〈딥러닝 기말 Final Project〉

4팀 고준서 권혜현 김하림 전규리

Ⅰ. 서론

ⅰ. 문제 정의

① 문제 정의와 필요성

심전도 데이터는 검사가 용이하고 시간이 많이 걸리지 않기 때문에 간단한 검사만으로 쉽게 측정할 수 있다. 그러나 심전도 데이터를 통해 특정 질병을 정확하게 진단하는 것은 많은 시간과 전문지식을 필요로 한다. 그러므로 저희 팀은 머신러닝/딥러닝 방법론을 활용하여 심전도 데이터를 통해 질병을 예측하는 모델을 만들어 전문가들의 질병 진단에 보조적인 역할을 수행하고자 한다.

② Repolarization Abnormalities

심전도 데이터를 활용하여 분석 및 진단할 수 있는 질병은 부정맥, 조기 재분극 현상, 빈혈, 심실 비대증 등이 있으나 특정 질병을 정확하게 진단하는 것은 고도의 의학적 전문지식을 요구한다. 따라서 이번 프로젝트에서는 주어진 데이터셋에 포함된 label들을 최대한 활용하여 진단할 수 있는 Repolarization Abnormality를 분석할 질병으로 선정하게 되었다. 재분극(repolarization)이란 심장 세포가 탈분극의 과정을 거친 후 안정되며 극성이 회복되는 상태를 의미한다. 이번 ECG데이터 분석을 통해 재분극에서 비정상적인 패턴을 보이는 것을 의미하는 repolarization abnormalities를 검출해내고자 한다.

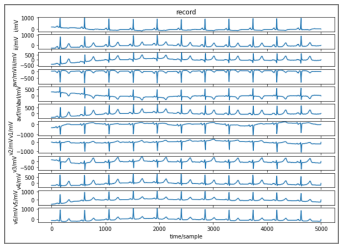
ⅱ. 관련 연구

기존의 많은 연구들은 심전도를 machine learning의 기법들로 분석하고자 하였다[1]. 그러나 심전도의 특성상 직접 작성해야 하는 feature가 너무 많다는 단점이 존재한다[2]. 또한, ECG신호를 그대로 머신러닝에 사용할 경우 차원의 저주가 발생할 수 있다. 다른 연구[3]를 참조할 때, 오토인코더를 통해 전처리를 해준 데이터를 활용하여 성능을 향상시켰다는 결과를 확인할 수 있다. 이 방법을 적용하여 오토인코더로 신호데이터인 ECG데이터를 처리하여 사용할 경우, 향상된 성능을 기대해볼 수 있다. 그 외에도 LSTM[4], Xgboost[5]를 기반으로 한 ECG데이터 분석이 기존에 수행되어 왔다. 이번 연구에서는 이러한 기존의 방법론을 결합하여 ECG data로부터 특정 질병을 예측해 내는 연구를 수행할 예정이다.

Ⅱ. 본론

ⅰ. 데이터 분석

사용할 데이터는 Lobachevsky University ECG Database의 데이터이다. ECG는 심전도 데이터로 심장의 전기적 활동을 해석한 데이터이다[6]. 데이터를 보면 12개의 방식에 따른 신호를 10초 동안 초당 500개로 샘플링한 데이터이다. 12개의 방식은 표준 유도인 i, ii, iii, 사지유도인 avr, avl, avf 마지막으로 흉부유도인 v1 ~ v6까지이다. 환자의 수는 총 200명이다. 이를 시각화하면 아래의 그림과 같다.



[Figure1 - ECG Data Visualization]

해당 데이터를 딥러닝과 데이터분석에 사용하기 위해 넘파이 형식으로 불러오면 [환자수 x 샘플링한 신호길이 x 신호종류 = 200 x 5000 x 12] 가 된다. ECG데이터의 경우 연속적인 데이터이기 때문에, 한 줄의 신호에서 비슷하거나 거의

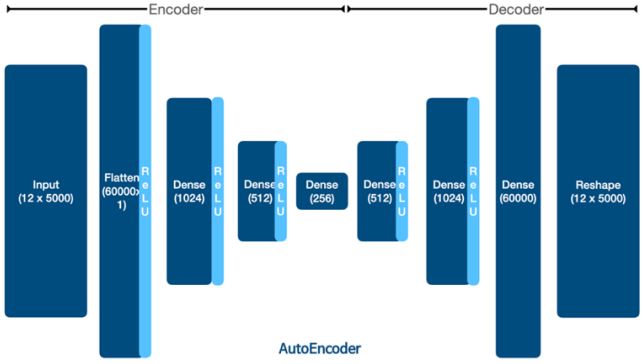
차이가 없는 데이터들이 많이 있어 차원의 저주 문제가 발생할 수 있다. 차원의 저주란 학습데이터의 수가 차원의 수보다 작아져서 발생하는 문제로, 차원이 커짐에 따라서 불필요하게 남는 공간이 발생하게 되고, 학습에 방해가 되는 경우가 발생하는 것이다. 이런 차원의 저주로 인해서 공간 대비 데이터의 양이 적을 경우, 오버피팅이 발생할 가능성이 높아지기 때문에, 아래의 전처리 과정에서 차원의 저주 문제를 방지하기 위해 택한 방식에 대해 서술할 예정이다.

또한, 모든 실험을 위해서 데이터셋을 공통적으로 분할해주고 성능을 평가하였다.

Train Data	153 x 12 x 5000
Test Data	20 x 12 x 5000
Validation Data	27 x 12 x 5000

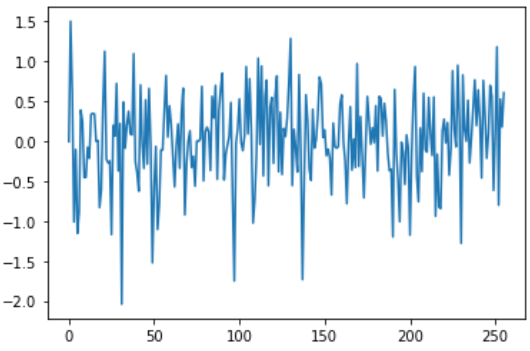
ii . 데이터 전처리

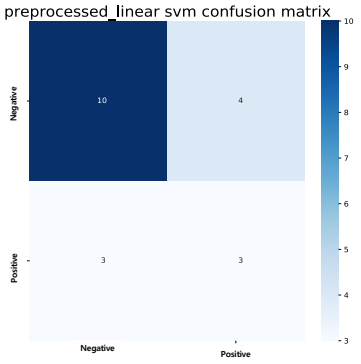
① AutoEncoder를 활용하여 차원 축소

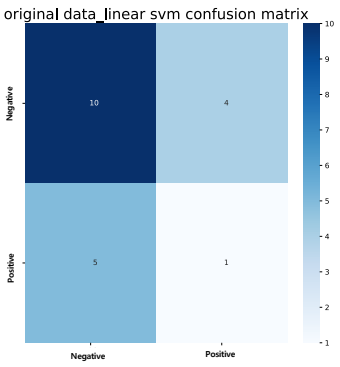


사용한 오토인코더 구조

차원의 저주를 극복하기 위해서 AutoEncoder를 사용해주었다. Autoencoder는 입력을 출력으로 복사하는 신경망이다. 그러나 복사하기까지 여러가지 제약을 줌으로써 복잡한 신경망으로 만든다. 인코더 부분에서는 데이터의 차원을 축소하고 디코더에서는 이 축소한 차원을 복구하는 역할을 한다. 이 축소하고 복구하는 네트워크의 loss를 줄여나가도록 학습을 수행한 후, 인코더 부분만을 활용하여 차원을 축소하도록 했다. 이런 과정을 통해서 한 환자 당 [5000 x 12]의 데이터를 [256 x 1]로 차원을 축소할 수 있었다. 더 적은 차원으로 줄이는 것도 시도해보았지만, 더 적은 차원으로 인코딩할 경우 디코딩 시 인코딩된 벡터를 적절하게 해석하여 원래대로 복사하는 것이 어려워 성능이 안좋았고, 더 많은 차원의 경우 차원의 저주를 극복하는데 어려움이 있었다. 그러므로 이진수인 256차원을 선택하게 되었다. 간단하게 SVM을 통해서 오토인코더의 성능을 확인해보았다. 딥러닝보다 머신러닝으로 성능을 확인해본 이유는, 간단하게 테스팅해볼 수 있고, random seed를 사람이 설정하지 않아도 성능이 일정하게 나오기 때문이었다. 왼쪽은 환자 하나의 데이터(12x5000)의 데이터에 대해서 오토인코더를 거친 결과이다.







두 개 모두 gridsearchCV를 통해서 여러 파라미터 중 최선의 것을 선택해 준 것이다. 보다시피 인코딩을 한 오른쪽 결과가 왼쪽 결과보다 높은 성능을 보이는 것을 알 수 있다. 특히 의료관련 데이터에서 중요한 positive를 positive로 예측하는 sensitivity의 정확도가 높아지는 것을 확인할 수 있다. 오토인코더를 활용한 데이터에서 더 높은 성능을 확인할 수 있음을 통해 오토인코더를 활용한 성능 향상을 증명하였다.

② Label 설정

라벨의 경우, 서론에서 언급한 바와 같이 Non-specific repolarization abnormalites를 검출해낼 것이다. wfdb를 통해 comments를 뽑아내어 그 중 해당 질병명이 포함되어 있으면 1, 없으면 0으로 설정해주었다.

iii. 사용한 모델과 그 이유

Baseline ① - Machine Learning Methods

1) Logistic Regression

로지스틱 회귀는 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 통계 기법으로, 종속 변수가 범주형 데이터를 대상으로 하며 입력데이터에 대한 결과가 특정 분류로 나뉜다는 특징을 가진다. 데이터마이닝, 통신 등의 분야뿐만 아니라 의학 연구에서도 사용되는데, 특히 질병의 위험인자나 요인을 밝히는 연구 등에 쓰인다는 점에서 착안하여 이를 시도해보기로 하였다[7].

2) XGBoost Classifier

XGBoost는 Extreme Gradient Boost를 줄여 이르는 말로, 의사결정 나무 기반의 앙상블 머신러닝 알고리즘이다. 많은 캐글 경진대회 우승자들이 이 알고리즘을 사용하여 우승한 만큼, 빠른 속도에 좋은 성능을 보인다. 병렬처리를 사용하는 모델의 특성상 학습과 분류가 빠른 편이고, 유연성이 좋은 것으로 알려져 있다. 이러한 장점들과 더불어 분류 문제에서 뛰어난 성능을 발휘하는 것으로 알려져 있기 때문에 XGBoost를 이용하게 되었다[8].

Baseline ② - Simple One Layer Neural Network

Simple One Layer Neural Network은 딥러닝의 가장 기본적인 신경망이다. 이진분류, 다중분류에 모두 적용가능한데, 다른 딥러닝보다 다소 단순한 구조로 되어 있어 간편하게 사용할 수 있는 장점이 있다.

Baseline ③ - LSTM

LSTM의 경우, 단방향 노드를 가지고 있는 네트워크이다. 성능을 높이기 위해 사용한 베이스 라인 중 하나이다. LSTM은 방향 노드를 가지기 때문에 다음 레이어로 gradient가 전달되어 Vanishing gradient 문제가 해소된다. 음성이나 영상, 자연어 등의 연속성 있는 데이터에 많이 활용된다[9].

iv. 각 모델의 성능을 높이기 위한 노력

① GridSearchCV

하이퍼 파라미터를 튜닝하는 방법 중 하나로, 여러가지 파라미터들을 지정하고 코드를 돌리면 모델에 가장 적합한 파라미터의 조합을 찾아 모델을 피팅하고 예측할 수 있게 한다. 파라미터 튜닝은 모델의 성능을 높이기 위해 중요하므로 그리드 서치를 통해 모든 조합의 경우 중 가장 높은 성능을 내도록 한다[10].

② AutoEncoder

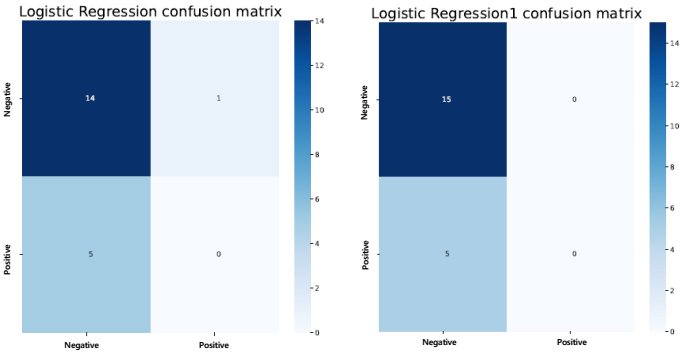
위에서 자세하게 언급하였다. 해당 방법을 통해 성능 향상을 보여주었다.

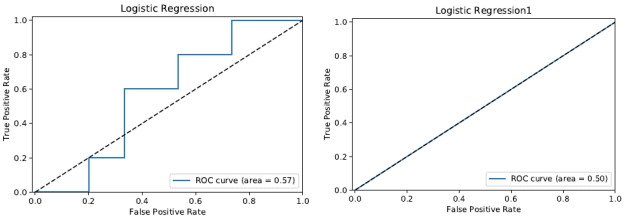
Ⅲ. 결론

ⅰ. 각 모델의 성능 리포트

1) Baseline ① - Machine Learning Methods

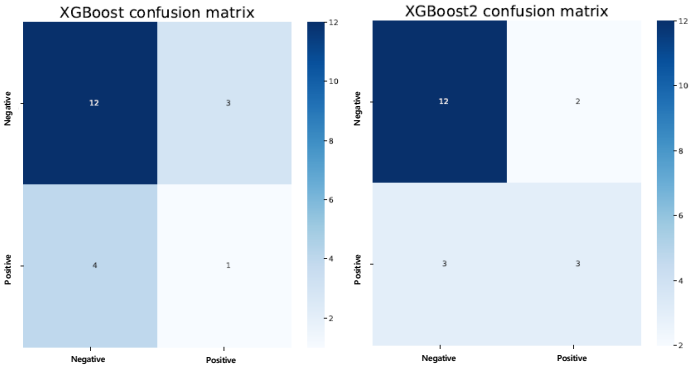
(1) Logistic Regression

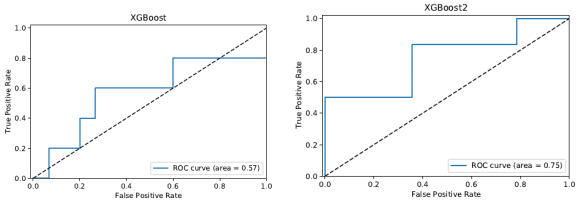




다른 베이스라인 모델들과 달리 AutoEncoder를 사용했을 경우, 분류가 아예 되지 않았고 역으로 정확도가 떨어지는 결과를 보였다. 복잡한 비선형 데이터 처리에 약세를 보인다고 알려진 LR 모델의 특성 상, AutoEncoder로 전처리된 벡터가 기존 데이터에 비해 선형성을 잃어버리도록 인코딩 되어 전처리 후의 데이터가 LR에 적합하지 않아 이런 결과가 나왔을 것이라고 생각한다. 실험 결과 역시 이를 뒷받침 해주는데, Logistic Regression으로 표시된 원본 데이터를 처리한 결과가 Logistic Regression1으로 표시된 인코딩된 벡터를 처리한 결과보다 좋지 못함을 확인할 수 있다.

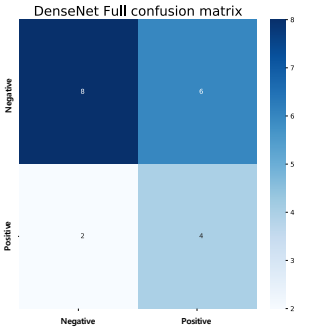
(2) XGBoost Classifier

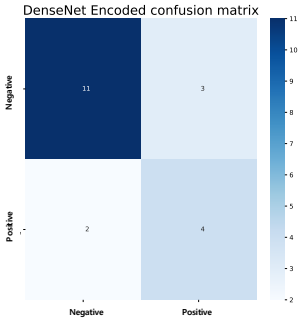


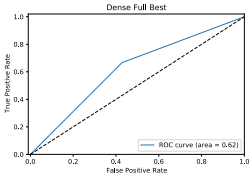


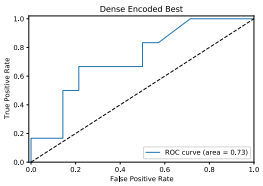
GridSearchCV와 Stratified K-fold를 이용해서 성능을 높이는 시도를 해봤다. 이외에도 Early stopping과 Evaluation set 지정을 해보면서 모델 학습을 진행하였다. 후술될 모델들과 마찬가지로 XGBoost 또한 AutoEncoder를 통해 전처리된 데이터를 사용했을 때 보다 높은 성능을 보였고, 정확도가 0.75까지 상승하는 것을 확인하였다.

2) Baseline ② - Simple One Layer Neural Network



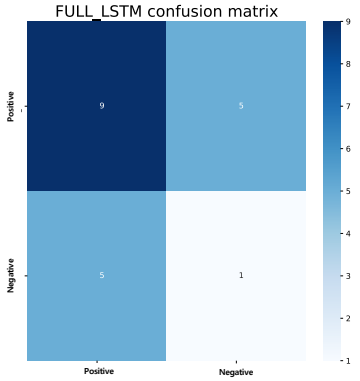


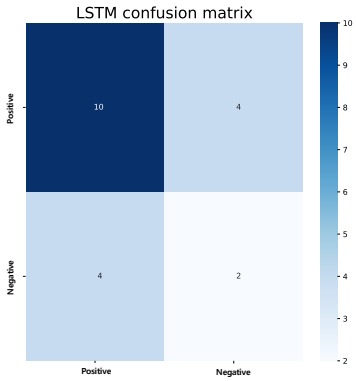




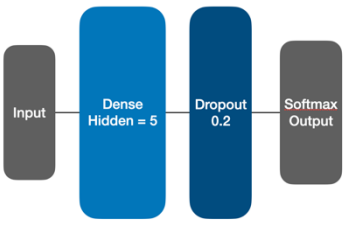
해당 베이스라인 모델도 높은 성능을 보인다. 인코딩한 벡터로 할 경우, 75%의 성능을 확인할 수 있다. XGboost Classifier과 비교했을 때, accuracy는 동일하지만, sensitivity에서 강세를 보이는 것을 확인할 수 있다. 정확한 수치로는 XGBoost가 0.5, NN이 0.67을 보인다. 그러므로 의료데이터인 현재의 데이터를 분류하는데 해당 모델이 더 높은 성능을 보인다고 간주할 수 있다. 그러나, AUROC커브를 보면, XGboost의 성능이 더 높은데, 헷갈리는 데이터에 대해서 확률적으로 더 정확하게 맞춘다는 것을 유추해볼 수 있다. 만약 의사 선생님들과 같이 협업하여 threshold 등에 대해서 조절해줄 수 있으면 XGboost가 더 좋은 성능을 보여줄 수 있지만 현재의 상황에서 가장 좋은 모델은 동일 정확도에서 sensitivity가 높은 이 모델이라 할 수 있다.

3) Baseline ③ - LSTM





LSTM의 결과 역시 AutoEncoder의 성능이 준수하다. 그러나, dense-layer를 활용한 단층 신경망보다는 성능이 떨어진 것을 볼 수 있다. 오토인코더에서 성능이 떨어진 이유는 오토인코더를 사용하면서 연속성을 잃어버렸거나 벡터 간의 연속성보다 그 벡터 하나의 값이 더 중요했기 때문일 것으로 판단한다. 두개 모두 성능이 저하된 이유는 아무래도 LSTM이 단순 dense layer보다 파라미터가 많은데, 그에 비해서 데이터셋의 양이 크지 않아 온전하게 학습이 진행되지 않은 것으로 추측된다.



결론적으로 Simple One Layer Neural Net을 사용했을 때, sensitivity와 accuracy 두가지 측면에서 모두 정확한 모습을 보였다. 왼쪽의 모델은 우리가 사용한 최종 모델이다. 우리는 75%의 성능을 얻을 수 있었고, 0.67의 sensitivity를 확보하였다. 이 결과만으로 우리는 해당 질병을 검출하는 데 있어서 의사를 완벽하게 보조하기는 어렵다. 그러나 추가적인 엔지니어링을 통해 성능을 향상시킨다면, 충분히 보조적인 역할을 수행할 수 있을 것이다.

Ⅳ. REFERENCE

[1] Lyon Aurore, Ariga Rina, Mincholé Ana, et al. (2018). Distinct ECG Phenotypes Identified in Hypertrophic Cardiomyopathy Using Machine Learning Associate With Arrhythmic Risk Markers. Frontiers in Physiology. 2018(9).

[2] B Pyakillya et al. (2017). J. Phys.: Conf. Ser. 913 012004

[3] https://dl.acm.org/doi/10.1145/1390156.1390294

[4] Saeed Saadatnejad, Mohammadhosein Oveisi, and Matin Hashemi. (2019). LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices.

[5] Li, Hao et al. (2019). 'Research on Massive ECG Data in XGBoost'. 1 Jan. 2019 : 1161 - 1169.

[6] https://ko.wikipedia.org/wiki/심전도

[7] https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/

[8] https://xgboost.readthedocs.io/en/latest/python/python_api.html

[9] https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/

[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html