KIPS Trans. Softw. and Data Eng.
Vol.11, No.5 pp.203~210
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.5.203

Deep Learning-Based Model for Classification of Medical Record Types in EEG Report

Kyoungsu Oh† · Min Kang†† · Seok-hwan Kang††† · Young-ho Lee††††

ABSTRACT

As more and more research and companies use health care data, efforts are being made to vitalize health care data worldwide. However, the system and format used by each institution is different. Therefore, this research established a basic model to classify text data onto multiple institutions according to the type of the future by establishing a basic model to classify the types of medical records of the EEG Report. For EEG Report classification, four deep learning-based algorithms were compared. As a result of the experiment, the ANN model trained by vectorizing with One-Hot Encoding showed the highest performance with an accuracy of 71%.

Keywords : Deep Learning, EEG Report Classification, Natural Language Processing

EEG Report의 의무기록 유형 분류를 위한 딥러닝 기반 모델

오 경 수† · 강 민†† · 강 석 환††† · 이 영 호††††

요 약

보건의료 데이터를 사용하는 연구 및 기업이 늘어나며 세계적으로 보건의료 데이터 활성화를 위한 노력을 진행 중이다. 하지만 기관에 따라 사용하는 시스템과 서식이 다르다. 이에 본 연구는 EEG Report의 의무기록 유형을 분류하는 기저 모델 구축을 통해 향후 다기관의 텍스트 데이터를 유형에 따라 분류하는 기저 모델을 구축하였다. EEG Report 분류를 위해 4가지의 딥러닝 기반 알고리즘에 대해 비교하였다. 실험 결과 One-Hot Encoding으로 벡터화하여 학습한 ANN 모델이 71%의 정확도로 가장 높은 성능을 보였다.

키워드 : 딥러닝, EEG Report 분류, 자연어처리

1. 서 론

전자 건강 기록(Electronic Health Records, EHR)은 환자의 종이 차트의 필기 문제로 인한 오류를 줄이고자 작성되기 시작하였다. 실제 미국 보건의료 정보기술조정국의 통계에 따르면 2014년 기준 미국 97%의 병원에서 인증된 EHR 시스템을 보유하고 있다[1]. 이처럼 대부분 병원에서 EHR 시스템에 사용으로 EHR 데이터가 증가하였다. 특히 EHR 데이터 중 뇌파에 대한 활용이 많아지고 있다.

뇌파검사(electroencephalography, EEG)는 뇌의 각 고유 영역의 기능을 객관적으로 평가하기 위한 검사이다. EEG를 통해 뇌파 신호가 기록되고 임상 전문의에 의해 EEG 결과 보고서(EEG Report)가 작성된다. 이러한 EEG를 통해 주로 간질, 뇌졸중, 치매 등과 같은 뇌 질환과 원인 모르는 의식장애, 정신 질환 등을 진단할 수 있다[2]. 또한, EEG를 통해 발생한 기록된 뇌파 신호를 활용하여 간질 유형을 분류하는 연구[3,4]도 활발히 진행되고 있다.

Gao Y는 EEG 신호를 PSDED(Power Spectrum Density Energy Diagrams)로 변환하였고, DCNNs(Deep Convolutional Neural Networks) 및 전이 학습을 통해 PSDED의 특징을 추출하여 간질 상태를 분류하였다[3]. Ranghu S는 간질 발작 유형을 분류하기 위해 CNN(Convolutional Neural Networks)를 사용한 방법을 제안하였다[4]. 하지만 EEG 신호 데이터를 활용한 연구는 활발하지만 EEG Report 데이터를 활용한 연구는 많지 않다. 따라서 본 연구는 EEG Report를 활용을 위한 연구를 진행하였다.

텍스트로 된 EHR 데이터를 활용한 기저 연구[5,6]에서는

※ 이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT센터육성지원사업의 지원으로 연구를 수행하였음(IITP-2021-2017-0-016630).
† 준 회 원 : 가천대학교 컴퓨터공학과 학사과정
†† 비 회 원 : 가천대학교 IT융합공학과 석사과정
††† 비 회 원 : 가천대학교 컴퓨터공학과 부교수
†††† 비 회 원 : 가천대학교 컴퓨터공학과 교수
Manuscript Received : September 9, 2021
Accepted : October 12, 2021
* Corresponding Author : Young-ho Lee(lyh@gachon.ac.kr)

환자의 상태를 분류하였다. Whang Y는 SVM, 랜덤 포레스트, MLPNN(Multilayer Perceptron Neural Networks), CNN에 대해 Mayo Clinic에서 1차 진료를 받은 환자 코호트의 흡연 상태 분류하거나 근위 대퇴골 골절 발생 여부를 분류하였고 i2b2 2006 흡연 상태 분류 데이터에 대해 테스트를 진행하였다[5]. Goodwin TR은 EEG Report 데이터를 Word2vec를 통해 벡터화시켜 DAN(Deep Averageing Network)과 LSTM(Long Short-Term Memory) 모델에 넣어 학습 시켜 EEG를 한 환자의 정상 유무를 분류하였으며 약 91%의 정확도를 얻었다[6]. 하지만 EEG 검사를 한 환자가 간질 검사, 외래 진료 환자 등 구분이 이루어지지 않았다. EEG 환자를 구분하기 위해서는 EEG Report의 텍스트 분류가 이루어져야 한다.

텍스트 분류는 단어, 문장 또는 문서 전체를 이용하여 텍스트의 카테고리를 분류하는 작업이다. Kim은 SNS의 포스터의 스팸 여부를 분류하는 연구를 진행하였으며[7], Kim은 한국어 신문 기사의 카테고리를 분류하는 연구를 진행하였다[8]. Liu는 7개의 감성 분석 데이터에 대해 성능을 비교하여 AC-BiLSTM(Attention-based Bidirectional Long Short-Term Memory) 모델을 제시하였다[9]. 이처럼 다양한 분야에서 텍스트 분류를 위한 연구가 진행되고 있다. 하지만 텍스트로 작성된 의무기록의 경우 환자의 특정 상태를 분류하는 연구는 많이 이루어졌지만, 의무 기록의 발생한 곳이나 유형에 대한 분류는 거의 이루어지지 않고 있다[10,11].

진료기록을 스캔하여 보관하는 과정에서 다양한 진료기록이 혼재되는 문제를 극복하고자 의료 기록 이미지와 텍스트를 사용하는 연구가 진행되었다[10]. 또한, 의학 텍스트는 의학 분야의 개념 또는 약어를 나타내는 의학 용어가 포함되며, 문법적으로 문제가 있는 문장들로 인해 발생되는 분류의 어려움을 해결을 위한 연구도 진행되었다[11]. 이처럼 의학 텍스트의 전문 용어 및 보관 과정에서 혼재되는 문제를 해결할 필요성은 높아졌지만, 환자의 상태 분류 연구에 비해 부족한 실정이다. 특히 EEG Report의 경우 외래, 입원, 집중치료실, 간질 모니터링 등 다양하게 작성되지만 환자의 상태 분류를 위한 연구만 진행되었다.

보건의료 분야에 인공지능 등의 기술이 적용되며 여러 기업에서 활용 중이다[12]. 보건의료 데이터가 많은 곳에서 활용됨에 따라 미국은 데이터를 쉽게 찾을 수 있고, 접근이 용이하며, 상호 호환할 수 있고, 재활용 가능해야 한다는 연구데이터 개방 원칙 FAIR[13]을 발표하였다. 또한, 미국 국립보건원은 바이오메디컬 데이터 세트 접근 및 컴퓨팅에 대한 경제적, 기술적 장벽을 낮추는 것을 목표로 STRIDES Initiative[14]를 발표하였다.

국내의 경우 보건의료 데이터 활성화를 위해 2020년 1월 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」 개정안이 국회 본회의를 통과하였다[15]. 또한, 보건의료 데이터 결합을 지원하는 결합 전문기관으로 국민건강보험공단, 건강보험심사평가원, 한국보건산업진흥원을 지정하여 보건의료 데이터 활성화에 노력하고 있다. 하지만 의료기관이 폐업했을 경우 기록물 관리 및 보존에 있어서 폐업 의료기관이 보건소로 기록을 이관하지 않거나 전자의무기록 시스템이 의료기관마다 사용하는 시스템 및 서식이 다른 문제 등이 있다[16].

의료기관마다 사용하는 시스템 및 서식을 다른 부분을 해결하고자 표준화에 필요성이 대두되고있다[17]. 하지만 표준 모델 도입에 있어 비용, 데이터 비표준화, 전문 인력 부재 등의 어려움을 겪고 있다[18]. 이에 본 연구는 EEG Report의 의무기록 유형을 분류하는 모델을 제시하였다. 이를 통해 향후 여러 보건의료기관의 텍스트로 된 데이터를 유형, 서식 등에 따라 분류하여 활용성을 높이는 기저 연구로써 기반을 구축하고자 한다.

2. 관련 연구

2.1 의무기록

의무기록은 환자에게 일관성 있고 지속적인 치료를 제공할 수 있는 근거 자료로서 과거 병력과 치료의 내용을 알려줌으로써 진단과 치료의 방향을 쉽게 설정하게 해 준다[19]. Y 병원의 경우 의무기록은 업무에 따라 응급진료기록, 초진 기록, 입원기록 등 다양하게 존재[20]한다. 종류에 따라 서식, 기재 항목 등의 내용이 다르게 나타난다.

본 연구에서 사용한 EEG Report의 의무기록 유형은 외래 진료 환자(Outpatient)의 기록, 입원 진료 환자(Inpatient)의 기록, 집중치료실 환자(Intensive Care Unit, ICU)의 기록, 간질 모니터링 환자(Epilepsy Monitoring Unit, EMU)의 기록을 종속변수로 사용하였다.

2.2 EEG Report

EEG Report는 의무기록의 한 종류로 두피 전체의 정해진 위치에 전극 크림을 이용해 전극을 두피에 붙이고 진행된 검사를 신경과 전문의가 검토하여 작성된다. 미국의 경우 임상신경 생리학 학회(American Clinical Neurophysiology Society, ACNS)의 EEG Report 작성 지침[21]에 따라 다음 세 가지 주요 부분으로 구성되어야 한다.

1) 서론

서론은 기록 시작 시 환자의 의식상태, 금식 여부, 일반 약물과 관계없이 환자가 받는 약물 등에 대한 내용이 포함되어야 한다. 또한, 사용된 전극의 수가 10-20시스템의 표준 21이 아닌 다른 생리학적 매개변수의 모니터링이 사용되는 경우 도입부에 이를 명시해야 한다. 특별한 이유로 인해 총 기록 시간이 ACNS의 권장보다 길거나 짧은 경우 총 기록 시간에 대해 보고하는 것이 권장된다.

2) 기록에 대한 설명

EEG에 대한 정상 및 비정상 기록의 모든 특성을 포함하

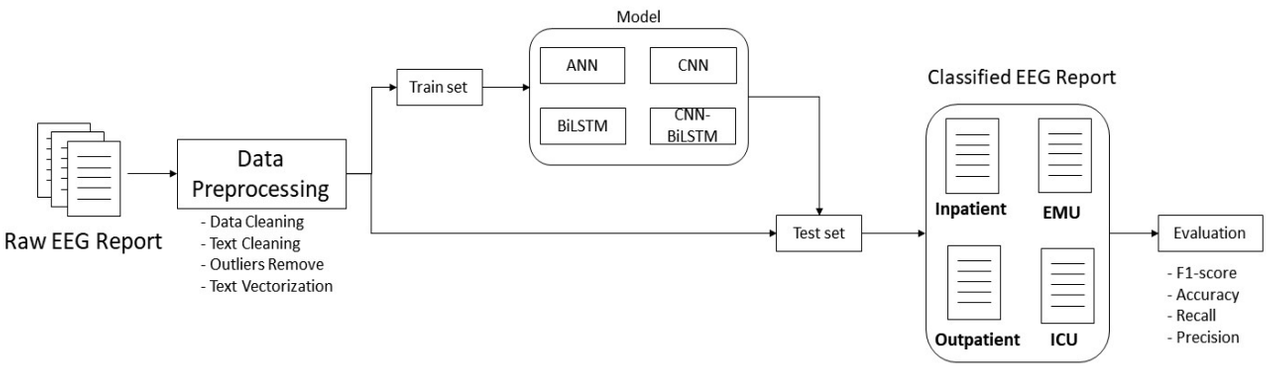

Fig. 1. Research Pipeline

며, 객관적으로 작성해야 한다. 다른 전문의가 EEG를 보지 않고도 EEG Report만으로 정상 또는 비정상 정도에 관한 결론에 도달할 수 있게 작성된다. 설명에서는 진폭, 대칭 여부 등의 내용을 포함하며 주파수는 Hz 또는 초당 사이클 단위로 지정한다.

3) 해석

EEG 기록의 정상 유무에 대한 주관적인 기록으로 가능한 간결하게 작성해야 한다. 또한, EEG 소견이 전체 임상 사진에 어떻게 맞는지 또는 맞지 않는지를 설명하려는 임상적 상관관계에 대해 작성하여야 한다.

3. EEG Report 분류 모델

3.1 실험 환경 구성

본 연구에서는 EEG Report의 의무기록 유형을 분류하기 위해 구글에서 제공하는 클라우드 실험 환경인 코랩(Colaboratory)을 사용하였다. 코랩은 컴퓨터의 사양과 관계없이 웹 브라우저를 통해 동일한 환경에서 Python을 실행할 수 있다. 본 연구에서 사용한 코랩의 사양은 “Python 3.7.11, Ubuntu 18.04.5 LTS, Intel® Xeon® CPU @ 2.00GHz, MemTotal 13302928kB" 이다. 본 연구에서 사용한 플랫폼은 데이터 전처리를 위해 Pandas 1.1.5와 nltk 3.2.5를 사용했으며, 모델 구축에는 Keras 2.6.0을 사용하였다. 모델의 성능 비교를 위해 Scikit-learn 0.22.2.post1을 사용하였다.

3.2 EEG Report 분류 모델 설계

Fig. 1는 본 연구의 파이프라인이다. 먼저 EEG Report의 4가지 유형을 분류하기 위해 EEG Report에 대한 데이터 전처리를 진행한다. 전처리된 데이터 중 20%는 평가를 위해 사용하며 80%는 학습을 위해 사용한다. 학습은 인공신경망(Artificial Neural Network, ANN), CNN, BiLSTM (Bidirectional Long Short-Term Memory), CNN과 BiLSTM을 결합한 CNN-BiLSTM 모델에 대해 이루어졌다. 각 모델의 성능을 비교하기 위해 F1-score, 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)에 대해 평가하였다.

CLINICAL HISTORY: This is an 82-year-old woman who presented with unresponsiveness initially with a Glasgow score of 3, found to have a hemiparesis, intubated and gradually improved. Initial CT scans were negative for stroke.
MEDICATIONS: PHT, Ativan
INTRODUCTION: Digital video EEG is performed at the bedside in the ICU using standard 10-20 system of electrode placement with one channel of EKG. The patient is intubated. Her face is turned away from the monitor and, therefore, it is difficult to see any motor activity on the video.
DESCRIPTION OF THE RECORD: The background EEG is diffusely slow. It is relatively continuous and primarily theta frequency activity from the left with frequency periodic activity or spike and slow wave complexes from the right hemisphere. The patient is noted to have some spontaneous movements of her head. In this section of the record, it is not possible to see the face twitch in part because of the limitation of visualization of that part of the anatomy.
HR: 84 BPM.
IMPRESSION: Abnormal EEG due to:
Frequent right hemispheric spike and slow and poly spike and slow wave complexes, which were reportedly associated with focal facial twitching.
CLINICAL CORRELATION: This tracing supports a partial or focal mechanism for seizures. In addition, a focal abnormality on the right should be considered.

Fig. 2. EEG Report of Example

1) 연구 데이터

본 연구에서는 미국 펜실베이니아에 위치한 Temple University Hospital(TUH)에서 2002년부터 2015년까지 아카이브 기록에서 수집된 TUH EEG Corpus(TUEG) v1.1.0을 사용하였다[22]. TUH는 IRB에 대한 승인을 받아 데이터셋을 수집하였으며, HIPAA 규정에 따라 식별자를 제거하여 데이터를 공개하였다. TUEG v1.1.0은 크게 EEG 전극 신호 데이터와 해당 EEG에 대한 Report로 구성되어있다. Fig. 2은 TUEG v1.1.0의 EEG Report의 예시로 서론, 기록에 대한 설명, 해석이 포함되어있다. TUEG v1.1.0의 EEG Report는 23,000개가 있으며, 본 연구에서는 전처리 이후 20,652개의 EEG Report를 사용하였다.

2) 데이터 전처리

Table 1은 종속 변수별 데이터 수를 나타낸 것이다. 8,027개는 Outpatient이지만, 2개의 데이터가 outpatient로 작성되었고 5,727개는 ICU이지만, 1개의 데이터가 iCU로 작성되었다. 이는 작성자의 실수로 판단하여 Outpatient와 ICU로 변경하였다. 또한, 결측 데이터와 중복된 데이터를 제거하여 21,260개를 사용하였다.

텍스트 데이터는 숫자, 알파벳, 작은따옴표('), 줄표(-)를 제외한 문자는 제거하였다. 또한, 알파벳의 대문자는 소문자로 모두 변경하였다. 변경된 텍스트에 대해 띄어쓰기 단위로

Table 1. Count of Data by Value of Target Variable

Target variable	Count
Outpatient	8,027
Inpatient	8,003
ICU	5,727
EMU	1,240
outpatient	2
iCU	1
Total	23,000

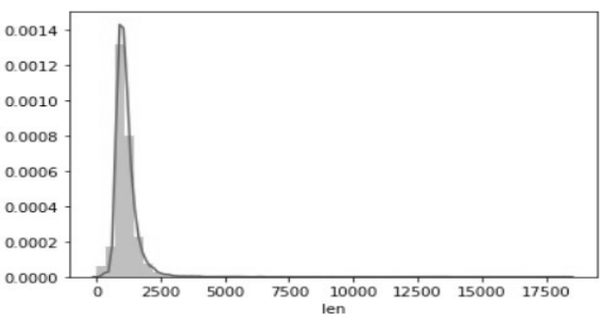

Fig. 3. Distribution of Length of Text

토큰화하였으며, NLTK 영어 불용어 목록을 사용하여 179개의 불용어를 제거하였다[23,24].

Fig. 3는 텍스트의 길이에 대한 분포를 나타낸 것으로 최소 0부터 최대 18,419였으며, 평균 1,137.98로 구성되었다. 길이가 짧은 텍스트는 이상치로 판단하여 제거하여 20,652개의 텍스트를 사용하였으며 각 텍스트의 토큰을 기반으로 단어 사전을 만들었다.

만들어진 사전을 통해 3가지 방법으로 벡터화를 진행하였다. 우선 사전의 단어가 존재 여부를 나타내는 원-핫 인코딩(One-hot encoding, One-hot)[25]을 사용해 벡터화시켰다. One-hot은 각 텍스트에 토큰의 유무를 나타내는 벡터로 표시되며 문장에서의 토큰의 순서를 고려하지 않는다.

두 번째로는 사전의 단어가 텍스트에 몇 번 존재하는지를 나타내는 Bag of Words(BoW)[26]를 사용해 벡터화시켰다. BoW는 각 텍스트에 토큰이 몇 번 나타나는지를 벡터로 표시되며 문장에서의 토큰의 순서를 고려하지 않는다.

마지막으로 특정 문서 내에서 해당 토큰이 얼마나 중요한 것인지에 따라 가중치를 주는 TF-IDF(Term Frequency-Inverse Document Frequency)[27]를 통해 벡터화하였다. TF-IDF는 각 문서에서 등장한 토큰들에 대해 전체 문서에서 해당 토큰을 고려하여 가중치를 준다.

3) EEG Report 분류 모델 구축

본 연구에서 ANN, CNN, BiLSTM, CNN-BiLSTM 알고리즘을 사용하여 EEG Report의 의무기록 유형을 분류하는 모델의 성능을 비교하였다. 모델은 학습 데이터 중 20%를 평가 데이터로 사용하여 최적의 epoch 값을 구하였으며, batch는 128로 학습하였다.

Table 2. ANN Model Structure

Layer	Output Shape	Param #
dense (Dense)	(None, 32)	488448
dense_1 (Dense)	(None, 32)	1056
dense_2 (Dense)	(None, 4)	132

Table 3. CNN Model Structure

Layer	Output Shape	Param #
embedding 1 (Embedding)	(None, 1500, 128)	1953664
conv1d (Conv1D)	(None, 1494, 32)	28704
max_pooling1d (MaxPooling1D)	(None, 298, 32)	0
conv1d_1 (Conv1D)	(None, 292, 32)	7200
global_max_pooling1d (GlobalMaxPooling1D)	(None, 32)	0
dense_4 (Dense)	(None, 4)	132

Table 4. BiLSTM Model Structure

Layer	Output Shape	Param #
embedding (Embedding)	(None, None, 32)	488416
bidirectional (Bidirectional)	(None, 64)	16640
dense_3 (Dense)	(None, 4)	260

Table 2은 ANN 모델의 구조를 나타낸다. 입력 레이어와 히든레이어는 32개의 노드와 활성화 함수로 Relu를 가진다. 출력 레이어는 4개의 노드와 Softmax를 활성화 함수로 가진다. ANN 모델의 최적화 함수는 RMSprop를 사용하였다.

Table 3는 CNN 모델의 구조를 나타낸다. 임베딩 레이어의 입력 길이는 1,500이며 Convolution 레이어는 32개의 필터와 padding의 크기는 7, 활성화 함수로 Relu를 사용한다. Maxpooling 레이어는 padding의 크기를 5로 하였다.

Table 4는 BiLSTM 모델의 구조를 나타낸다. 임베딩 레이어와 BiLSTM 레이어의 노드는 32이고 출력 레이어는 4개의 노드를 사용하며 활성화 함수로 softmax를 사용한다. BiLSTM 모델의 최적화 함수로 RMSprop를 사용하였다.

Table 5은 CNN-BiLSTM 모델의 구조를 나타낸다. 이는 CNN 모델과 BiLSTM 모델을 결합한 것으로 Convolution 레이어 2개와 Maxpooling 레이어 1개, BiLSTM 레이어로 구축하였다. 각 레이어의 노드와 활성화 함수는 본 연구에서 사용한 CNN, BiLSTM의 레이어와 같으며 최적화 함수도 동일하다.

Table 5. CNN-BiLSTM Model Structure

Layer	Output Shape	Param #
embedding 1 (Embedding)	(None, 1500, 128)	1953664
conv1d (Conv1D)	(None, 1494, 32)	28704
max_pooling1d (MaxPooling1D)	(None, 298, 32)	0
conv1d_1 (Conv1D)	(None, 292, 32)	7200
bidirectional_1 (Bidirectional)	(None, 32)	0
dense_4 (Dense)	(None, 4)	132

4) EEG Report 분류 모델 평가

본 연구에서 구축된 EEG Report 분류 모델의 학습에 사용되지 않은 나머지 20%의 데이터로 평가하였다. 평가에 사용된 데이터는 학습에 사용되지 않았다. 평가 기준은 정확도(Accuracy)와 매크로 평균과 마이크로 평균을 사용한 정밀도(Precision), 재현율(Recall), F1-score로 성능을 평가하였다.

이진 분류에서 실제 정답과 분류 결과에 따라 참 긍정(True Positive, TP), 거짓 긍정(False Positive, FP), 거짓 부정 (False Negative), 참 부정(True Negative)으로 구분되어 진다. 정확도(1)는 전체 데이터 중 TP와 FN으로 예측한 것의 비율로 수치가 높을수록 정확히 분류한 데이터가 많음을 의미한다. 정밀도(2)는 모델이 참으로 분류한 것 중 실제 정답이 참인 것의 비율로 TP를 FP와 TP의 합으로 나눈 값이다. 재현율(3)은 실제값이 참인 데이터를 모델이 얼마나 참으로 예측하였는지를 나타내며 TP를 TP와 FN의 합으로 나눈 값이다. F1-score(4)는 정밀도와 재현율의 조화평균을 나타내며 데이터 불균형을 이루는 모델을 평가에 주로 사용된다. 이를 식으로 나타내면 다음과 같다.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\,score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \quad (4)$$

본 연구는 다중 분류를 하는 모델임으로 정밀도, 재현율, F1-score를 각 클래스의 샘플의 빈도를 고려한 마이크로 평균 방법과 빈도를 고려하지 않는 매크로 평균 방법[28]을 사용하였다.

4. 실험 결과

4.1 데이터 전처리 결과

Table 6은 데이터 전처리 이후 EEG Report의 의무 기록 유형별 데이터 수를 나타낸 표이다. Outpatient이 7,749개로 가장 많았으며, EMU가 798개로 가장 적었다. Fig. 4는 데이터 전처리 이후 텍스트의 길이 분포를 나타낸 것으로 텍스트의 길이는 최소 570에서 최대 18,419였으며, 평균 1166.23였다.

4.2 EEG Report 분류 모델 결과 비교

Table 7은 EEG Report 분류 모델의 성능을 비교한 결과표이다. 딥러닝 기반의 ANN, BiLSTM, CNN, CNN-BiLSTM을 사용하였으며 One-Hot, BoW, TF-IDF로 토큰을 벡터화하였다. One-Hot을 사용하여 벡터화한 ANN 모델에서 매크로 평균을 이용한 재현율을 제외한 모든 성능이 가장 높았다.

One-Hot을 사용한 ANN 모델이 71%로 가장 높았다. 그 다음으로는 BoW를 사용한 ANN이 70%, TF-IDF를 사용한 ANN이 67% 순으로 나타났다. CNN과 BiLSTM을 각각 사용하여 예측한 정확도보다 둘을 결합한 CNN-BiLSTM 모델의 정확도가 모든 벡터화 방식에서 약 4%에서 최대 14%까지 높았다. BoW와 TF-IDF로 벡터화한 데이터로 학습된 Bi-LSTM의 정밀도를 제외한 모든 평가에서 마이크로 평균 방법이 높거나 같았다.

Fig. 5은 성능이 가장 높았던 One-Hot을 사용한 ANN 모델의 혼동 행렬을 나타낸 것으로 EMU의 경우 대체로

Table 6. Count of Target Variables with Data Preprocessed

Target variable	Count
Outpatient	7,749
Inpatient	7,191
ICU	4,914
EMU	798
Total	20.652

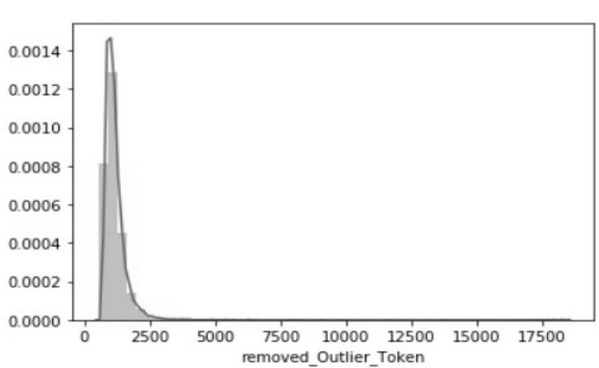

Fig. 4. Distribution of Reomved Outlier in Text

Table 7. Deep Learning based Model Result(%)

Model	Vectorization	Accuracy	Precision		Recall		F1-score	
			micro	macro	micro	macro	micro	macro
ANN	One-Hot	71	71	69	71	60	70	62
	BoW	67	69	63	67	58	66	58
	TF-IDF	70	69	62	70	62	68	61
CNN	One-Hot	53	61	55	53	46	44	41
	BoW	58	59	55	58	51	54	49
	TF-IDF	62	59	59	62	52	57	49
BiLSTM	One-Hot	56	54	43	56	45	54	43
	BoW	50	53	62	50	39	45	37
	TF-IDF	55	57	62	55	44	51	44
CNN-BiLSTM	One-Hot	62	62	58	62	52	61	52
	BoW	64	65	63	64	53	63	53
	TF-IDF	59	57	45	59	47	57	45

Actual / Predict	Outpatient	Inpatient	ICU	EMU	Total
Outpatient	1276	194	4	2	1476
Inpatient	228	968	212	14	1422
ICU	32	384	880	19	1315
EMU	39	142	38	52	271
Total	1575	1688	1134	87	4484

Fig. 5. Confusion Matrix of ANN Model using One-Hot

Inpatient로 예측한 경향을 보인다. 또한, 다른 카테고리에서도 Inpatient로 잘못 예측하는 경우가 가장 많았다. 데이터는 Outpatient이 Inpatient보다 많았지만, 모델이 예측한 결과는 Inpatient로 예측한 데이터가 약 100개 많았다.

5. 결 론

본 연구는 여러 보건의료기관의 텍스트로 된 데이터를 유형, 서식 등에 따라 분류하여 관리 및 활용을 위한 기저 연구로써 기반을 구축하고자, EEG Report의 의무기록 유형을 분류하였다. 이를 위해 딥러닝 기반 알고리즘을 사용하여 구축된 분류 모델의 예측 정확도를 산출하여 비교하였다. 비교 결과 One-Hot으로 벡터화하여 학습한 ANN 모델의 성능이 71%로 가장 높은 것을 확인하였다.

CNN-BiLSTM 모델은 텍스트의 순서를 고려하는 이점이 있는 모델이다. 하지만 본 연구에서는 벡터화에서 단어의 의미와 문장에서 위치를 고려하지 않아 CNN-BiLSTM 모델이 64%의 정확도로 ANN 모델보다 낮은 성능을 보였다. 이에 향후 Word2Vec[29]나 Doc2Vec[30]을 사용하여 단어의 의미와 순서를 반영한다면 더 높은 예측 결과를 도출할 수 있을 것이다.

본 연구에서 사용한 데이터의 경우 데이터 불균형이 나타나고 있다. 이를 극복하기 위해 오버샘플링(Oversampling)의 경우 의료 기록의 특성상 실제로 존재할 수 없는 데이터 생성될 수 있기에 적합하지 않다고 판단하여 사용하지 않았다. 하지만 최근 실제와 더욱 유사한 데이터로 oversampling 하는 Generative Adversarial Network 기반 알고리즘[31]을 활용하여 오버샘플링을 적용한다면 더 높은 예측 결과를 도출할 수 있을 것이다.

보건의료 데이터는 아직 개방된 데이터가 적어 본 연구에서도 한 기관의 데이터를 사용하였다. 이는 모델의 일반화 가능성을 낮추고 과적합되기 쉽게한다. 향후 다기관 연구를 통해 일반화된 모델을 제시할 것이다.

References

[1] D. Charles, M. Gabriel, and M. F. Furukawa, “Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2014,” Washington, DC: Office of the National Coordinator for Health Information Technology, No.23, pp.1-10, 2015.

[2] Korea University Anam Hospital, EEG, curious about that [Internet], http://anam.kumc.or.kr/info/examInfoView.do?BNO=13&cPage=1&BOARD_ID=S003.

[3] Y. Gao, B. Gao, Q Chen, J. Liu, and Y. Zhang, “Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification,” *Frontiers in Neurology*, Vol.11, 2020.

[4] S. Raghu, N Sriraam, Y Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, Vol.124, pp.202-212, 2020.

[5] Y. Wang et al., "A clinical text classification paradigm using weak supervision and deep representation," *BMC Medical Informatics and Decision Making*, Vol.19, No.1, pp.1-13, 2019.

[6] T. R. Goodwin and S. M. Harabagiu, "Deep learning from EEG reports for inferring underspecified information," in *AMIA Joint Summits on Translational Science Proceedings*, pp.112-121, 2017.

[7] J. Kim, D. Seo, H. Kim, and P. Kang, "Facebook spam post filtering based on instagram-based transfer learning and meta information of posts," *Journal of the Korean Institute of Industrial Engineers*, Vol.43, No.3, pp.192-202, 2017.

[8] D. Kim and M. W. Koo, "Categorization of Korean news articles based on convolutional neural network using Doc2Vec and Word2Vec," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.44, No.7, pp.742-747, 2017.

[9] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, Vol.337, pp.325-338, 2019.

[10] Y. Chen, X. Zhang, and T. Li, "Medical records classification model based on text-image dual-mode fusion," in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, pp.432-436, 2021.

[11] L. Qing, W. Linhong, and D. Xuehai, "A novel neural Network-Based method for medical text classification," *Future Internet*, Vol.11, No.12, pp.255, 2019.

[12] I. Y. Jung, "Startup trends using domestic and foreign health and medical big data" [Internet], https://repository.hira.or.kr/handle/2019.oak/1473.

[13] M. D. Wilkinson, M. Dumontier, and I. J. Aalbersberg, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, Vol.3, No.1, pp.1-9, 2016.

[14] National Institutes of Health, STRIDES Initiative [Internet], https://datascience.nih.gov/strides.

[15] S. A. Kim, "Meanings and tasks of the three revised bills which ease regulations on the use of personal information," *Jouranl of Information and Security*, Vol.20, No.2, pp.59-68, 2020.

[16] K. J. Kim, B. Y. Jang, J. Y. Jung, and O. W. Park, "The coming of the 4th industrial revolution and the HRD Issues for nurses - prospects and challenges," *Korean Journal of Resources Development*, Vol.21, No.3, pp.137-159, 2018.

[17] Health Insurance Review & Assessment Service, Current status of introduction and development of electronic medical records in Korea [Internet], https://www.hira.or.kr/bbsDummy.do?pgmid=HIRAA030096000000&brdScnBltNo=4&brdBltNo=623.

[18] Korea Health Information Service, 2020 Health and Medical Informatization Survey Results Report [Internet], https://www.k-his.or.kr/board.es?mid=a10306040000&bid=0005&act=view&list_no=283&tag=&nPage=1.

[19] J. Y. Lee, Y. Kim, and G. Kim, "A study on the analysis and methods to improve the medical records management in a large university hospital," *Journal of Korean Society of Archives and Records Management*, Vol.13, No.1, pp.107-134, 2013.

[20] E. M. Lee, M. Kim, and J. Hee, "A study on the current status and tasks of medical records management: Focused on applying the KS X ISO 15489 to the Y hospital," *Journal of the Korean Society for information Management*, Vol.29, No.3, pp.257-285, 2012.

[21] American Clinical Neurophysiology Society, "Guideline 7: guidelines for writing EEG reports," *American Journal of Electroneurodiagnostic Technology*, Vol.46, No.3, pp.231-235, 2006.

[22] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers in Neuroscience*, *Data Report*, Vol. 10, No.196, 2016.

[23] S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp.69-72, 2006.

[24] S. Bird, E. Klein, and E. Loper, "Natural language processing with Python: Analyzing text with the natural language toolkit," O'Reilly Media, Inc., 2009.

[25] Google Developers, Step 3: Prepare Your Data [Internet], https://developers.google.com/machine-learning/guides/text-classification/step-3.

[26] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE Transactions on Fuzzy Systems*, Vol.26, No.2, pp.794-804, 2018.

[27] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, Vol.38, No.3, pp.2758-2765, 2011.

[28] M. Bihis and S. Roychowdhury, "A generalized flow for multi-class and binary classification tasks: An azure ML approach," *2015 IEEE International Conference on Big Data (Big Data)*, pp.1728-1737, 2015.

[29] S. S. Lim, H. Lee, and Y. M. Yoon, "Prediction of new drug-side effect relation using Word2Vec model-based word similarity," *Journal of Korean Institute of Information Technology*, Vol.18, No.11, pp.25-33, 2020.

[30] A. M. Shah, X. Yan, and A. Qayyum, "Social network analysis of an online smoking cessation community to identify users' smoking status," *Healthcare Informatics Research*, Vol.27, No.2, pp.116-126, 2021.

[31] J. Xu, X. Ren, J. Lin, and X. Sun, "DP-GAN: Diversity-promoting generative adversarial network for generating informative and diversified text," *arXiv preprint arXiv: 1802.01345*, 2018.

오 경 수

https://orcid.org/0000-0002-3177-8793
e-mail : ba8745@gachon.ac.kr
2017년 ~ 현 재 가천대학교 컴퓨터공학과 학사과정
관심분야 : 인공지능, 자연어처리, 데이터 분석

강 민

https://orcid.org/0000-0002-0548-170X
e-mail : km8846@gachon.ac.kr
2021년 가천대학교 컴퓨터공학과(학사)
2021년 ~ 현 재 가천대학교 IT융합공학과 석사과정
관심분야 : 딥러닝, 머신러닝, 데이터분석

강 석 환

https://orcid.org/0000-0002-8076-0290
e-mail : shkang@gachon.ac.kr
1987년 ~ 1988년 대우통신㈜ 중앙연구소 연구원
1989년 ~ 2014년 Ericsson-LG(LG정보통신) 중앙연구소 실장
2014년 ~ 2017년 KJ컴텍㈜ 연구소장
2017년 ~ 현 재 가천대학교 컴퓨터공학과 부교수
관심분야 : 모바일 시스템(LTE, 5G), 인공지능, U-Healthcare, 프로젝트 관리(Aglie)

이 영 호

https://orcid.org/0000-0003-0720-0569
e-mail : lyh@gachon.ac.kr
2007년 아주대학교 의료정보학(박사)
2008년 ~ 현 재 대한의료정보학회 기획이사
2010년 ~ 현 재 산업통상자원부 국가표준화위원
2014년 ~ 2015년 Virginia Tech Research Fellow
2016년 ~ 현 재 보건복지부/보건산업진흥원 PM제도위원
2017년 ~ 현 재 가천대학교 전산정보원장
2002년 ~ 현 재 가천대학교 컴퓨터공학과 교수
관심분야 : 의료 빅데이터 분석, 인공지능, 모바일 헬스케어