsik Lee, Woongjae Lee, and Taikyeong Jeong)

요 약 본 논문은 최근 다양한 종류의 웨어러블 디바이스가 헬스케어 도메인에 급증하여 사용되고 있는 상황에서 최신 첨단 기술이 실제 메디컬 환경에서 개인의 질병예측이라는 관점을 바라본다. 사용자 참여형 웨어러블 디바이스를 통하여 임상 데이터와 유전자 데이터, 라이프 로그 데이터를 병합하여 데이터를 수집, 처리, 전송하는 과정을 걸쳐 딥뉴럴 네트워크의 환경에서 학습모델의 제시와 피드백 모델을 연결하는 과정을 제시한다. 이러한 첨단 의료 현장에서 일어나는 메디컬 IT의 임상시험 절차를 걸친 실제 현장의 경우 대사 증후군에 의한 특정 유전자가 질병에 미치는 영향을 측정과 더불어 임상 정보와 라이프 로그 데이터를 병합하여 서로 각기 다른 이종 데이터를 처리하면서 질병의 특이점을 확인하게 된다. 즉, 이종 데이터의 딥뉴럴 네트워크의 객관적 적합성과 확실성을 증빙하게 되고 이를 통한 실제 딥러닝 환경에서의 노이즈에 따른 성능 평가를 실시한다. 이를 통해 자동 인코더의 경우의 1,000 EPOCH당 변화하는 정확도와 예측치가 변수의 증가 값에 수차례 선형적으로 변화하는 현상을 증명하였다.

핵심주제어: 딥러닝, 웨어러블 디바이스, 디지털 헬스케어, 질병 예측, 유전자, 라이프 로그

Abstract This paper aims to look at the perspective that the latest cutting-edge technologies are predicting individual diseases in the actual medical environment in a situation where various types of wearable devices are rapidly increasing and used in the healthcare domain. Through the process of collecting, processing, and transmitting data by merging clinical data, genetic data, and life log data through a user-participating wearable device, it presents the process of connecting the learning model and the feedback model in the environment of the Deep Neural Network. In the case of the actual field that has undergone clinical trial procedures of medical IT occurring in such a high-tech medical field, the effect of a specific gene caused by metabolic syndrome on the disease is measured, and clinical information and life log data are merged to process different heterogeneous data. That is, it proves the objective suitability and certainty of the deep neural network of heterogeneous data, and through this, the performance evaluation according to the noise in the actual deep learning environment is performed. In the case of the automatic encoder, we proved that the accuracy and predicted value varying per 1,000 EPOCH are linearly changed several times with the increasing value of the variable.

Keywords: Deep learning, Wearable device, Digital healthcare, Disease prediction, Genome, Life-log

* Corresponding Author: ttjeong@sehan.ac.kr
+ This work was supported by the Technology Innovation Program no. 20002781, funded by the Ministry of Trade, Industry & Energy(MOTIE). This work was supported by Sehan Univ. grant in 2020.

Manuscript received November 06, 2020 / revised December 07, 2020 / accepted December 14, 2020
1) CHA Univ. Dept. of Integrated Medicine, 제1저자
2) Seoul Women's Univ. Dept. of Digital Media, 제2저자
3) Sehan Univ. Dept. of Artificial Intelligence, 교신저자

1. 서 론

현대사회의 인류는 질병에 대한 두려움과 염려로 말미암아 과학기술과 메디컬 IT의 영역의 발전과 새로운 신약 물질 또는 신약 후보물질을 발굴하여 질병 퇴치 및 감염 예방의 결과로 성과를 이루고 있다 (Mark and Pichika. 2019; Koul et al., 2011). 최근 불어오는 인류 감염병 확산과 문제로 인하여 포스트 코로나-19 (COVID-19)를 대비하려는 연구가 많은 사람들의 공감을 불러일으키고 있다. 이런 가운데 질병의 위협으로부터 자유롭게 하는데 별반 이견을 달리하지 않고 있는 실정이다(Klok et al., 2020).

특히 최신 첨단 기술을 실제 메디컬 환경에서 적용하고 이를 바탕으로 임상시험에 들어가고자 하는 많은 메디컬 영역에서의 각고의 노력이 과학기술의 발전을 이루게 되어 헬스케어기기와 관련 소프트웨어가 많은 발전을 이루고 있는 다수의 예가 있다. 인공지능 또는 딥뉴럴네트워크 (이하 딥러닝 또는 Deep Neural Network)를 인류 사회의 최신 헬스케어 도메인에 적용하여 메디컬 IT 영역을 도와주게 되는 중요한 목적을 가지고 바라볼 때 인공지능을 통하여 누구도 생각하지 못할 과학적 발견과 발명을 이룰 수 있게 된다 (Hamet and Tremblay, 2017).

이에 본 논문에서는 다수의 많은 사용자가 참여하고 웨어러블 디바이스를 통하여 측정된 개인의 헬스케어 데이터 (Personal healthcare record; PHR)를 딥러닝 모델을 통한 여러 개의 피드백을 주고받으며 처리하는데 주안점을 맞추어 연구하고자 한다 (Du et al., 2018). 또한, 각종 데이터를 처리함에 있어서 각기 다른 도메인에서 존재하게 되며 성격이 전혀 이종 데이터 (Heterogenous data)를 수집하고 처리하는데 수반되는 데 따른 기술적 가치를 형성하고 이를 통해 공공 데이터 기반의 건강예측 모델과 정밀의료의 한 축으로서 디지털 헬스케어의 중요성을 두각 시킬 수 있도록 하는데 목적을 가지고 있다 (Freudenberg and Propping. 2002).

특히 첨단 의료의 현장에서 일어나는 많은 의료기기와 임상 데이터들이 혼재되어 있는 상황에서 메디컬 IT의 특성상 불특정 다수에게 제약 없이 참여하게 하지 못하는 제한사항을 수반하기도 한다. 이러한 특성으로 인해 임상데이터와 개인의 헬스케어 데이터를 자율적으로 또는 제한적으로 참여를 유도하여 최신 메디컬 IT 기술을 적용하기는 많은 제약사항이 따르게 된다. 이를 해결하기 위하여 전향적 코호트 (Cohort) 연구를 통하여 일반인은 물론 환자의 임상 데이터 (Clinical data), 유전자 데이터 (Gene data), 개인 헬스케어 데이터 (PHR), 라이프로그 데이터 (Life-log data) 를 동시에 수집하였다. 이를 통해 건강관리 시스템에 적용하기 위하여 임상시험계획서 (IRB)와 같은 절차를 모두 승인하에 진행하여야 하는 것이다 (Lee et al., 2019).

이러한 웨어러블 디바이스와 같이 실제 일반인과 환자들의 여러 가지 데이터를 혼합하여 가장 최적화된 질병 예측모델을 구축해 내기 위해서는 데이터 모델링에서부터 수집, 처리, 모니터링의 단계를 거쳐 최종적으로 질병 예측 모델의 성능을 확인할 수 있어야 한다. 동시에 각종 공공데이터를 통한 정밀의료 (Pittman et al., 2004)의 한 축을 기본 데이터로 사용하여 최종적으로 딥러닝에 의한 건강예측 모델의 우수성을 증명할 수 있어야 한다.

이를 위한 여러 가지 딥러닝 모델과 예측 방법을 모두 살펴볼 필요가 있다. 모든 딥러닝 모델을 살펴보기에는 방대한 유전자 데이터와 라이프 로그 데이터의 변화하는 패턴을 모두 학습해야 하는데 이는 시간적, 공간적 제약으로 현실적으로 불가능하다. 이에 임상시험계획서 (IRB)에 명시되어 참여한 사용자를 대상으로 최종적인 모델을 도출하기 위하여 각종 데이터 수집과 선정 작업을 실시하였다.

Fig. 1에서 보이는 바와 같이 사용자들의 웨어러블 디바이스를 통한 각종 데이터를 모두 수집하고 이를 통한 이종 데이터들 간의 Training 데이터 (Training data)와 테스트 데이터 (Test data)로 구분된 모델과 1차 피드백 모델, 2차 피드백 모델로 구분하여 최종적으로 딥러닝 질병예측모델을 도출할 수 있도록 하였다. 특히 이 중에서 특정 질병 군을 측정하기 위하여 혈압, 공복혈당, 중성지방, HDL-콜레스테롤 (High-density Lipoprotein), 허리둘레, BMI (Body mass index)등과 같은 대사증후군의 데이터를 최우선적으로 적용하기로 하였다.

임상시험계획서 (IRB)에서 승인된 치매유전자에 따른 라이프 로그 데이터를 확보하는 작업 (Dolpj et al., 2017)과 유방암 질병을 경험한 환자군

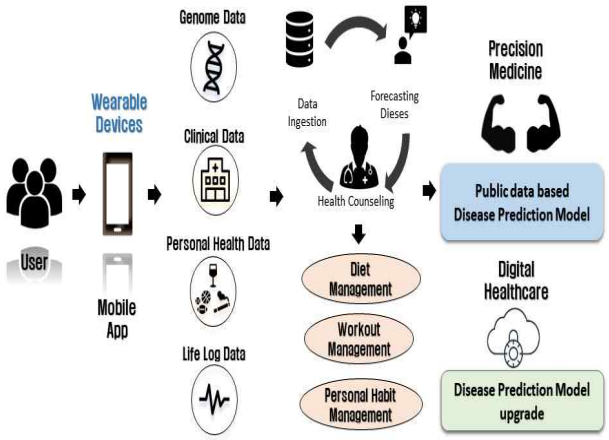

Fig. 1 Overall System Design Diagram of Various Public Data and Health Prediction Models through Wearable Devices

의 임상 데이터를 활용하여 질병을 예측 (Cireşan et al., 2013)하고 환자의 예후를 예측할 수 있는 딥 뉴럴 네트워크의 단계까지 학습할 수 있도록 딥러닝 모델로 발전할 수 있도록 하는데 본 연구의 목적을 두고 있다. 즉, 웨어러블 착용 군과의 비교 대조군을 형성하는 기초 실험 환경의 조성에서부터 최종적인 딥러닝 질병예측 모델의 상관관계 분석 및 성능을 비교하는 단계까지 임상데이터의 수집에서부터 라이프 로그 데이터의 병합과 처리라는 중요한 과학적 단계를 거치게 된다. 각종 임상 데이터는 물론 이종 데이터의 처리방법론 (Jeong, 2020)을 새롭게 도출하였으며, 딥러닝에 의한 질병예측 모델의 원천기술의 2번째 참고 문헌 (Jeong, 2020)을 보유하게 되었다.

본 논문은 최신 첨단 기술 중의 하나인 웨어러블 디바이스를 확대 보급하는 일을 통해 일반인 또는 환자들의 임상 경험 데이터를 수집하고 처리하는데 그치지 않고, 각기 다른 라이프 로그 데이터와 유전자 데이터를 병합하여 대상 데이터를 새롭게 수집, 처리하는 과정에 중심을 둔다. 이로서 건강 예측 모델로 확산하기 위한 기초 과학적 메디컬 IT의 개념과 임상적 의미를 확보하기 위하여 노력을 기울인다. 사용자들의 참여로 라이프 로그 데이터를 수집하고 이를 기반으로 서비스 플랫폼의 다양화가 이루어지며 개인 맞춤형 헬스케어로 발전하게 되는 메디컬 IT의 영역에서 최신 헬스케어 디바이스의 개발 목적과 딥러닝에 의한 질병예측 모델의 개발이라는 대명제를 완성할 수 있도록 하고자 한다.

제 2장에서는 웨어러블 디바이스를 통한 관련 연구와 국제적인 흐름과 동향을 살펴보고, 제 3장에서는 디지털 헬스케어 전반의 웨어러블 디바이스의 데이터 전송과 처리과정을 설명하고자 한다. 제 4장에서는 딥러닝 질병예측 모델에 대하여 설명하고 제 5장에서는 각종 데이터를 처리하여 성능을 평가하는 실험 환경과 절차, 그리고 결과를 논하기로 한다.

2. 관련연구

사용자 참여형 웨어러블 디바이스에 의한 데이터 전송 및 딥러닝 예측 모델의 중요한 몇 가지 연구를 나열할 수 있다. 예를 들면 모바일 스마트폰으로 연동하는 웨어러블 디바이스는 국제적인 글로벌 기업들이 이미 많은 버전의 센싱 하드웨어, 펌웨어 버전, 최신 알고리즘을 앞 다투어 시장에 내놓고 있다. 특히 종전에 없었던 새로운 심전도 모니터링 (Electrocardiogram monitoring), 혈중 산소 포화도 (Saturation Pulse Oxygen; SPO2) 등을 기존의 혈압 측정 등의 기능이 새로이 추가되었다 (CNP, 2020).

이 중에서 심전도 측정의 중요성과 함께 산소포화도를 사용자에게 알려주는 기능은 글로벌 기업이 최신으로 앞 다투어서 개발하는 기능으로 헬스케어의 중요성과 함께 메디컬 IT의 도메인에서 접근할 수 있는 첨단기술의 영역으로 뽑을 수 있다. 사용자들의 일상적인 생활 패턴을 측정하여 수집하는 데이터는 여러 방면에서 인류의 건강에 도움을 주기 때문에 이를 딥러닝과 연계하려는 많은 연구들이 있다. 특히 수면을 취하는 동안의 혈중 산소 농도를 웨어러블 디바이스로 측정하는 연구의 경우와 같이 중요한 임상적 소견을 보이기도 한다 (Sathyanarayana et al., 2016). 이는 저산소증이 미리 알람을 줄 수 있는 경우 (산소 포화도 ≤95%)에 호흡기 질환이나 폐렴과 같은 소견을 보일 수 도 있기 때문이다. 또한 일상적 웨어러블 디바이스를 통한 심전도 모니터링의 경우에 심장전문의 수준의 부정맥을 측정할 수 있는

정도의 정확성을 보여주는 높은 수준의 연구가 진행되었다 (Hannun et al., 2019).

Fig. 2에서는 사용자들이 참여하여 각종 임상 데이터, 유전자 데이터, 라이프 로그 데이터를 수집하여 전송하게 되는 과정을 모식도로 보여주고 있다. 이때 측정되고 모니터링 되는 라이프 로그 데이터와 임상 데이터 중에는 유전자 데이터가 포함되어 있는데 이를 서로 다른 도메인 데이터이므로 데이터 처리과정에서 보이는 정확성을 높이기 위한 방법이 필요하게 된다. 유전자 데이터의 경우에는 딥뉴럴 네트워크의 초기에만 상관되지만 전체적인 라이프 사이클에서는 반드시 필요한 중요 특징점 (Features)로 신경망에서 다루어지기 때문에 명확한 출력 값의 하나로 명시적으로 포함하여야만 한다 (Grapov et al., 2018).

동시에 웨어러블 디바이스는 기업들이 첨단 최신 기종을 보여주는 현장이기도 하다. 예를 들면 최신 기종 중에서 스마트워치용 피부전기활동 (EDA) 센서, 다중 입력에 의한 센싱 정보의 배분, 유기발광 다이오드, 통신수단 및 기술 등에서 최신 연구 동향을 볼 수 있다. 이를 구체적으로 명시하면 기업중심의 경제 판도에서 웨어러블 디바이스의 판매상황이 높아지고 있는 것을 볼 수 있으며 (Walker, 2014), 멀티채널에

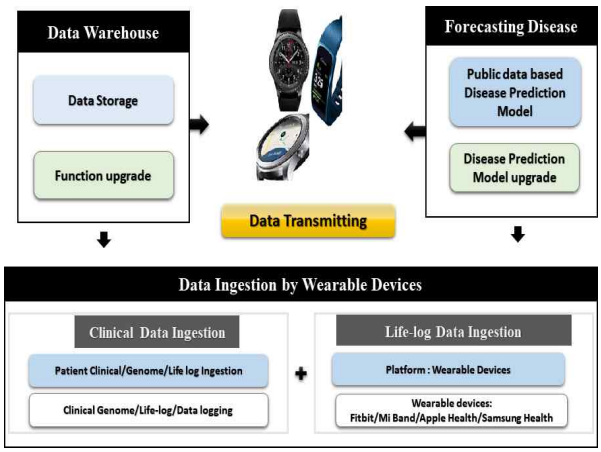

Fig. 2 Schematic Diagram of the Data Transmission Process according to the Transmission and Collection Procedures of Various Clinical, Gene, and Life-logging Data

의한 시계열 데이터에 의한 결과 값을 표현하는 방식으로서 혈압과 심박동을 동시에 처리하여 결과 값으로 보여주는 기술적인 표현과 더불어 헬스케어 데이터의 규제 해제와 더불어 많은 진전을 보이고 있다 (Jeong, 2020).

3. 디지털 헬스케어

3.1 웨어러블 디바이스와 데이터 전송

사용자들의 데이터를 수집하는 방법은 데이터의 유형과 형태에 따라 분석 기술이 다르게 적용된다. 수집 방법은 웨어러블 디바이스로 임상 데이터와 함께 크롤링, ETL, 로그수집, 센싱 정보로 가능하다, 일반인 또는 환자들의 24시간동안 모든 개인 헬스케어 데이터를 DBMS를 통해 SQL 등을 활용하여 정형 데이터를 가져 오고 수집 및 저장을 진행한다. 임상 시험 계획서 (IRB)에 명시된 범위 안에서 주기적인 측정을 통하여 임상의에게 주문된 임상 데이터를 모두 스크립트 형태로 제공하는 반정형 데이터와 함께 데이터 형태로 기록되어 비교할 수 있게 된다. 기 구축된 임상데이터 보다는 전향적 코호트 연구를 실시하도록 하여 정상 군과 비교 군의 데이터를 모두 수집하기 위하여 오픈 API (Application protocol interface)를 기반으로 한 데이터 수집을 실시하였다.

일반적으로 웨어러블 디바이스 착용 군을 통하여 설문조사와 유전자 조사를 통하여 개인의 헬스케어 데이터를 수집하고 이를 유전적 위험도 (Model 1) 와 생활습관 위험도 예측 모델 (Model 2)을 구분하여 개발하기로 하였다. 이때 발생하는 임상시험 기준 등은 모델별로 상이하여 제 4장에서 깊게 다루기로 한다.

Fig. 3에서 보는 바와 같이 유전적 위험도 예측 모델 (Model 1)은 공공데이터 수집의 단계에서부터 유전자 검사에 이루기까지 모든 단계에서 활용되는 데이터를 저장하여 특정 값을 추출하게 된다. 마찬가지로 생활습관 위험도 예측 모델 (Model 2)은 웨어러블 디바이스를 기반으로 하는 라이프 로그 데이터와 임상 데이터의 상관관계를

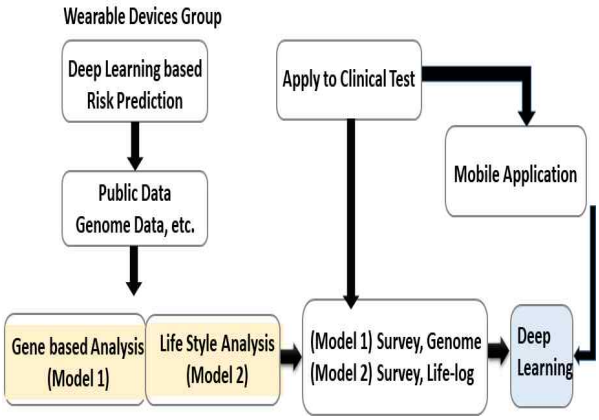

Fig. 3 Procedure and Flow Chart of Mobile Implementation Clinical Trial Application by Merging Predictive Model 1 and Predictive Model 2 for Deep Learning-based Disease Risk Prediction

모델링하여 설정하여 순서도를 제시한다.

데이터 분석 및 질병예측 모델의 수립 측면에서 라이프 로그 데이터 수집은 분석의 내용과 질병예측 모델의 성능을 결정하고 이후 응용에 있어서도 많은 영향을 미치기 때문에 어떻게 각기 다른 이종 데이터를 분석하고 예측모델에 사용할 것인지 결정된다면 수집할 개인 헬스케어 데이터에 대한 사전 탐색을 정하고, 이때 데이터 수집의 난이도와 비용, 안정성을 고려해야 한다. 본 논문에서는 데이터의 분석, 1차 모델링, 2차 모델링, 그리고 모델을 위한 성능 평가의 단계로 진행하였다.

3.2 데이터 분석과 딥러닝

임상 데이터를 포함하는 웨어러블 디바이스의 라이프 로그 데이터를 선정할 때 우선적으로 고려되어야 하는 사항은 수집 가능성이며 아무리 좋은 데이터가 있다 하더라도 수집이 어렵다면 연구 및 진행에 위험요인이 된다. 기초 임상정보와 같은 혈당, 콜레스트롤, 혈압 등의 원천 데이터에서 일반인 또는 환자들의 데이터를 각기 다른 방식으로 데이터 웨어하우스에 저장하고 이를 다시 라이프 로그 데이터와 함께 운동, 심박, 수면, 식이기록으로 분류하여 전처리과정을 걸쳐 Jason 형식의 메타정보를 뽑아내게 된다.

앞서 언급한 웨어러블 디바이스에서 수집되는 메타정보와 기본 데이터를 중심으로 모두 공통 형식으로 할 필요는 없으나 데이터 보간법을 적용하여 새롭게 이종 데이터간의 일어날 수 있는 데이터 엉킴과 충돌을 사전에 제거하여야 한다 (Jeong, 2020). 이를 통하여 구축되는 이종 데이터간의 연결모델은 제 4장과 제 5장에서 자세히 다루기로 한다.

실험을 위한 웨어러블 디바이스는 Fitbit를 사용하여 실험 군을 최초 46명 모집하고 비교군 46명을 대조하였다. 이후 증가된 실험 군을 위하여 계속적인 확보를 이루고 있다.

Fig. 4는 사용자들에 착용된 웨어러블 디바이스의 수집된 데이터들이 모바일 서버에서 응용단계로 넘어가 하이브리드 서버에 저장된 후 2차적으로 운동, 심박, 수면, 식이기록과 같은 데이터와 병합되고 최종적으로 라이프 로그 데이터와 같은 형태로 라이프 로그 전용 서버에 전송되어 처리되는 과정을 보인다. 이때 임상 데

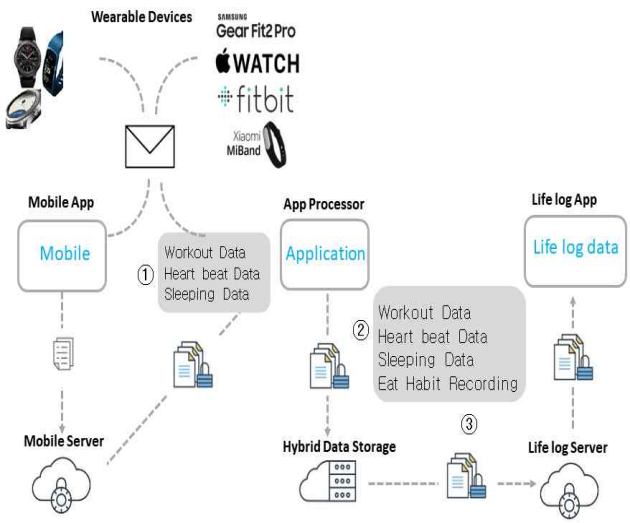

Fig. 4 A General Wearable Device that Transmits Basic Clinical Information (Blood Sugar, Cholesterol, Blood Pressure, etc.) and Life-log Data, and Deep Learning-based Data Collection Method and Transmission Schematic Diagram

이터, 유전자 데이터 및 라이프 로그 데이터를 처리하는 과정을 순차적으로 설명하면 아래와 같이 시간에 의하여 열거할 수 있다.

(1) 상용 웨어러블 디바이스로 운동, 수면, 심박 데이터 수집
(2) 수집된 데이터를 라이프 로그 모바일 앱으로 가져옴
(3) 라이프 로그 모바일 앱으로 식사기록 수집. 수집하고 전송된 데이터를 모바일 클라우드 서버에 전송
(4) 모바일 클라우드 서버에서 하이브리드 서버에서 최종 전송
(5) 전송된 데이터는 유전자 데이터와 통합, 분석되어 의료기관에서 환자 관리 및 연구목적으로 사용

또한, 임상 데이터 및 라이프 로그 데이터의 크기에 따라서 모바일 및 하이브리드 서버의 데이터와 모델이 달라지고 예상하지 못했던 문제가 발생할 수 있어 통일된 방식으로 데이터를 수집하게 할 필요가 있었다. 이 문제는 이종 데이터의 사이에서 발생하는 예측하지 못한 확장성에 관여하여 다음 데이터로 강제 입력이 필요한 경우와 강제 삭제가 필요한 경우로 나뉘기도 한다 (Jeong, 2020). 마찬가지로 무시하는 경우와 같이 이종 데이터의 처리는 중요한 변수로 질병예측 모델의 구현에 영향을 미치게 된다. 그러므로 이종 데이터와 처리에 관련하여서는 기술적으로 제어할 수 있는 문제와 제어할 수 없는 문제가 발생하기 때문에 최대한 통제할 수 있는 방향으로 데이터를 수집하기 위해서 수집 절차를 최대한 간소화하고 구분되어 작업을 해서 오류를 줄이는 방향으로 개선하였다.

4. 딥러닝 질병예측 모델

개인이 지니고 있는 임상 정보의 중요성에 비추어 볼 때 실시간으로 수집되는 모든 헬스케어 데이터와 본연의 유전자 데이터가 병합되어 최종적인 질병예측이 가능하기 위하여서는 많은 요소들을 고려해야 한다. 앞서 언급했던 것처럼 최종적인 유전적 위험도 (Model 1)과 생활습관 위험도 (Model 2)의 각기 다른 데이터 즉, 이종 데이터간의 병합을 이루기 위해서는 학습 모델별로 조건이 달라야 하기 때문이다. 이때 학습 모델이란 딥뉴럴 네트워크에서 분석하고자 하는 대상으로서 조건 값에 기인한 결과 즉, 예측치를 발생하는 원리를 나타내는 것으로 의사결정 트리 구조의 회귀분석의 특성을 모두 가질 수 있다 (Mythili et al., 2013).

딥뉴럴 네트워크에서 사용자 참여에 의하여 수집되는 이종 데이터의 병합으로 인해 학습 모델이 수시로 바뀌는 문제가 발생하기 때문에 이를 인공지능의 학습의 대상으로 삼아 딥러닝의 출력 값으로 산출하고자 한다. 우선 이종 데이터들의 전처리 과정을 수반하여 여러 데이터 보간법을 적용하여 제시하여야 한다 (Zhang and Wu, 2006).

Fig. 5와 같이 최초에 사용자들의 참여로 수집되는 데이터는 임상 데이터로부터 각 개인들의 유전자 검사에 의한 데이터를 포함한다. 이때 Training 데이터 (Training data)와 테스트 데이터 (Test data)를 분할하기 위한 공공데이터를 포함하여 딥러닝 모델에 동시에 적용하여 첫 번째 예측치 (1^{st} Prediction)를 발생하게 된다. 이를 1^{st} Prediction 값으로서 첫 번째 모델 즉, 모델의 첫 번째 피드백 모델 (1^{st} Feedback

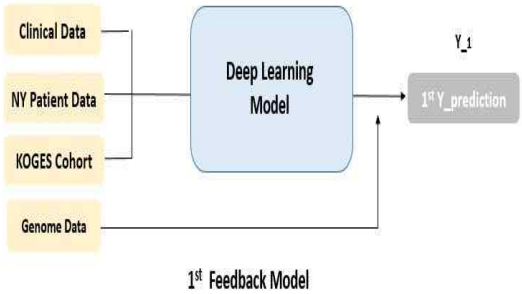

Fig. 5 The First Feedback Model Procedure in the Process of Generating the First Prediction Value of the Deep Learning Model by Inputting Various Data

Model)이 형성되게 된다.

이를 첫 파일럿 테스트를 시행하기 위하여 상당히 많은 공공 데이터와 함께 각 개인의 헬스케어 정보를 무작위로 선택하여 데이터 보간법을 거치게 된다. 임상 시험 계획 (IRB)에서 승인된 각 개인들의 데이터를 딥러닝으로 실행하였을 때 나오는 결과치의 신빙성을 높여주는 방식으로 향후 예측치의 정확도를 사용자 분석의 관점에서 접근하기 위한 것으로 증명된다.

Fig. 6에서는 첫 번째 피드백 모델에서 발생한 1st Prediction 값이 Y_1 으로 동일시되고, 라이프 로그 데이터가 입력 값으로 동작한다. 이렇게 현저하게 다른 이종 데이터의 경우 병합되는 과정에서 문제를 발생할 수 있게 되어 학습 모델의 신빙성이 떨어지게 되는 경우를 방지하기 위하여 조건 값을 적용할 수 있다.

결국, 1차 예측치를 생성에 의한 라이프 로그 데이터의 병합과정과 최종적인 2차 예측치 (2nd Prediction 값)의 생성과정에서 두 번째 피드백 모델 (2nd Feedback Model)이 나오게 된다.

이렇듯 1차 피드백 모델과 2차 피드백 모델을 거치는 동안 딥뉴럴 네트워크의 심층 신경망이 강화되고 각기 다른 이종 데이터의 예측치의 의사결정트리가 강화되면서 최종적인 결과물로서

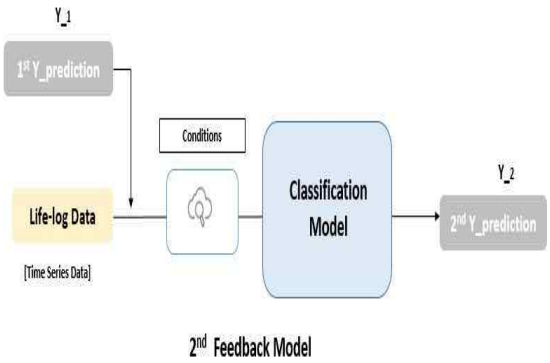

Fig. 6 The Procedure Diagram of the Second Feedback Model in the Process of Merging Life Log Data by Generating the First Prediction Value and the Process of Generating the Second Prediction Value

각 개인들의 유전 정보를 바탕으로 라이프 로그의 상관관계를 볼 수 있고 동시에 임상 정보와의 관계에 대한 객관적인 증명이 가능하게 된다.

본 논문에서는 대사증후군 중에서 혈압, 공복혈당, 중성지방, HDL-콜레스테롤 (High-density lipoprotein), 허리둘레, BMI (Body mass index) 관계를 객관적으로 표현하기 위하여 이종 데이터 병합과 1차 피드백 모델, 2차 피드백 모델을 파일럿 테스트로 구축하였다. 본 딥러닝 모델이 가지는 학습 모델의 정확도를 높이기 위한 조건 값으로서 실제 및 가상 데이터를 구분하게 된다. Fig. 7에서 보는 바와 같이 각기 다른 이종 데이터의 상관관계를 의사결정 트리의 중앙값에서 판별하여 해당 칼럼에 입력하는 방식의 강제 입력 (Force input)과, 그렇지 않을 경우 강제 삭제 (Force remove)로 구분하여 조건을 생성한다. 그 외의 경우에는 강제 무시 (Don't care)의 조건 값을 형성하게 되어 학습모델의 상관관계를 나타낸다 (Jeong, 2020). 예를 들면 대사증후군에서 공통적인 요소라고 할 수 있는 흡연이 폐질환에 어떠한 영향을 미치는 것인가에 대한 판단을 하고자 SNP (Single nucleotide polymorphism) (Furberg et al., 2010)를 통해 실험한다. 최초 흡연은 설문 조사시 기입한 환자의 흡연 여부를 문답식으로 0 과 1로 구분하였다. 이때 지노-타입 (Geno-type)의 유전적 성질 (Reesink et al., 2008)을 '15q24 susceptibility locus 염색체'의 영향이 라이프 로그 데이터와 얼마만큼의 영향을 미치는지

```
if relative with Data :
        Force Input
else if not relative with Data :
        Force Remove
else:
        Don't Care
```

Fig. 7 Condition Value Formation Process of Forced Input, Deletion and Disregard of Heterogeneous Data Processing

(Lambrechts et al., 2010) 살펴보았다. 또한 Bierut et al. (2010), Raaschou-Nielsen et al. (2008)에서 언급된 '15q24 susceptibility locus 염색체'의 경우 임상 데이터에 얼마만큼의 영향을 미치는지 여부를 객관적 지표 즉, 파라미터와 연관된 특정 영향 요소를 과학적으로 증빙할 수 있는 기초를 제시 할 수 있게 된다. 또한 데이터의 내용과 포맷의 형태는 설문데이터, 유전자 데이터의 염색체이름으로 실례로 들었다. 본 논문을 통한 이종 데이터의 과학적 분석 체계와 딥러닝 학습모델을 통한 객관적 우수성이 증명된다.

5. 성능 평가

결국 임상 데이터와 유전자 데이터를 포함하는 웨어러블 디바이스의 라이프 로그 데이터를 수집하여 데이터베이스에 전송하게 될 때 가장 중요한 사실은 학습 모델의 파라미터 값이 딥뉴럴 네트워크의 신경 (Neuron)이 제대로 동작(Bias)하는 가의 문제이다. 이를 위하여 인코딩되는 부분의 자동 인코더 (Denoise autoencoder)를 많은 문헌에서 대표적으로 사용하고 성능을 평가하고 있다 (Vincent et al., 2010).

이처럼 자동 인코더에서 추출된 이미지를 Training 데이터 (Training data)와 테스트 데이터 (Test data)로 구분하여 자동 인코더가 동작하는 경우를 살펴보게 된다. 이때 동작하는 입력 값으로서 흡연이 폐질환에 미치는 영향 요소와 지표를 객관적으로 살펴볼 수 있게 된다.

Fig. 8은 자동 인코더 (Denoise autoencoder)의 성능을 객관적으로 확인하기 위하여 노이즈(Noise)를 추가하여 해상도에 영향을 끼치는 경우의 이미지를 보여주고 있다 (우측 상단) (Azzeh et al., 2018). 이때 객관적 성능을 확인할 수 있는 벤치 마커 (Bench maker)를 사용하여 'SaltAndPepper' 노이즈를 추가하는 경우의 절차를 보여주어 자동 인코더가 동작하는 경우를 살펴보게 된다 (Esakkirajan et al., 2011).

결국 일반적으로 이미지에 생성되는 노이즈의 한 종류인 'SaltAndPepper' 노이즈를 적용하여

Fig. 8 Sample of the Object to which 'SaltAndPepper Noise', an Image Noise Addition Method for Denoise Autoencoder (Above Right)

'SaltAndPepper' 노이즈가 발생한 이미지에는 무작위적인 희고 검은 점이 나타나는 것을 증명한다. 일반적으로 이런 잡음을 줄이기 위해서는 중간 값 필터를 사용하게 되고 이를 실제 대사증후군중 흡연에 적용하게 되어 중간에 필터를 사용하여 노이즈를 줄어드는 것을 볼 수 있다 (Fu et al., 2019).

Fig. 9에서 보이는 바와 같이 학습 모델에 사용된 파라미터들의 관계를 나타내는 데에 이용되는 이종 데이터와 실제 예측을 할 칼럼을 추

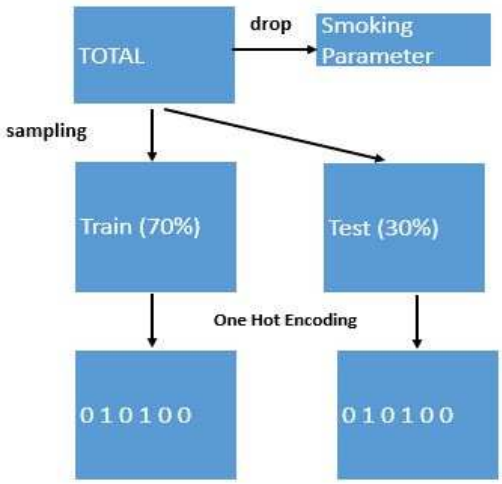

Fig. 9 The Actual Predicted Value of the Heterogeneous Data Result by Smoking in Metabolic Syndrome (for Smoking) Data Preprocessing Procedure Diagram

출 (대사증후군중 흡연의 경우의 특성 값의 예시)하는 경우를 제시하고 있다.

실제 딥뉴럴 네트워크를 진행하기 위하여 Training을 진행하여 피드백 모델의 결과 값을 예측하여 보도록 하였다. 실제 예측을 할 칼럼을 추출하여야 하고 이 경우에 흡연 (Smoking)의 경우를 전제로 하여 Training 데이터 (Training data)와 테스트 데이터 (Test data)를 7 : 3의 비율로 샘플링을 진행 하였다. 각각의 Training 데이터 (Training data)와 테스트 데이터의 원-핫 인코딩 (One-hot encoding)을 하여 Training이 가능하도록 전처리를 진행하였다.

일반적으로 집합의 크기를 벡터의 차원으로 하고, 표현하고 싶은 단어의 인덱스에 1의 값을 부여하고, 다른 인덱스에는 0을 부여하는 단어의 벡터 표현 방식을 일컫는데 이를 최종적으로 압축하여 표현할 수 있기 때문이다 (Buckman et al., 2018).

Fig. 10은 자동 인코더 (Denoise autoencoder)의 인코딩 과정에서의 데이터 압축과 1 EPOCH 당 파라미터 (W)값 전달 과정의 절차도를 시각화하였다. 이중에서 자동 인코더는 특정 알고리즘을 가진 지적재산권으로 출원 하였다. 즉, 자동 인코더의 Graph에서 Training 데이터와 테스트 데이터를 7 : 3의 비율로 각각 파라미터로 전달되고 이를 자동 인코더안의 신경 (Neuron)에 해당하는 DAE Graph 입력 직전에 1 EPOCH 당 'SaltAndPepper' 노이즈를 추가하고 자동 인코더 통과 이후 압축 데이터 반환하게 된다.

결국 자동 인코더 이후 Fig. 11에서 보이는 바와 같이 자동 인코더는 Graph 생성 이후 압축된 데이터를 반환 받아 입력 값 (Input) 으로, 기존의 남아있는 데이터인 흡연 (Smoke) 여부에 대한 데이터를 출력 값 (Output)으로 학습을 진행하게 된다. 그러므로 딥뉴럴 네트워크의 결과로 예측치 (Prediction data)와 정확도 (Accuracy)를 반환하게 된다.

딥뉴럴 네트워크의 신경망 모델 실행을 위한 입력 값 (흡연)과 출력 값 (예측치)의 생성 단계별 학습과정의 정확도 생성을 증명하였다. 딥러닝 모델은 앞서 언급한 바와 같이 이종 데이터의 1차, 2차 피드백 모델을 걸쳐 그 정확도를

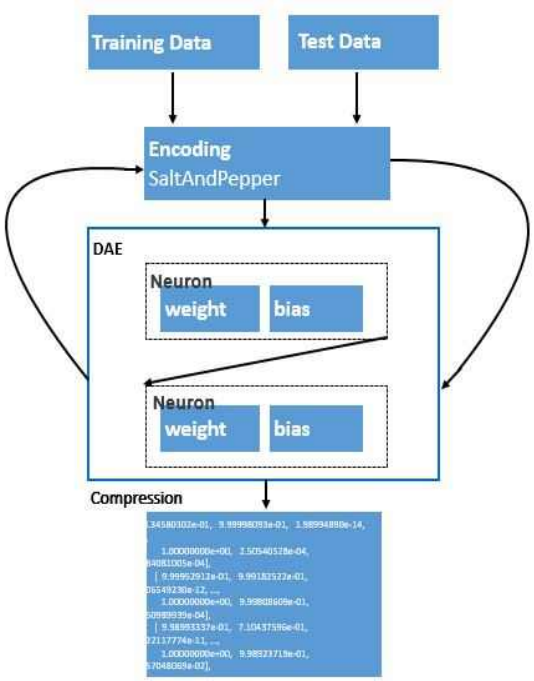

Fig. 10 Denoise AutoEncoder's Encoding Process for Data Compression and Parameter (W) per Epoch Transfer Process

높이게 되는 장점을 가지게 된다. 이때 구현된 딥뉴럴 네트워크의 신경망에서 신경 (Neuron)의 함수의 옵션들은 Fig. 12와 Fig. 13에서 비교하여 표현하고 있다.

Fig. 12에서 사용된 실제 데이터 (흡연의 경우) 노이즈 추가에 의한 변수 10에 의한 예측치 성능 비교를 제시하고 있다. 딥뉴럴 네트워크의 정확도를 측정하기 위하여 Epoch을 1000씩 간격으로 증가 시켜 보았을 때 약 39.5 %에서 41.5 %로 증가하는 것을 볼 수 있었다. 동시에 딥뉴럴 네트워크와 자동 인코더 (DAE + DNN)를 같은 방식으로 EPOCH을 1,000씩 간격으로 증가 시켜 보았을 때도 마찬가지로 71.2%에서 79.8%이상으로 증가 하는 것을 측정할 수 있었다.

마찬가지로 자동 인코더는 Graph 생성 이후 압축된 데이터를 반환 받아 예측치와 정확도를 계산하기 위한 노이즈 추가 변수를 10에서 40으로 증가 시켜 예측치의 성능을 비교할 필요가 생기게 된다. 1 Epoch 당 노이즈를 추가하고 자

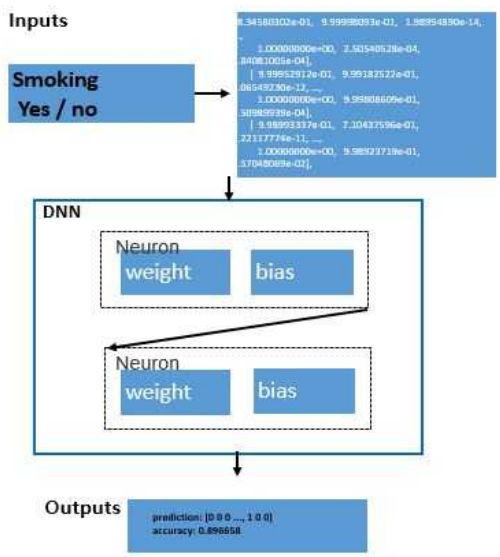

Fig. 11 Generation of Input Values (Smoking) and Output Values (Predicted Values) for the Execution of Neural Networks of Deep Neural Networks

동 인코더 통과 이후 압축 데이터 반환하게 되고 이에 미치는 영향이 예측치의 정확도와 관계가 있는지 살펴보게 된다. 우리는 본 실험을 통하여 실험군의 흡연여부를 설문에서 수집하여 이를 노이즈와 EPOCH 당 변화를 통하여 신경망 모델에서 미치는 영향을 살펴보게 된다.

그러므로, Fig. 13에서 사용된 실제 데이터 (흡연의 경우) 노이즈 추가에 의한 변수 40에 의한 예측치 성능 비교를 제시하고 있다. 딥뉴럴 네트워크의 정확도를 측정하기 위하여 Epoch을 1,000씩 간격으로 증가 시켜 보았을 때 약 79.2 %에서 86.5 %로 증가하는 것을 볼 수 있었다. 동시에 딥뉴럴 네트워크와 자동 인코더 (DAE + DNN)을 같은 방식으로 Epoch을 1000씩 간격으로 증가 시켜 보았을 때도 마찬가지로 86.3%에서 92.7%이상으로 증가 하는 것을 측정할 수 있었다.

언급한 바와 같이 Fig. 12와 Fig. 13에서 보이는 딥뉴럴네트워크의 정확도만을 이야기할 경우와 딥뉴럴네트워크와 자동 인코더를 합하여 노이즈를 추가하면서 EPOCH당 더 좋은 결과 값을 가지게 되는 것을 증명하였다.

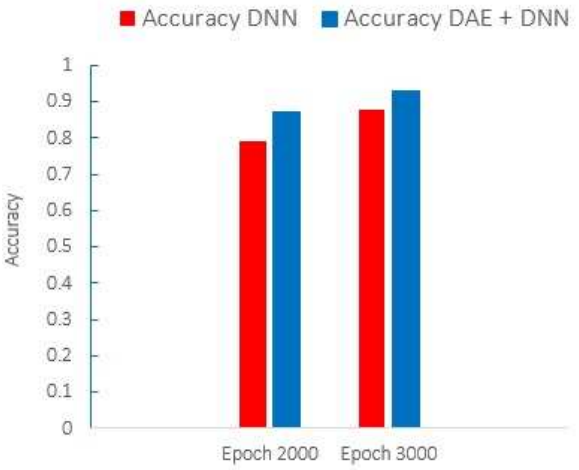

Fig. 12 Deep Learning-based Real Data for Disease Prediction (for Smoking) Comparison of Predicted Value Performance by Variable 10 by Adding Noise Level

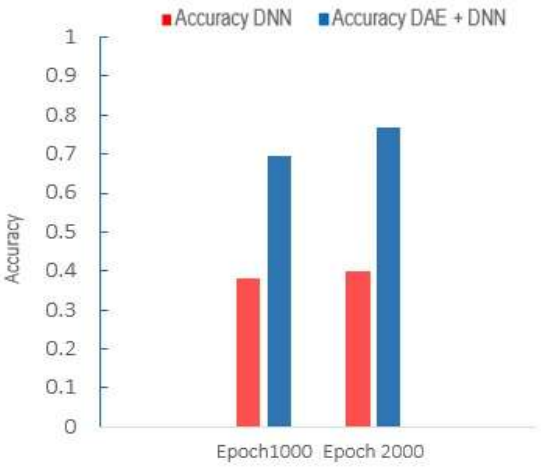

Fig. 13 Deep Learning-based Real Data for Disease Prediction (for Smoking) Comparison of Predicted Value Performance by Variable 40 by Adding Noise Level

6. 결 론

현대사회의 질병에 대한 불안과 두려움이 과학기술의 발전으로 말미암아 해소되고 있는 시

점에서 개인의 질병에 대한 명확한 치료를 위하여 딥뉴럴 네트워크상의 분석과 예측이 시행되고 있다. 이를 위한 딥러닝 학습 모델이 개발되고 개인의 웨어러블 디바이스의 질병에 대한 실시간 모니터링이 지속적으로 제공되고 있는 시점에서 헬스케어의 관점에서 질병예측 모델을 제시하도록 하였다.

과학기술과 메디컬 IT 도메인의 영역을 모두 아우를 수 있는 디지털 헬스케어의 영역의 발전과 새로운 신약 물질 또는 신약 후보물질을 인공지능의 관점에서 발굴하는 것이 전 세계의 추세이기 때문에 앞으로도 많은 발전과 획기적인 이익이 나올 것으로 보인다.

본 논문에서 제시한 딥뉴럴 네트워크를 통하여 제시된 웨어러블 디바이스의 착용 군과 비착용 군의 변화에 대한 객관적 성능 검증 방법으로 데이터의 전송과 딥러닝 피드백 모델을 접근하는 방식을 제시하였다. 이를 위해 웨어러블 디바이스의 데이터 전송에서 시작하여 질병예측을 위한 딥러닝 기반의 실제 데이터의 노이즈 추가에 의한 변수에 의하여 성능 비교를 실시하였다. 딥러닝을 통한 데이터 수집, 처리, 분석을 구현하였고 다시 이를 실제 딥뉴럴 네트워크에서 일반적으로 소개된 노이즈 추가의 방법으로 신경망 구성과 제시된 변수의 객관화된 벤치마커로서 시각화하는 기술적 구현을 완성하였다.

본 논문에서 사용한 딥러닝의 구현과 자동 인코딩의 일반적 방법은 직접 딥뉴럴네트워크를 제어하여 대사 증후군 중에서 흡연이 미치는 영향을 분석하였고, 멀티 채널에 의한 수집된 이종 데이터를 처리하여 사용자 참여를 유도하여 상호 관계를 보여주었다. 과학기술의 발전으로 말미암아 웨어러블 디바이스의 센싱 비율이 더욱 더 높아지게 되면 향후 질병예측 모델의 정확도가 높아지는 방향으로 앞으로의 연구의 초점을 맞추고자 한다.

References

Azzeh, J., Zahran, B., and Alqadi, Z. (2018). Salt and pepper noise: Effects and removal, JOIV, *International Journal on Informatics Visualization*, 2(4), 252-256.

Bierut, L. J. (2010). Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25, *Trends in Pharmacological Sciences*, 31(1), 46-51.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. (2018). Thermometer encoding: One hot way to resist adversarial examples, *International Conference on Learning Representations*.

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). *Mitosis detection in breast cancer histology images with deep neural networks. In International conference on medical image computing and computer-assisted intervention,* Springer, Berlin, Heidelberg. 411-418.

CNP. (2020). *COVID-19 Health monitoring function enhancement*, https://www.chosun.com/economy/tech_it/2020/10/03/F56YLXSXKFHMFOI3DLMPDTBP24/ (Accessed on Oct. 4th, 2020)

Dolph, C. V., Alam, M., Shboul, Z., Samad, M. D., and Iftekharuddin, K. M. (2017). Deep learning of texture and structural features for multiclass Alzheimer's disease classification, 2017 *International Joint Conference on Neural Networks (IJCNN)* 2259-2266.

Du, M., Liu, N., Song, Q., and Hu, X. (2018). Towards explanation of dnn-based prediction with guided feature inversion, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1358-1367

Esakkirajan, S., Veerakumar, T., Subramanyam, A. N., and PremChand, C. H. (2011). Removal of high density salt and pepper noise through modified decision based unsymmetric trimmed median filter,

IEEE Signal processing Letters, 18(5), 287-290.

Freudenberg, J., and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes, *Bioinformatics*, 18 (suppl_2), S110-S115.

Fu, B., Zhao, X., Li, Y., Wang, X., and Ren, Y. (2019). A convolutional neural networks denoising approach for salt and pepper noise, *Multimedia Tools and Applications*, 78(21), 30707-30721.

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., ... and Absher, D. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior, *Nature Genetics*, 42(5), 441.

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine, *Omics: A Journal of Integrative Biology*, 22(10), 630-636.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, *Nature Medicine*, 25(1), 65.

Hamet, P., and Tremblay, J. (2017). Artificial intelligence in medicine, *Metabolism*, 69, S36-S40.

Lee, H-S. and Jeong, T., (2019) *Institutional Review Board*, no. 201901-HR-003-02. 2019. 4. 26. pp. 1- 15.

Jeong, T. (2020). Time-series Data Classification and Analysis associated with Machine Learning Algorithms for Cognitive Perception and Phenomenon. *IEEE Access,* 12(3), 1-10, https://doi.org/10.1109/ACCESS.2020.3018477

Jeong, T. (2020). Deep Neural Network Algorithm Feedback Model with Behavioral Intelligence and Forecast Accuracy, *Symmetry*, 12(9), 1465-1476, https://doi.org/10.3390/sym12091465

Koul, A., Arnoult, E., Lounis, N., Guillemont, J., and Andries, K. (2011). The challenge of new drug discovery for tuberculosis, *Nature*, 469(7331), 483-490.

Mak, K. K., and Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects, *Drug Discovery Today*, 24(3), 773-780.

Mythili, T., Mukherji, D., Padalia, N., and Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL), *International Journal of Computer Applications*, 68(16), 134-142

Klok, F. A., Boon, G. J., Barco, S., Endres, M., Geelhoed, J. M., Knauss, S., ... and Siegerink, B. (2020). The Post-COVID-19 Functional Status scale: a tool to measure functional status over time after COVID-19, *European Respiratory Journal*, 56(1), 97-106

Lambrechts, D., Buysschaert, I., Zanen, P., Coolen, J., Lays, N., Cuppens, H., ... and Wijmenga, C. (2010). The 15q24/25 susceptibility variant for lung cancer and chronic obstructive pulmonary disease is associated with emphysema, *American Journal of Respiratory and Critical Care Medicine*, 181(5), 486-493.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., ... and Nevins, J. R. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes, *Proceedings of the National Academy of Sciences*, 101(22), 8431-8436.

Raaschou-Nielsen, O., Sørensen, M., Overvad, K., Tjønneland, A., and Vogel, U. (2008). Polymorphisms in nucleotide excision repair

genes, smoking and intake of fruit and vegetables in relation to lung cancer, *Lung Cancer*, 59(2), 171-179.

Reesink, H. W., Engelfriet, C. P., Schennach, H., Gassner, C., Wendel, S., Fontão Wendel, R., ... and Pham, B. N. (2008). Donors with a rare pheno (geno) type. *Vox sanguinis*, 95(3), 236-253.

Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., and Taheri, S. (2016). Sleep quality prediction from wearable data using deep learning, *JMIR mHealth and uHealth*, 4(4), e125.

Walker R. J. (2014). Seeing ourselves through technology: How we use selfies, blogs and wearable devices to see and shape ourselves, Springer Nature.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *Journal of Machine Learning Research*, 11(12), 187-199

Zhang, L., and Wu, X. (2006). An Edge-guided Image Interpolation Algorithm via Directional Filtering and Data Fusion, *IEEE Transactions on Image Processing*, 15(8), 2226-2238.

이 현 식 (Hyunsik Lee)

- 연세대학교 MBA 석사
- (현재) 차의과학대학교 대학원 의학과 박사과정
- 관심분야 : 인공지능, 메디컬 IT, etc

이 웅 재 (Woongjae Lee)

- 연세대학교 전기공학과 학사
- University of Illinois Computer Science, 석사
- Illinois Institute of Technology Computer Science, 박사
- (현재) 서울여자대학교 디지털미디어학과 교수
- 관심분야 : 인공지능, WSN, 자연어처리, 멀티미디어, etc.

정 태 경 (Taikyeong Jeong)

- 정회원
- University of Maryland, UC Computer Science 학사
- University of Texas at Austin Computer Eng.석사, 박사
- Cisco Systems, Inc. Engineer
- 2014년~2018년 서울여자대학교 컴퓨터학과 교수
- 2018년~2020년 차의과학대학교 교수
- (현재) 세한대학교 인공지능학과 교수
- 관심분야 : 인공지능, 알고리즘, etc.