



인공지능 시대의 마취통증의학 연구

이형철 · 정철우

서울대학교 의과대학 서울대학교병원 마취통증의학교실

Anesthesia research in the artificial intelligence era

Hyung-Chul Lee and Chul-Woo Jung

Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Korea

Received May 29, 2018

Accepted June 4, 2018

Corresponding author

Chul-Woo Jung, M.D.
Department of Anesthesiology
and Pain Medicine, Seoul National
University College of Medicine, Seoul
National University Hospital, 101
Daehak-ro, Jongno-gu, Seoul 03080,
Korea
Tel: 82-2-2072-0640
Fax: 82-2-747-5639
E-mail: spss@snuh.org

ORCID

http://orcid.org/0000-0001-7876-8659

A noteworthy change in recent medical research is the rapid increase of research using big data obtained from electrical medical records (EMR), order communication systems (OCS), and picture archiving and communication systems (PACS). It is often difficult to apply traditional statistical techniques to research using big data because of the vastness of the data and complexity of the relationships. Therefore, the application of artificial intelligence (AI) techniques which can handle such problems is becoming popular. Classical machine learning techniques, such as k-means clustering, support vector machine, and decision tree are still efficient and useful for some research problems. The deep learning techniques, such as multi-layer perceptron, convolutional neural network, and recurrent neural network have been spotlighted by the success of deep belief networks and convolutional neural networks in solving various problems that are difficult to solve by conventional methods. The results of recent research using artificial intelligence techniques are comparable to human experts. This article introduces technologies that help researchers conduct medical research and understand previous literature in the era of AI.

Key Words: Artificial intelligence, Big data, Machine learning, Medical research.

서론

의학 연구에서 최근 눈에 띄는 변화 중 무엇보다 먼저 거론될 것은 빅데이터와 인공지능이다. 2000년대 초반부터 각 병원에 도입된 전자의무기록(electrical medical record, EMR), 원무전산 시스템(order communication system, OCS), 의료 영상전송 시스템(picture archiving and communication system, PACS) 등은 기존의 수기 기록에서 볼 수 없었던 엄청난 양의 데이터를 축적할 수 있게 해 주었다. 그러나 이러한 빅데이터는 그 규모와 복잡성 때문에 기존의 고전적인 통계적 기법이나 한정된

성능의 통계 프로그램들에서 해석하기가 쉽지 않았다.

2016년 등장한 알파고(AlphaGo)에 의해 촉발된 딥러닝(deep learning)의 대유행은, 인공지능(artificial intelligence, AI) 기법에 대한 폭발적 관심을 이끌었다. 한 해 뒤인 2017년 발표된 일련의 의학 논문들은 빅데이터와 인공지능을 이용해 인간 전문가의 수준의 결과를 달성함으로써[1-3] 의료 인공지능(medical artificial intelligence)이라는 새로운 분야의 가능성을 보여주었다.

향후 인공지능 분야의 기술이 의학 연구에 광범위하게 도입되어 변화를 가져올 것은 분명해 보인다. 이들이 안전하고 유효하

게 만들어지기 위해서는 타당한 방법으로 만들어진 데이터를 적절한 기법으로 분석하는 것이 무엇보다도 중요하다. 또한, 이 과정에 의료 전문가들이 참여하기 위해 기존의 통계적 기법들과 다른 인공지능 알고리즘의 특성을 이해할 필요가 있다.

본 종설에서는 새로운 분야에 도전하는 역량 있는 연구자들에게 도움을 제공하고자, 인공지능 알고리즘의 전반적인 개념적 소개와 더불어 이와 관련된 마취통증의학 분야의 최신 연구들을 소개하고자 한다.

마취 영역에서의 인공지능 이용

인공지능이란 일반적으로 인간이 하는 지적인 일을 대체할 수 있는 시스템 혹은 알고리즘을 통칭하여 일컫는다. 이 중 규칙에 기반한 전문가 시스템, 피드백 컨트롤 기반 시스템 등은 매우 오래전부터 연구되었을 뿐만 아니라 실제 임상에서의 적용도 시도되어 왔다.

마취통증의학 영역에서는 이미 1950년에 Bickford [4]는 뇌파에 기반한 자동 마취를 성공시켰음을 보고하였다. 최근의 메타 분석에서 이중 분광 지수(bispectral index)에 기반한 프로포폴의 자동 투여는 마취 유도 시 요구량, 목표 마취 심도의 달성 비율, 짧은 각성 시간의 측면에서 마취과 의사보다 우월하다고 보고된 바 있다[5]. 또한, 이러한 인공지능 시스템을 이용한 자동 마취가 각성 시 섬망을 유의하게 줄일 수 있음이 보고 되어 예후 개선의 가능성 또한 제시되었다[6].

목표 지향 치료(goal directed therapy, GDT)는 술 후 합병증과 재원 기간을 단축시킬 수 있지만, 그 적용률은 여전히 낮다[7]. 인공지능 시스템을 이용한 자동 수액 투여 시스템은 그 대안으로 이용될 수 있다. 연구 결과에 따르면 맥박 윤곽(pulse contour) 알고리즘을 이용한 피드백 GDT 시스템은 안전하면서도 술 후 합병증과 재원기간을 단축시킬 수 있다[8].

진통 통각 지수(analgesia nociception index, ANI)에 기반한 레미펜타닐 투여와[9] 비침습적 혈압에 기반한 페닐에프린 자동 투여[10] 또한 시도되었다. 현재 이러한 전문가 시스템은 수면, 진통, 수액 투여를 모두 자동화한 시스템이 시도되고 있는 수준이며[11] 조만간 임상에 적용될 가능성이 있다. 또한, 시장에서 철수한 Sedasys 시스템(Ethicon Inc., USA)의 실패로부터 교훈을 얻은[12] AutoTIVA 시스템(NeuroWave systems Inc., USA)은[13] 미 해군의 지원을 받으며 Food and Drug Administration (FDA) 승인 작업이 진행 중이다.

빅데이터의 중요성

앞서 언급한 인공지능 분야의 성과들은 최근 주목받는 딥러닝 기술이 등장하기 이전의 기술들로 구현된 것들이다. 그러나

최근의 인공지능 분야의 놀라운 결과들은 기계 학습(machine learning), 그중에서도 딥러닝을 빼놓고 설명할 수 없다(Fig. 1). 기계 학습이 최근엔 폭넓게 이용되고 있는 이유 중 한 가지는 빅데이터의 등장이다. 기계 학습의 성능은 데이터의 양과 품질에 의해 크게 좌우되는데, 최근 들어서야 필요한 수준의 데이터가 이용 가능해졌기 때문이다. 전 세계에서 생산되는 디지털 데이터의 양은 지속적으로 증가해 2025년에는 매년 163 제타바이트(163조 기가바이트)가 될 것이라고 한다[14]. 의학 분야 연구에 활용될 수 있는 빅데이터로는 90년대 이후 축적되어 온 EMR과 PACS 상의 대규모 데이터나 정부나 보험 청구자료, 국가 단위의 레지스트리 등이 있다. 대한민국 정부에서도 ‘보건 의료 빅데이터 개방 시스템’(<http://opendata.hira.or.kr>)을 통하여 환자의 인구학적 특성과 더불어 진단, 투약, 시술, 재원 일수 등의 정보를 공개하고 있다. 2010년에 발표된 입원 환자의 원내 사망을 예측하는 위험도 층화 지수(risk stratification index, RSI)는 이러한 데이터를 활용한 연구의 모범 사례이다[15]. 이 지수는 미국 Medicare에서 공개한 Medicare provider analysis and review (MedPAR) 상의 2001년부터 2006년까지 입원 환자 35,179,507명의 진단, 수술코드를 이용해 재원 기간과 사망을 예측하였다. 개발된 알고리즘은 2007년부터 2012년까지 39,753,036명 데이터를 이용하여 외적 검증되었는데 원내 사망의 예측에 대한 수신자 조작 곡선 아래 면적(area under receiver-operating characteristic curve, AUROC)은 0.94였다[16]. 진단과 수술코드에 사망에 대한 중요한 힌트들이 들어가 있기는 하지만 이러한 결과는 매우 놀랍다. 당시에는 기계 학습 기법이 널리 쓰이기 전으로 저자들은 통계적 분석에 어려움을 겪었으며 결국 데이터의 1%만을 사용하여 P value 0.2 이상인 변수를 추출하고 10%에서 P value 0.05로 재추출한 후 최종적으로 로지스틱 회귀를 수행하는 방법을 이용하였다.

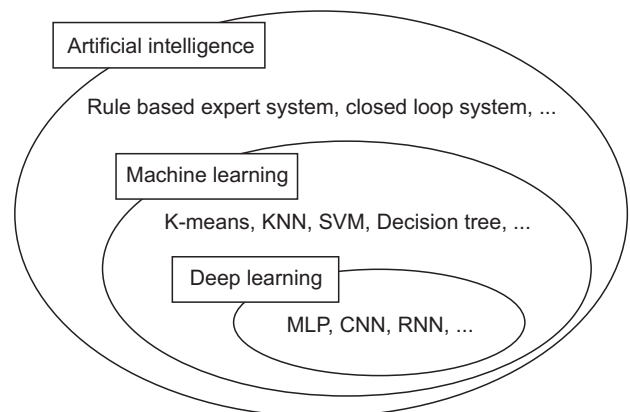


Fig. 1. A Venn-diagram of artificial intelligence. KNN: K-nearest neighbor, SVM: support vector machine, MLP: multi-layer perceptron, CNN: convolutional neural network, RNN: recurrent neural network.

마취통증의학 분야의 주요 연구 주제인 생체 신호 분야에서 대규모 데이터로는 매사추세츠공과대학교(Massachusetts Institute of Technology, MIT)에서 2001년부터 수집하여 공개한 Medical Information Mart for Intensive Care (MIMIC) 데이터베이스(<https://mimic.physionet.org>)에서 제공하는 중환자 데이터가 있으며, 최근 발표된 MIMIC-III의 경우 약 4만 명의 중환자실 환자 데이터를 포함하고 있다[17]. 수술 환자의 생체 신호 데이터베이스로는 저자들이 발표한 VitalDB (<https://vitaldb.net>)가 있다[18]. VitalDB 사이트에서는 다수의 환자 감시 장치로부터 기록된 6,388례의 수술 중 생체 신호 데이터와 함께 데이터베이스 작성에 사용된 데이터 수집 프로그램(Vital Recorder)이 무료로 제공되고 있다.

기계 학습의 소개

빅데이터의 분석에는 흔히 기계 학습 기법이 사용된다. 기계 학습이란 데이터로부터 자동 학습을 통하여 스스로 알고리즘의 성능을 높여가는 방법을 말한다. 고전적 통계 기법이 빅데이터에서 잘 사용되지 않는 이유는 빅데이터 내부의 많은 변수들 간의 복잡한 비선형 관계를 다루기에는 너무 단순하고, 많은 오차나 결측치를 포함하고 있는 데이터를 잘 처리하지 못하기 때문이다.

기계 학습 기법에서는 최종 모델의 가치를 평가하기 위해 P value 대신 손실 함수(loss function)를 사용한다. 흔히 사용되는 손실 함수로는 연속형 변수를 위한 평균 절대 에러(mean absolute error), 평균 제곱 에러(mean squared error) 등이 있고 범주형 변수를 위한 교차 엔트로피(cross entropy)가 있다.

학습 데이터로부터 손실 함수를 최소화하는 모델 파라미터를 찾는 과정을 학습이라 하는데 흔히 경사 하강법(gradient descending algorithm)이 사용된다. 경사 하강법은 손실 함수의 경사(gradient)를 계산하여 이를 따라 손실을 최소화하는 방향으로 파라미터를 조정한다. 이때 조정의 강도를 학습률(learning rate)이라 한다. 기계 학습 시 데이터 내의 여러 샘플을 사용하여 경사를 한 번에 계산한 후 그 평균을 이용해 파라미터를 한번에 조정하는 것을 배치 학습(batch learning)이라 한다. 배치 학습을 사용하면 수행 속도를 매우 높일 수 있는데, 그래픽 카드에 내장된 그래픽 프로세싱 유닛(graphic processing unit, GPU)과 같은 병렬 처리 장치를 연산에 사용할 때 특히 그러하다. 또한, 배치 학습 시 매 배치의 데이터를 자신의 평균과 표준편차로 정규화하는 것을 배치 정규화(batch normalization)라 한다. 배치 정규화를 사용하면 학습 데이터의 내부 분산이 줄어들어 초기에 학습이 옳은 길로 가게 되는 효과가 있다. 또한, 최종 모델의 예측력이 높아진다[19].

손실 함수에 일정 값을 얻어 패널티를 가하는 방법을 규제

(regularization)라고 하는데 대표적으로 L1 (lasso), L2 (ridge) 규제가 있다. 규제는 모델 파라미터의 개수가 많아지거나 몇 가지 파라미터가 한쪽으로 편중되면 좋지 않을 경우 사용한다. 규제의 개념은 최근에 등장한 것이 아니다. 역사적으로 모델 선택에 사용되어온 아카이케 정보 기준(Akaike information criterion, AIC) [20], 베이시안 정보 기준(Bayesian Information Criterion, BIC) [21]도 최대 가능도(maximum log-likelihood) 손실 함수에 파라미터 개수에 관한 규제를 더한 것으로 해석될 수 있다.

기계 학습 모델의 학습 및 성능 비교에는 흔히 교차 검증이 사용된다. 전체 데이터를 학습(training), 검증(validation), 시험(testing)군으로 나눈 다음[22], 학습군 데이터로 경사 하강하여 파라미터를 최적화하고 일정 주기(epoch)마다 검증 데이터로 검증한다. 학습이 진행될수록 모델은 학습군에만 최적화가 되는데 이를 과적합(overfitting)이라 한다. 과적합이 일어나면 모델의 예측 성능이 떨어지므로 그 전에 학습을 조기 중단해야 한다. 검증군 데이터는 학습의 조기 중단에 사용되며 최소의 검증 오차를 보이는 모델이 사용된다.

만일 데이터가 부족하여 검증군 데이터를 학습에 사용하지 못하는 것이 문제가 될 경우 k-fold 교차 검증을 이용할 수 있다. 이때에는 데이터를 k군으로 나누어 한 군씩을 검증군으로 이용하고 평균 검증 오차가 가장 적은 모델을 사용한다. 시험군 데이터는 오직 최종 학습된 모델의 성능 평가에만 사용된다.

고전적 기계 학습 기법

현재 사용되고 있는 기계 학습 알고리즘 중 일부는 과거로부터 사용되어 온 것들이지만, 여전히 사용할 만하고 매우 효율적이다(Table 1). 이들 알고리즘은 해결하려는 문제에 따라 군집화, 분류, 회귀 알고리즘으로 나누어 볼 수 있다(Fig. 2).

군집화 알고리즘(clustering algorithms)

군집화 알고리즘은 데이터를 속성의 유사함에 따라 지정한 개수의 군집으로 분류하는 알고리즘으로 데이터의 속성값만 존재하고 라벨이 존재하지 않으므로 비지도 학습(unsupervised learning)이라 한다. k-평균(k-means) 알고리즘이 흔히 사용된다.

분류 알고리즘(classification algorithms)

k-최근접 이웃(k-nearest neighbor) 알고리즘은 입력 공간상에서 가장 가까운 데이터 k개를 추출한 후 이들의 군 중 최다 빈도인 군으로 자신의 군을 정하는 방법이다. 이 방법은 현재에도 누락 데이터를 대체하는데 요긴하게 사용할 수 있다.

Table 1. Commonly Used Machine Learning Algorithms

Types	Method	Strength	Weakness
Unsupervised learning			
K-means	Classifies the data by the distances from K centers and update the center with the average of classified data	Labeling of learning data is unnecessary	Difficulty in determining K value
Supervised learning			
KNN	Finds the K closest samples from the data and classifies by the major group of the samples	Easy to interpret; Nonlinear	Difficulty in determining K value; The higher the dimension of the input variable, the lower the performance
ANN	Relationship is modelled by the perceptrons in input, hidden, output layers	Strong in the analysis of very complicated relations and the processing of outliers.	Difficult to interpret the final mode; Overfitting possibility; Long training time; Requires memory and processing power to learn
SVM	Determines the decision boundary as far as possible from the samples	Fast learning	Kernel selection is difficult for nonlinear relationships
Decision tree	Performs optimal classification sequentially to the desired level	Easy to interpret; Suitable for ensemble learning	Complex trees are difficult to interpret; Small deviations of the data affect the final tree; Overfitting possibility

KNN: K-nearest neighbor, ANN: artificial neural network, SVM: support vector machine.

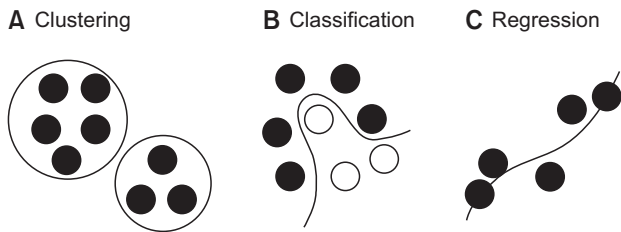


Fig. 2. Clustering, classification and regression methods.

Support vector machine (SVM)은 hinge 손실 함수에 ridge 규제를 더한 분류기이다[23]. 학습된 SVM 모델은 각 군 데이터로부터 최대한 멀리 떨어진 결정 경계를 갖는다. 비선형 관계를 갖는 데이터의 경우 커널(kernel) 함수를 이용해 입력 변수를 변형한 후 같은 방법을 적용한다.

의사결정나무(decision tree)는 특정 변수의 임계값을 기준으로 하는 이진 분류를 원하는 깊이까지 반복하는 분류기이다. 분류 기준 변수 및 값은 데이터로부터 자동으로 학습된다. 의사결정나무의 학습에는 경사 하강법이 아니라 CART (classification and regression tree) 알고리즘이 사용되는데 새넌 엔트로피(Shannon entropy) 혹은 지니 계수(Gini index)가 최소가 되도록 단계별로 노드를 추가하는 방법이다. 의사결정나무의 장점은 학습된 분류 결과를 사람이 쉽게 이해할 수 있다는 점이다.

분류 알고리즘을 적용할 때 주의해야 할 점은 군간 데이터 수의 불균형을 보정해야 한다는 점이다. 군간 불균형이 존재하면 데이터 수가 많은 군이 더 중요하게 취급되어 예측 성능이 떨어지게 된다. 적은 군의 데이터 수를 증가시키는 방법으로는 SMOTE (Synthetic Minority Over-sampling Technique) 알

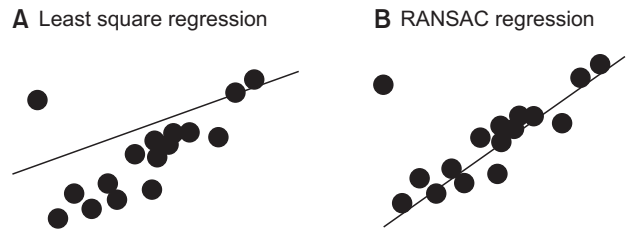


Fig. 3. Ordinary least square regression and RANSAC regression. RANSAC: RANDOM SAmple Consensus.

고리즘이 주로 사용된다[24].

회귀 알고리즘(regression algorithms)

고차원 모형에서 단계적 변수 선택에 기반한 회귀 모형은 그 성능이 매우 떨어진다. 따라서 변수 수를 늘릴 때마다 별점을 주는 규제의 방법을 사용하여 이를 보상한다. 대표적인 방법으로 라쏘 회귀(lasso regression)가 있다[25]. 데이터의 이상값 또는 특이점에 의한 영향을 줄이기 위해 데이터의 일부를 선택하여 학습시킨 후 이를 반복하는 로버스트 회귀(robustness regression) 기법을 이용할 수도 있다. 대표적인 방법으로 임의 추출 합의(RANDOM SAmple Consensus, RANSAC) 회귀가 있다(Fig. 3) [26].

앙상블 학습 기법(ensemble learning methods)

앙상블 학습은 데이터의 일부 샘플 혹은 일부 변수만을 이용

한 여러 모델을 학습한 후에, 이 모델들을 동시에 이용하여 단일 모형보다 좋은 예측 성능을 얻는 방법을 말한다. 앙상블 학습 기법의 대표로 배깅(bagging; bootstrap aggregating)과 부스팅(boosting)이 있다. 배깅은 1994년 Breiman [27]에 의해 제안되었으며, 학습 집합의 일부 샘플 혹은 일부 변수만 사용하여 결과 변수에 적합하는 과정을 여러 번 수행하여, 이들의 합으로 최종 출력을 결정하는 방법이다. 배깅이 좋은 성능을 내는 이유는 각 분류기의 개수가 증가할수록, 분류기의 예측값들이 내는 값들의 평균은 그 분산이 줄어들기 때문이다. 부스팅은 여러 분류기를 생성할 때 순차적으로 생성하는 방법을 말한다[28].

앙상블 학습 기법의 대표적인 예로 의사결정나무의 배깅인 랜덤 포레스트(random forest)가 있다[29]. 분류기의 성능을 비교 평가한 연구 결과에 따르면 121개 데이터 셋에 대해 랜덤 포레스트는 비교 대상인 179개 분류기 중 나머지를 압도하는 성능을 보였다[30]. 부스팅 모델의 대표적인 예로는 경사 부스팅(gradient boosting) 모델이 있다. 우수한 예측 성능을 갖는 모형의 상당수가 이 기법을 이용하므로 반드시 자신의 연구에 적용해 보기를 추천한다.

예를 들어 2016년 발표된 패혈증 예측 모델인 InSight의 4시간 전 예측값의 AUROC는 0.74였다[31]. 그러나 2018년에 같은 연구팀이 의사결정나무의 경사 부스팅을 적용한 결과, AUROC는 0.96으로 증가하였다[32]. 2018년 발표된 또 다른 연구는 병원 내 급성 신손상(acute kidney injury, AKI)을 예측하기 위해 의사결정나무의 경사 부스팅을 사용하였으며 AUROC는 0.90였다[33].

딥러닝 기법

딥러닝은 여러 층의 인공신경망을 사용한 기계 학습 기법을 말한다. 1943년 McCulloch과 Pitts [34]에 의해 소개된 인공 신경망(artificial neural network, ANN)은 뇌의 구조를 모방하는 알고리즘으로 개발되었다. 그러나 초기의 모델의 예측 성능은 그리 좋지 못하였고 이내 인공 신경망에 대한 관심이 줄어들게 되었다. 인공 신경망이 새롭게 부각된 것은 2000년대 후반 심층 신뢰망(deep belief network)과 컨볼루션 신경망(convolutional neural network)이 등장하면서부터다(Fig. 4).

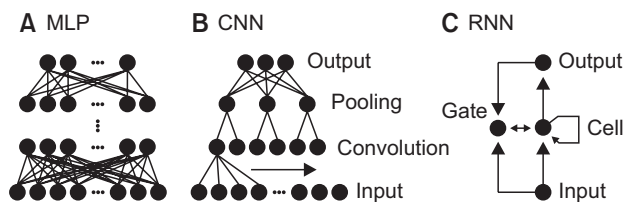


Fig. 4. Types of deep neural networks. MLP: multi-layer perceptron, CNN: convolutional neural network, RNN: recurrent neural network.

다층신경망(Multi-layer Perceptron, MLP)

인공 신경망을 이용한 1990년대의 많은 연구들은 1-2층의 은닉층(hidden layer)을 가지는 얇은 신경망(shallow network)을 이용하였다. 그러나, Hinton 등[35]이 빅데이터의 해석에 깊은 신경망이 효과적으로 작용하는 것을 보이고, 2010년 이후 드롭아웃(dropout) 기법, ReLU (Rectified Linear Unit) 함수 등이 개발되어 보다 깊은 층의 인공 신경망 학습이 가능해짐으로써 비로소 딥러닝의 시대가 열리게 되었다.

드롭아웃은 신경망의 학습 시 일부 노드를 고의로 누락하여 학습시키는 방법으로 앙상블 모델과 같은 효과를 나타낸다[36]. ReLU 함수는 오차의 지수적 감소를 없애 손실함수의 경사를 깊은 층까지 전달하는 역할을 한다[37].

단순히 다층 신경망을 쌓은 것만으로 좋은 결과를 얻기는 힘들지만 기본적인 회귀나 분류 성능을 확인하기 위해 신경망의 적층을 시도 해 볼만 하다. Lee 등[38]은 술 후 원내 사망을 예측하는 4층의 은닉층을 가지는 MLP 모델을 만들었으며 59,985명 환자의 46개 특징과 미국마취과학회(American Society of Anesthesiologists) 신체 등급 점수를 추출하여 학습시켰다. 그 결과 진단 코드 등의 중요한 힌트 없이도 원내 사망을 예측할 수 있었고 AUROC는 0.91였다. 그러나 그 연구에서 로지스틱 회귀 모형의 AUROC도 0.90으로 비슷한 결과를 보였다.

합성곱 신경망(Convolutional Neural Network, CNN)

Lecun 등[39]에 의해 소개된 CNN은 합성곱을 이용한 인공신경망으로 주로 음성, 이미지 데이터 처리에 사용된다. 특히 이미지 데이터에서 필요한 요소를 추출하기 위한 필터셋을 자동적으로 학습하는 것을 특징으로 한다. GoogLeNet과 같은 잘 학습된 CNN 모델에 내부의 필터셋들은 다른 문제 해결에도 그대로 활용될 수 있는데 이러한 방법을 전이 학습(transfer learning)이라고 한다. 전이 학습을 이용하면 기존에 학습된 모델을 이용함으로써 새로운 모델의 개발 시 적은 수의 데이터로도 매우 우수한 성능의 모델을 개발할 수 있다.

의학과 관련된 딥러닝 분야에서 2017년은 기억할 만한 해이다. 한 해 동안 발표된 일련의 논문들의 성과는 매우 놀랍다. 이들은 모두 CNN과 전이 학습을 이용하였고 병리과, 안과, 피부과, 영상의학과 영역에서 전문의를 능가하였다. 구글에서 발표한 당뇨병 망막병증을 예측하는 모델은 GoogLeNet에서 전이 학습한 모델을 사용하였으며 AUROC는 0.99였다[2]. 피부병변으로 피부암을 진단하는 모델 또한 GoogLeNet에서 전이 학습한 모델을 사용하였으며 AUROC는 0.96이었다[3]. 흉부 X선 사진으로부터 폐결핵을 진단하는 알고리즘에서 AlexNet과 GoogLeNet에서 전이 학습한 모델의 앙상블이 사용되었고

AUROC가 0.99였다[40]. 유방암 림프절 전이 여부에 대한 병리 판독 모델은 2개의 GoogLeNet에서 전이 학습된 모델의 앙상블을 사용하였고 AUROC는 0.99였다[1].

CNN이 영상 분석에만 쓰이는 것은 아니다. Rajpurkar 등[41]은 심전도 분석에 1D (1 dimension)-CNN을 적용하였고 레이어를 무려 34층이나 쌓았다. 결과적으로 심장 전문의보다 대부분의 부정맥 진단에서 더 정확한 심전도 판독기를 구현할 수 있었다.

순환 신경망(Recurrent Neural Network, RNN)

RNN은 내부에 자기 자신으로 되돌아오는 레이어가 있는 인공신경망으로 주로 시계열 데이터의 처리를 위해 사용된다. RNN에서 순환층의 연결 강도는 각 시계열 데이터에 대해 같게 적용되므로 손실 함수의 경사는 시계열의 길이가 길어질수록 지수적으로 변해 0 혹은 무한대가 된다. 이 문제 때문에 비교적 최근까지 깊은 RNN은 학습이 불가능했다. 그러나 1997년 발표된 장-단기 기억(long short-term memory, LSTM) [42] 모델과 gated recurrent unit (GRU) [43]가 발표되면서 상황이 바뀌었다. 이러한 모델들은 게이트(gate)를 두어 출력단의 오차를 상당히 깊은 층까지 전달함으로써 RNN의 학습이 가능하게 하였다.

시계열을 이용하는 대부분의 연구들에서 LSTM이 사용되고 있다. 본 저자들이 발표한 프로포폴과 레미펜타닐의 주입 이력으로부터 BIS를 예측하는 연구에서 LSTM은 propofol과 remifentanyl 투여 시 시간에 따른 투여 약물의 약동약력학적 변화를 모사하기 위해 사용되었으며 BIS 예측에 있어 기존의 response surface model 보다 오차를 절반 가량으로 줄일 수 있음을 보였다[44].

인공지능 기법을 이용한 연구의 수행

지금까지 알아본 인공지능 기법을 실제 연구에 적용하기 위해서는 먼저, 대규모 데이터가 필요하다. 양질의 데이터의 수집은 무엇보다 중요하다. 다음으로 수집한 데이터가 해결할 수 있는 임상적 문제에 관한 아이디어가 필요하다. 본 저자들이 생각하는 좋은 아이디어란 다음과 같다. 우선, 예측 변수의 임상적 중요성이 있어야 한다. 다음으로 입력 변수는 개입 가능한 인자여야 한다. 마지막으로 입력과 출력 간의 예측력이 충분히 높아야 한다. 문제에 따라 다르겠지만 최근 연구를 보면 AUROC나 accuracy 기준으로 최소 0.9 혹은 90% 이상이 대부분이다. 출력 변수에 대한 이 정도 정확한 예측이 가능하도록 모든 정보가 입력 변수에 들어가 있는지를 검토해야 한다.

다음으로 그 문제가 군집 문제인지, 분류 문제인지, 회귀 문제인지에 따라 적절한 인공지능 모델을 선택한다. 마지막으로 이

를 구현하고 검증한다. 현재까지 그래픽 유저 환경에서 데이터의 전처리와 기계 학습 기법의 적용까지 한번에 아우를 수 있는 적절한 프로그램이 없으므로, 효과적인 학습을 위해서는 연구자가 직접 Python이나 R 등의 프로그래밍 언어를 이용하여 코딩할 필요가 있다. 다행스럽게도 코딩의 편의를 위한 기계 학습 관련 라이브러리들은 대부분 무료로 공개되어 있다. Python은 TensorFlow, scikit-learn, keras 라이브러리가 흔히 사용된다. R 언어에서는 각 기법 별로 패키지가 나와있으며 종합적인 패키지로는 h2o, keras가 유명하다. 많은 파라미터를 가진 모델의 학습에는 오랜 시간이 요구되는데 GPU가 있으면 이를 매우 단축시킬 수 있다.

결론

기술의 발전은 의료에 적용되어 의료 현장을 완전히 바꾸어 놓는다. 1950년대 심박동 조율기 등 의용 전자 분야가 시작될 수 있었던 것은 이보다 앞선 시기에 트랜지스터의 발명으로 기기의 소형화가 가능했기 때문이다. 1990년대에 BIS 모니터, 목표 농도 주입 펌프(target-controlled infusion pump), 맥박 윤곽 알고리즘 기반 심박출량계 등 혁신적인 의료 장비들이 동시에 등장한 것은 앞선 시기에 마이크로프로세서의 발전이 있었기 때문이다. 몇십 년이 지난 후 현시대를 되돌아보면 어떠할까? 저자들은 현시대가 빅데이터에 기반한 인공지능 의학 연구가 태동한 시기로 기억되지 않을까 생각한다. 이러한 시대적 변화에 맞춰 최신 기술에 관심을 가지고 배움의 노력을 기울인다면, 인공지능 기술을 의학 연구에 강력한 도구로 활용할 수 있을 것이다. 역량 있는 연구자들이 새로운 시대의 연구 개발 도구인 인공지능 학습의 기법들에 익숙해지고 이와 관련된 연구에 큰 성취를 이룰 수 있기를 기대한다.

REFERENCES

1. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017; 318: 2199-210.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402-10.
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115-8.
4. Bickford RG. Automatic electroencephalographic control of general anesthesia. Electroencephalogr Clin Neurophysiol 1950; 2:

- 93-6.
5. Pasin L, Nardelli P, Pintaudi M, Greco M, Zambon M, Cabrini L, et al. Closed-loop delivery systems versus manually controlled administration of total IV anesthesia. *Anesth Analg* 2017; 124: 456-64.
6. Cotoia A, Mirabella L, Beck R, Matrella P, Assenzo V, Chazot T, et al. Effects of closed-loop intravenous anesthesia guided by bispectral index in adult patients on emergence delirium: a randomized controlled study. *Minerva Anestesiol* 2018; 84: 437-46.
7. Srinivasa S, Kahokehr A, Soop M, Taylor M, Hill AG. Goal-directed fluid therapy- a survey of anaesthetists in the UK, USA, Australia and New Zealand. *BMC Anesthesiol* 2013; 13: 5.
8. Joosten A, Coeckelenbergh S, Delaporte A, Ickx B, Closset J, Roumeguere T, et al. Implementation of closed-loop-assisted intra-operative goal-directed fluid therapy during major abdominal surgery: a case-control study with propensity matching. *Eur J Anaesthesiol* 2018. doi: 10.1097/EJA.0000000000000827. [Epub ahead of print].
9. De jonckheere J, Delecroix M, Jeanne M, Keribedj A, Couturier N, Logier R. Automated analgesic drugs delivery guided by vagal tone evaluation: interest of the analgesia nociception index (ANI). *Conf Proc IEEE Eng Med Biol Soc* 2013; 2013: 1952-5.
10. Ngan Kee WD, Tam YH, Khaw KS, Ng FF, Lee SWY. Closed-loop feedback computer-controlled phenylephrine for maintenance of blood pressure during spinal anesthesia for cesarean delivery: a randomized trial comparing automated boluses versus infusion. *Anesth Analg* 2017; 125: 117-23.
11. Joosten A, Jame V, Alexander B, Chazot T, Liu N, Cannesson M, et al. Feasibility of fully automated hypnosis, analgesia, and fluid management using 2 independent closed-loop systems during major vascular surgery: a pilot study. *Anesth Analg* 2018. doi: 10.1213/ANE.0000000000003433. [Epub ahead of print].
12. Goudra B, Singh PM. Failure of sedasys: destiny or poor design? *Anesth Analg* 2017; 124: 686-8.
13. Bibian S. Closed loop total intravenous anesthesia (TIVA) for combat casualty care - an auto-pilot for safe and effective TIVA delivery. Paper presented at: Society of Technology in Anesthesia Annual Meeting 2018; 2018 Jan 10-13; Aventura (FL), USA. Available from http://www.stahq.org/userfiles/ckfiles/files/STA_18AM_Handout_Bibian.pdf.
14. Reinsel D, Gantz J, Rydning J. Data age 2025: the evolution of data to life-critical. International Data Corporation [serial on the Internet]. 2017 [2018 Jun 5]. Available from <https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>.
15. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG. Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *Anesthesiology* 2010; 113: 1026-37.
16. Chamoun GF, Li L, Chamoun NG, Saini V, Sessler DI. Validation and calibration of the risk stratification index. *Anesthesiology* 2017; 126: 623-30.
17. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.
18. Lee HC, Jung CW. Vital recorder-a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. *Sci Rep* 2018; 8: 1527.
19. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Paper presented at: 32nd International Conference on Machine Learning; 2015 Jul 7-9; Lille, France. arXiv preprint arXiv:150203167.
20. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974; 19: 716-23.
21. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc. Ser B (Methodological)* 1974; 36: 111-47.
22. Rumelhart DE, Widrow B, Lehr MA. The basic ideas in neural networks. *Commun ACM* 1994; 37: 87-92.
23. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273-97.
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321-57.
25. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological)* 1996; 58: 267-88.
26. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 1981; 24: 381-95.
27. Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123-40.
28. Freund Y. Boosting a weak learning algorithm by majority. *Inf Computat* 1995; 121: 256-85.
29. Breiman L. Random forests. *Mach Learn* 2001; 45: 5-32.
30. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* 2014; 15: 3133-81.
31. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016; 4: e28.
32. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018; 8: e017833.
33. Koyner JL, Carey KA, Edelson DP, Churpek MM. The develop-

- ment of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med* 2018; 46: 1070-7.
34. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol* 1990; 52: 99-115; discussion 73-97.
35. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006; 18: 1527-54.
36. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929-58.
37. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Paper presented at: the 27th international conference on machine learning; 2010 Jun 21-24; Haifa, Israel. Paper ID: 432.
38. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology* 2018. doi: 10.1097/ALN.0000000000002186. [Epub ahead of print].
39. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86: 2278-324.
40. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; 284: 574-82.
41. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv.org* [serial on the Internet]. 2017 [2018 Jun 5]. Available from <https://arxiv.org/abs/1707.01836>.
42. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9: 1735-80.
43. Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv.org* [serial on the Internet]. 2014 [2018 Jun 5]. Available from <https://arxiv.org/abs/1412.3555>.
44. Lee HC, Ryu HG, Chung EJ, Jung CW. Prediction of bispectral index during target-controlled infusion of propofol and remifentanyl: a deep learning approach. *Anesthesiology* 2018; 128: 492-501.