2022년 8월

석사학위논문

음성 및 텍스트 데이터로부터 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단

조 선 대 학 교 대 학 원

전자공학과

조 아 현

음성 및 텍스트 데이터로부터 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단

Depression Diagnosis Based on 4-stream of Bi-LSTM and CNN from audio and text data

2022년 8월 26일

조 선 대 학 교 대 학 원

전자공학과

조 아 현

음성 및 텍스트 데이터로부터 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단

지도교수 곽 근 창

이 논문을 공학석사학위신청 논문으로 제출함.

2022년 4월

조 선 대 학 교 대 학 원

전자공학과

조 아 현

조아현의 공학석사학위논문을 인준함

위원장 조 선 대 학 교 교수 염 홍 기 인

위 원 조 선 대 학 교 교수 신 주 현 인

위 원 조 선 대 학 교 교수 곽 근 창 인

2022년 5월

조선대학교 대학원

목 차

표 목 차

도 목 차

ABSTRACT

Depression Diagnosis Based on 4-stream of Bi-LSTM and CNN from audio and text data

Jo, A-Hyeon
Advisor : Prof. Kwak, Keun Chang, Ph. D.
Dept. of Electronic Engineering,
Graduate School of Chosun University

Depression is a disease that causes changes in emotions, thoughts, and behavior, and it falls into severe depression by leaving the disease unattended, which can lead to various problems. Currently, the diagnosis of depression is based on inconsistent subjective opinions of clinicians. In addition, it is done in a way that patients directly tell their conditions through questionnaires. However, these methods have disadvantages in that they are limited and difficult to diagnose accurately because objective opinions are excluded. Therefore, we need a system that objectively and accurately diagnoses depression by considering various data and features. Recently, interest in automated system design has been increasing in the field of the affective computing community and artificial intelligence to efficiently diagnose depression. In particular, based on deep learning technology, research on depression diagnosis is being actively conducted

using multi-mode, which can utilize much information by fusion of multiple data rather than a single-mode using one data.

In this paper, we propose a depression diagnosis model based on a 4-stream of Bidirectional Long Short Term Memory(Bi-LSTM) and Convolutional Neural Network(CNN) from audio and text data. One-dimensional features of speech signals were extracted using Mel Frequency Cepstral Coefficient(MFCC) and Gammatone Cepstral Coefficients(GTCC). Also, two-dimensional features were extracted from the Bark spectrogram, ERB spectrogram, and Log-Mel spectrogram based on time-frequency transformation. These features were applied to Bi-LSTM, and CNN-based transport learning models such as VGGish, YAMNet, and OpenL3. For text data, word-encoding was used to map the text into sequences with numeric indices. And word embedding concepts were used to represent all words as dense numeric vectors. These texts were applied to Bi-LSTM, and the CNN model based on n-gram. Finally, the softmax values of the four deep learning models were ensembled using the late score fusion method to diagnose depression based on the 4-stream. The data used in the experiment is the Extended Distress Analysis Interview Corpus Wizard of Oz (EDAIC-WOZ) depression database designed to help diagnose people's psychological distress state. The noise was removed from speech data, and unnecessary words were cleaned up from text data through preprocessing to improve data quality. Also, extending the depression data solved the class imbalance problem. The experimental results showed that the performance was improved from min 1.22% to max 2.44% more when using the 4-stream model than the single model. In addition, the proposed model was more competitive than the 2-stream-based model of the previous study under the same data conditions. Likewise, the proposed model showed good performance when evaluating the performance using the EDAIC-WOZ database. These results proved that the proposed model is effective in diagnosing depression.

제1장 서론

제1절 연구 배경 및 목적

코로나19 바이러스의 확산으로 세계 인구의 절반 이상을 격리하고 많은 국가에서 보건 비상사태를 선포하는 등 전례가 없는 보건 조치가 시행되고 있다. 이로 인해 정상적인 일상생활이 어려워져 불안과 우울 증상을 호소하는 사람들이 많아졌고, 전 세계의 우울증 발생률이 2배 이상 증가하였다[1-2]. 이는 사람들에게 매우 부정적인 영향을 미치고 있다. 우울증은 일반적으로 집중력, 사고의 과정, 의욕, 흥미, 수면 등 전반적으로 정신적인 기능이 지속적으로 저하되어 일상생활에 악영향을 미치는 상태를 의미한다. 정신의학에서 말하는 우울한 상태란 이러한 과정에서 일시적으로 기분이 저하되는 상태를 뜻하는 것이 아니다. 즐거운 일이 있을 때 즐겁고, 슬픈 일이 있을 때 슬퍼하는 것이 자연스럽고 건강한 것인데, 우울증을 앓고 있는 사람들은 이것이 뜻대로 되지 않을 때가 대다수이다[3]. 이렇게 우울증은 심각한 질병이지만 이를 적절한 시기에 발견하고 정확하게 진단하는 것이 어려워 문제가 되고 있다.

현재 우울증을 진단하고 치료하기 위해서는 환자가 병원에 방문하여 직접 증상을 말하거나, 상담을 통한 의사의 주관적인 판단, 우울증 진단 관련 설문지를 통해 진단하는 방법이 사용되고 있다. 하지만, 이러한 방법은 객관적인 의견이 배제되어 제한적이고 정확한 진단이 어렵다는 단점이 있다. 또한 세계보건기구(WHO)의 분석에 따르면, 현재 적절한 환경과 전문적인 의료 종사자가 부족하고 우울증 환자가 증가함에 따라 의료 종사자들의 부담이 커져 객관적이고 효과적인 방법으로 우울증을 정확하게 진단하는 것이 힘들어지고 있다[4]. 따라서 우울증을 초기 단계에서 진단하고 적절한 시기에 조치하기 위해서는 객관적이고 정확하게 우울증을 진단하는 방법이 필요하다. 이를 위해 현재 인공지능(Artificial Intelligence, AI) 기술을 적용하여 우울증 진단에 객관적인 도움을 줄 수 있는 연구들이 활발하게 진행되고 있다.

최근 인공지능 기술이 점점 발전하면서 우리 주변의 생활 영역에서 다양하게 활용되고 있다. 인공지능 기술의 발전은 의료 영역에서 큰 변화를 이뤘고, 자세한

정보를 얻기 위한 수단으로 이용되고 있다. 이는 정신건강 분야의 다양한 연구에서 우울증을 포함한 여러 정신 장애를 정확하게 감지하고 객관화하는 데 유용하게 적용되고 있다[7-35]. 우울증 진단을 위한 다양한 인공지능 기술과 딥러닝 알고리즘을 사용하여 우울증 데이터 분석하고 우울증의 여부를 진단하는 몇몇 성공적인 연구들이 있었다[28-30]. 이는 개인의 우울증을 조기에 발견하고 궁극적으로 치료 과정에 도움을 주어 심각한 장애로 이어져 실생활에서 발생할 수 있는 피해를 줄일 수 있는 결과를 이끌었다. 이렇게 선행연구에서 AI를 이용한 우울증 진단 시스템에 대한 일부 결과가 계속 나오고 있지만, 더욱더 객관적이고 정확하게 우울증을 진단하기 위해서는 많은 연구와 개선점이 여전히 필요하다.

일반적으로 AI 기술을 적용한 우울증 진단은 얼굴과 뇌전도(electroencephalogram, EEG), 사람의 목소리, 행동, 텍스트를 등 많은 데이터를 이용한 연구가 진행되고 있다. 특히, 음성과 텍스트와 같은 언어적인 데이터에서 우울증 환자들의 특징이 잘 드러난다. 우울 증상을 가지고 있는 사람들은 말할 때 강도가 보통 사람들보다 낮고, 음역대가 감소하고 말하는 속도가 느리다는 음성적인 특징이 있다[5-6]. 또한 소셜 미디어의 발달로 SNS상에서 사용자들이 많은 글을 올려 본인의 감정을 솔직하게 표현함에 따라 텍스트 속에서도 우울증을 진단할 수 있는 특징들이 드러나고 있다. 이러한 음성과 텍스트를 이용하여 우울증에 대한 특징을 분석할 수 있지만, 우울증은 복합적인 정신질환이기 때문에 하나의 데이터를 이용해 특정 측면에서 분석하는 것은 효과적인 평가를 위해 충분하지 않을 수 있다. 여러 연구에 따르면 다양한 데이터를 융합하게 되면 우울증 진단의 정확도가 현저히 높아진다는 것이 증명되었다[26-35]. 따라서 다양한 데이터 및 특징들을 고려해 정확하게 우울증의 여부를 진단할 수 있는 연구가 필요하다. 본 논문에서는 다양한 요소 중 음성 및 텍스트 데이터를 고려하여 멀티모달 중 바이모달 접근 방식으로 우울증 데이터를 분석하고 우울증을 진단하는 모델을 제안한다. 이는 데이터를 융합하여 우울증 환자의 다양한 특징 및 정보를 담고 있는 딥러닝 모델을 설계한 뒤 우울증의 여부를 효과적으로 판단하는 것에 목적을 두고 있다.

제2절 연구 내용 및 구성

본 논문에서는 음성 및 텍스트 우울증 데이터로부터 딥러닝 모델 중 하나인 Bi-LSTM(Bidirectional Long Short Term Memory)과 CNN(Convolutional Neural Networks) 모델의 late score fusion 방법을 이용한 4-stream 기반 우울증 진단모델을 설계한다. 음성이나 텍스트, 생체신호 중 하나의 데이터만을 이용해 우울증을 진단하는 것은 정보가 부족할 수 있다는 문제가 있다. 이러한 문제를 해결하기 위해 많은 데이터 중 음성 및 텍스트 2가지 데이터 즉, 멀티모달(multi modal) 데이터를 우울증 진단모델의 학습에 사용한다. 두 가지 데이터 중 음성 데이터는 전처리를 통해 잡음을 제거하고 클래스 불균형을 해결하기 위해 우울증 데이터를 확장한다. 그 후 MFCC(Mel Frequency Cepstral Coefficient), GTCC(Gammatone Cepstral Coefficients)를 통해 음성신호의 1차원 특징을 추출하여 Bi-LSTM 모델을 기반으로 학습시킨다. 또한 Bark, ERB, Log-Mel 스펙트로그램과 같은 시간-주파수 변환 기반 2차원 특징을 추출하여 CNN 기반 전이학습 모델을 통해 학습시키고 성능을 비교 분석한다. 텍스트 데이터는 텍스트를 숫자형 시퀀스로 변환한 후 단어로 매핑하는 wordEmbedding 계층을 신경망에 포함해 Bi-LSTM과 CNN을 기반으로 한 우울증 진단모델의 성능을 비교 분석한다. 앞서 제안한 4개의 모델의 소프트맥스 계층에서 출력된 확률값을 late score fusion 방법으로 융합한 멀티모달 데이터를 이용한 4-stream 딥러닝 모델 기반 우울증 진단모델 설계한다. 본 논문에서는 잡음을 제거하고 우울증 데이터 확장을 통해 데이터 질 향상 및 클래스 불균형을 해결하여 단일 모델의 성능이 개선되었고, 이는 최종 결과에 긍정적인 영향을 미쳤다. 최종적으로 음성 및 텍스트 학습에 사용된 4개의 모델을 late fusion 함으로써 제안된 4-stream 기반 딥러닝 모델의 성능이 단일 데이터를 사용한 모델과 기존 2-stream 기반 모델보다 개선되었음을 증명한다.

본 논문은 다음과 같이 구성된다. 2장에서는 딥러닝 기반 우울증 진단에 관한 관련 연구를 음성, 텍스트를 이용한 딥러닝 기반 우울증 진단, 멀티모달 데이터를 이용한 앙상블 모델 기반 우울증 진단 세 부분으로 나누어 기술한다. 3장에서는 본 논문에서는 제안하는 음성 및 텍스트 데이터로부터 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단에 대한 전체적인 흐름과 방법론적인 것을 기술한다. 첫 번째로 인터뷰 형식의 음성으로 이루어진 우울증 데이터를 이용한 우울증 진단모델에 대

한 기본 개념과 설계한 딥러닝 모델에 대해 설명한다. 이는 1차원 데이터에서 특징을 추출하는 방법과 본 연구에 사용된 모델에 관해 기술한다. 또한 1차원인 신호를 2차원 시간-주파수로 변환하는 방법과 이를 이용해 우울증 진단에 사용된 CNN 기반 전이학습 모델에 관해 기술한다. 두 번째로 텍스트 데이터를 이용한 우울증 진단모델에 대해 마지막으로 음성 및 텍스트 데이터를 이용한 4-stream 기반 우울증 진단모델을 제안하고 설명한다. 4장에서는 제안하는 4-stream 기반 우울증 진단모델의 성능을 검증하기 위해 공개 우울증 데이터 세트를 이용하여 실험한 결과를 비교 및 분석한다. 마지막 5장에서는 본 연구의 최종적인 결론과 향후 연구 계획을 기술하고 마무리한다.

제2장 관련 연구

제1절 음성 데이터를 이용한 딥러닝 기반 우울증 진단

여러 연구에서 언어 패턴과 우울증 사이의 상관관계를 발견했으며 조음, 음높이, 말하기 속도 및 크기가 모두 우울증 환자와 건강한 대조군에서 다른 요인을 가진다고 밝혀졌다[5-6]. 그 이후로 음성 특징을 사용하여 환자의 우울증을 진단하는 자동화 프로그램 및 AI 시스템 개발에 관한 많은 연구가 진행되었다. 우울증 환자의 목소리는 일반적으로 음역대가 낮고, 천천히 말하거나 말을 더듬거리거나 속삭이듯 말하는 특징이 있다. 이러한 음성신호에서 운율적 특징, 스펙트럼 및 켑스트럴 특징을 추출하여 순환신경망(Recurrent Neural Network, RNN), 장단기 메모리(Long-Short Term Memory, LSTM), 양방향 장단기 메모리(Bidirectional Long-Short Term Memory, Bi-LSTM), 합성곱 신경망(Convolutional Neural Networks, CNN) 등과 같은 딥러닝 모델을 기반으로 우울증 분류 또는 점수 예측을 진행한다.

음성 데이터를 이용한 우울증 분석 및 진단 연구에서는 1차원 신호 혹은 2차원 신호 특징을 이용한 우울증 진단 크게 2가지 유형으로 나누어진다. 먼저 1차원 신호를 이용한 딥러닝 기반 우울증 진단 연구는 다음과 같이 진행되었다. E. Rejaibi[7]는 우울증을 진단하고 그 심각성을 평가하기 위해 MFCC 방법을 이용한 RNN 모델을 제안한다. 음성신호가 사전 처리되고 이 신호에 대한 MFCC 특징이 추출된 후 정규화된다. MFCC 계수는 연속적인 LSTM 계층을 가진 심층 RNN을 통해 학습된다. 학습 데이터의 부족과 과적합 문제를 극복하기 위해 학습 데이터를 확장했고, 제안된 아키텍처는 DAIC-WOZ 코퍼스에 대해 평가되고 좋은 결과를 달성하였다. Y. Zhao[8]의 연구에서는 DAIC-WOZ과 MODMA corpora 데이터를 사용하여 음성 신호의 핵심 정보를 강조하기 개선된 어텐션 기반 LSTM 우울증 진단모델을 제안한다. 먼저 LSTM에 음성 시퀀스의 고유한 감정 정보를 유지할 수 있는 프레임 수준의 특징을 적용한다. 또한 중요한 고유 정보를 활용하기 위해 LSTM 출력에 시간 도메인의 멀티 헤드 어텐션을 적용한다. 멀티 헤드 어텐션은 축소된 차원을 가진 다양한 컨텍스트 벡터에 대해 LSTM 모델의 출력을 서로 다른 공간에 선형으로 투영하는데 도움이 된다. D. Sztahó[9]는 연속적인 언어에서 음성 장애를 감지하기

위해 스펙트로그램 특징을 입력으로 사용하는 Multi-Task Learning과 함께 LSTM 기반 Autoencoder Hybrid 모델을 제안한다. 이 방법의 장점은 데이터 중심적이어서 음향 및 음성신호에 특별한 전처리가 필요 없다는 것이다. 음성 우울증 데이터 세트에 대해 90%의 정확도를 얻음으로써 제안된 모델의 타당성을 입증하였다.

1차원 신호를 이용한 연구보다 2차원 시간-주파수 표현을 이용한 딥러닝 기반 우울증 진단에 관련된 연구가 더 활발하게 진행되었다. L. He[10]는 음성신호에서 우울증의 심각한 정도를 효과적으로 분류하기 위해 hand-craft 및 딥러닝 특징을 이용하여 이를 융합한 DCNN(Deep Convolutional Neural Networks) 모델을 제안한다. 음성 우울증 데이터베이스로는 AVEC 2014(Audio/Visual Emotion Challenge and Workshop 2014) 데이터 세트를 사용했고, 우울증 점수 예측을 개선하기 위해 AVEC 2013 및 AVEC 2014 데이터 세트를 결합하여 새로운 확장 데이터베이스를 얻었다. 먼저 DCNN은 원래 음성신호와 음성신호를 스펙트로그램으로 변환하여 심층 학습된 특징을 학습하고 스펙트로그램에서 강력한 확장 로컬 이진 패턴 중앙값을 추출하도록 설계하였다. 마지막으로, 원시 및 스펙트로그램 DCNN을 결합하기 위해 조인트 튜닝 계층을 적용하는 것을 제안하여 우울증 분류 성능을 향상시켰다. hand-craft 모델과 딥러닝 모델을 모두 사용했을 때 조인트 튜닝의 성능이 향상되었다.

X. Ma[11]은 AVEC 2016에서 음성 기반 우울증 분류를 위한 DepAudionet 모델을 제안한다. 이 방법은 2개의 CNN 계층, 1개의 LSTM 계층 및 2개의 완전 연결 계층을 쌓아 네트워크를 설계하여 음성신호의 대상이 우울한지를 분류하는 것이다.

A. V. Romero[12]는 음성신호를 이용하여 합성곱 신경망 앙상블 평균화 방법 기반 우울증 자동 분류 모델을 제안한다. 이는 2016년 AVEC-2016의 Depression Classification Sub-Challenge(DCC)에서 제공된 데이터와 실험 프로토콜을 사용하여 평가된다. 전처리 단계에서 음성신호를 로그 스펙트로그램으로 변환하고 양성 및 음성 샘플의 균형을 맞추기 위해 무작위로 샘플링된다. 분류모델은 1차원 합성곱 신경망(1d-CNN)을 이용한 X. Ma[11]의 연구에서 제안된 DepAudioNet을 기반으로 하지만 마지막 LSTM 계층을 포함하지 않는다. 1d-CNN은 로그 스펙트로그램을 입력으로 사용하며, 입력 계층 1개, 은닉계층 4개 및 출력계층 1개로 구성된다. 몇몇 1d-CNN 기반 모델은 서로 다른 초깃값으로 학습된 다음 앙상블 평균 알고리즘을 사용하여 개별 예측값을 융합하고 화자별로 결합하여 최종 결정을 내린다. 제안된 앙상블 모델은 DepAudionet 기반 모델과 SVM(Support Vector Machines) 분

류기와 비교했을 때 그보다 더 좋은 결과를 얻었다.

L.Yang[13]의 연구에서는 우울증 진단을 위한 음성 데이터가 부족하여 특징을 생성하기 위해 DCGAN(Deep Convolutional Generative Adversarial Net)을 기반으로 하는 데이터 확장 접근 방식을 제안한다. AVEC2016(Audio Visual Emotion Challenge) 우울증 데이터 세트의 음성신호를 이용했고, 특징은 GeMAPS(Geneva Minimalistic Acoustic Parameter Set) 및 INTERSPEECH Challenges 특징 세트 이용하여, 총 238개의 LLD(저수준 설명자)를 먼저 추출한 다음 각 음성 세그먼트에 대한 6,902차원 특징 벡터를 생성하였다[14-15]. 그 후 DCGAN을 이용하여 특징 데이터를 확장하고 VGG-SVM을 기반으로 학습을 진행한 뒤 성능을 평가한다. 이는 우울증 진단 시스템의 성능 향상을 위한 새로운 관점을 제공하였다.

M. Muzammela[16]는 인공지능(AI) 기반 임상 우울증 진단 및 언어 평가 응용 프로그램인 깊은 음소 수준의 분석 및 학습이 진행되고 모음, 자음 공백 및 모음과 자음의 융합을 이용한 3개의 스펙트로그램 기반 CNN 네트워크를 제안하였다. 자음 기반 CNN 아키텍처는 모음 CNN 네트워크보다 성능이 뛰어나다. 두 네트워크를 결합하면 결과가 크게 향상되고 성능이 눈에 띄게 향상된다. 이러한 결과는 제안된 딥러닝 기반 모음과 자음 분석이 임상적 우울증의 자동 진단에 매우 신뢰성이 있음을 보여준다.

R. Flores[17]는 음성을 이용해 학습된 CNN 기반 전이학습 모델 중 하나인 VGGish의 임베딩에 대해 LSTM 계층을 포함하고 self-attention 메커니즘을 추가하여 변형된 VGGish Attention 모델을 제안한다. 먼저 VGGish 모델을 통해 음성 특징 임베딩을 생성한다. VGGish는 음성 클립을 로그 멜 스펙트로그램으로 변환하여 음성의 초당 크기가 128인 임베딩 벡터를 추출하여 분류에 사용할 수 있는 2D 배열을 형성하는 다층 합성곱 신경망에서 처리된다. VGGish 모델은 긴 시퀀스에 적용될 때 성능이 떨어지지만 제안된 VGGish Attention 모델의 성능은 긴 음성 시퀀스에서 크게 저하되지 않았다. 또한 VGGish Attention 모델은 VGGish 모델보다 성능이 뛰어나다는 것을 증명하였다.

V. Ravi[18]의 연구에서는 음성신호로부터 DepAudioNet을 기반으로 우울증 진단을 하기 위한 프레임 레이트 기반의 데이터 확장(Frame Rate Based Data Augmentation, FrAUG) 기법을 제안한다. 프레임 속도 매개변수를 변경하여 모델에 다양한 해상도를 가진 시간-주파수 특징을 학습시켰다. 학습 데이터는 프레임 이동 매개변수와 프레임 너비를 변경하여 생성된 새로운 특징 샘플로 확장되었습니

다. 제안된 접근 방식은 성대 또는 음성 소스 관련 매개변수를 수정하지 않았으므로 MDD 모델링에 중요한 음향 정보를 보존하였다. FrAUG는 DAIC-WOZ 데이터 세트의 멜 스펙트로그램 특징을 사용하여 훈련된 DepAudioNet 및 CONVERGE의 음성 데이터의 MFCC 특징을 사용하여 사전 훈련된 모델에서 생성된 x-벡터 임베딩으로 훈련된 다운스트림 네트워크의 분류 성능을 개선하였다. 이렇게 음성 신호를 이용한 딥러닝 기반 우울증 진단 연구들이 다양한 방면으로 활발하게 진행되었다.

제2절 텍스트 데이터를 이용한 딥러닝 기반 우울증 진단

인터넷의 발달로 사람들은 트위터, 페이스북, 인스타그램과 같은 다양한 소셜 미디어 플랫폼을 통해 자신의 감정, 의견을 표현하고 일상을 공개하는 경향이 있다[19]. 이러한 소셜 미디어 플랫폼에 많은 사용자가 있기 때문에 감정이나 우울증을 분석할 수 있는 데이터가 많다. 그중 가장 많이 사용되는 커뮤니케이션 형식은 텍스트이고, 이러한 광범위한 텍스트 데이터를 이용하여 우울증 진단 시스템을 구축하는 여러 연구가 수행되었다.

Seojeong Park[20]의 연구에서는 한국어 소셜 미디어 텍스트를 활용한 딥러닝 기반의 우울 경향 모델을 제안한다. 네이버 지식인, 블로그, 하이닥, 트위터에서 데이터를 수집했고, DSM-5 주요 우울 장애 진단 기준을 이용하여 우울 증상 개수에 따라 클래스를 구분하였다. 이후 구축한 TF-IDF 분석과 동시 출현 단어 분석을 통해 데이터의 클래스별 특징을 살펴보았다. 또한, 다양한 텍스트 특징을 활용하여 우울 경향 분류모델을 생성하기 위해 단어 임베딩과 사전 기반 감성 분석, LDA 토픽 모델링을 수행하였다. 이를 통해 문헌 별로 임베딩된 텍스트와 감성 점수, 토픽 번호를 산출하여 텍스트 특징으로 사용하였다. 그 결과 임베딩된 텍스트에 문서의 감성 점수와 토픽을 모두 결합하여 KorBERT 알고리즘을 기반으로 우울 경향을 분류하였을 때 83.28%의 높은 정확성을 증명하였다.

A. H. Orabi[21]의 연구에서는 우울증 검출과 예측을 위한 딥러닝 알고리즘을 제안한다. 이는 우울증이 있는 사용자들과 우울증이 없는 사용자들의 트위터 문자를 수집하여 우울증 진단을 수행하였다. 먼저, Word2Vec 모델을 사용하여 텍스트를 임베디드하고, 임베디드 텍스트를 기반으로 하는 양방향 LSTM, CNN 및 RNN 기반 알고리즘과 같은 딥러닝 알고리즘을 사용하여 우울증 진단모델의 성능 비교를 시도하였다. 기계학습 알고리즘인 SVM 기반 우울증 분류모델의 성능에 비해 CNN 알고리즘을 이용한 분류모델의 성능이 크게 향상되었으며, 딥러닝 알고리즘이 우울증 검출에 적합한 것으로 나타났다.

F. M. Shah[22]는 사용자의 텍스트 포스트를 분석하여 우울증을 진단할 수 있는 하이브리드 모델을 제안한다. CLEF eRisk 2017[23]에서 우울증의 조기 탐지(Early Detection of Depression in CLEF eRisk 2017)를 위해 제시된 reddit의 데이터 세트를 이용해 TrainableEmbed, GloveEmbed, Word2Vec Embedding, Fasttext Embedding, Meta data 특징을 추출해 다양한 방법으로 실험을 진행하였다. Embed

특징들은 Bi-LSTM의 입력으로 들어갔고 출력 차원이 300인 은닉계층과 연결된다. 정규화된 Meta data 특징은 출력 차원이 10인 은닉계층과 연결된다. 그 후, 이 2개의 은닉층에서 출력된 벡터가 연결되고, 연결 계층의 출력 차원은 310이다. 과적합을 방지하기 위해 0.2의 드롭아웃 계층을 추가하고 마지막으로 이진 분류를 수행할 때 출력 차원이 1이고 활성화 함수가 시그모이드인 은닉층을 추가한다. 이러한 하이브리드 모델을 이용하여 텍스트 기반 우울증 진단 결과 Word2VecEmbed+Meta 특징을 입력으로 넣었을 때, 가장 좋은 성능을 보였다.

A. Amanat[24]는 텍스트 데이터를 처리하고 우울 특성을 감지하는 데 중점을 두었다. 트위터 데이터에서 우울증 증상과 그 사람의 정서를 나타내기 위해 One-Hot 인코딩 방법과 PCA(주성분 분석)를 사용하여 텍스트 데이터 세트에서 특징을 추출한다. 데이터 학습을 위해 LSTM을 사용하는 딥러닝 모델을 제안하는데, LSTM 유닛은 두 개의 은닉상태와 바이어스를 가지고 있고, 우울하고 우울하지 않은 샘플 데이터를 이용해 두 개의 은닉계층이 있는 RNN 모델에 학습시켜 우울증을 진단한다. 제안된 접근 방법은 수많은 소셜 미디어 가입자의 감정에서 우울증을 조기에 인식하는 탁월한 결과를 달성함으로써 RNN 및 LSTM의 실행 가능성을 나타낸다.

A. Shankdhar[25]의 연구에서는 트윗을 활용하여 자연어 처리(NLP) 도구를 기반으로 사용자가 우울증에 걸릴 가능성을 분류하는 하이브리드 BiLSTM+CNN 모델을 제안한다. 텍스트 데이터는 이모티콘을 텍스트 의미로 대체하는 Emoji Processing, 데이터의 크기를 줄이기 위한 General Pre-processing, Word Embedding을 통해 전처리 된다. 분류모델은 BiLSTM과 CNN이 융합된 모델로 Bi-LSTM 계층은 자연어 특징을 인코딩하는 데 사용되고, CNN 계층은 범주적 특징을 인코딩하는 데 사용된다. 하이브리드 BiLSTM + CNN 모델은 높은 정확도를 보임으로써 트윗에서 우울한 사용자를 식별할 수 있음을 증명하였다.

제3절 멀티모달 데이터를 이용한 앙상블 기반 우울증 진단

최근 몇 년 동안 많은 연구자가 표정, 오디오 특징, 텍스트 등과 같은 생리적 지표를 분석하여 자동 우울증 진단을 위한 객관적인 AI 도구를 찾기 위해 노력하고 있다. 멀티모달 분석을 통해 포괄적이고 강력한 특징을 얻을 수 있어 멀티모달 분석이 널리 주목받고 있으며, 이에 따라 단일 모드 우울증 진단에 관한 연구는 점차 멀티모달 우울증 진단으로 옮겨가고 있다. 우울증은 복잡하고 다인자 장애이기 때문에 우울증 진단을 위한 멀티모달 접근은 중요하며 자동 우울증 진단을 처리하는 데 다양한 양식의 정보를 고려하는 것이 필수적이다. 우울증 진단을 위해 개발된 일부 멀티모달 접근법은 다음과 같이 연구되어왔다.

J. Park[26]은 단일 데이터를 이용한 우울 탐지 모델의 낮은 정확도 문제를 개선하기 위해 멀티모달 데이터 기반 어텐션 메커니즘 우울 진단모델을 제안한다. 제안된 모델은 자연어 분석을 위한 BERT-CNN(Bidirectional Encoder Representations from Transformers-Convolutional Neural Network)의 양방향 인코더, 음성신호 처리를 위한 CNN-BiLSTM(CNN-Bidirectional Long Short-Term Memory), 우울증 진단을 위한 멀티모달 분석 및 융합 모델로 구성된다. 이는 DAIC-WOZ의 음성 및 텍스트 데이터를 사용하여, 음성 데이터에서는 로그-멜 스펙트로그램으로 변환하여 CNN 모델을 통해 특징을 추출하고 어텐션 메커니즘을 적용한 Bi-LSTM을 기반으로 학습된다. 텍스트 데이터는 BERT 토크나이저 이용하여 임베딩 벡터로 변환하고 사전 훈련된 BERT-CNN 모델을 미세 조정하고 학습하여 특징 벡터를 얻는다. 제안된 모델을 통해 단일 데이터 사용으로 인한 급격한 손실 증가를 해결하고 향상된 정확도를 보여주었다.

J. Xiao[27]는 오디오 및 텍스트 시퀀스를 기반으로 하는 자동 우울증 감지를 위한 새로운 접근 방식을 제안한다. 오디오 특징추출을 위해 C-CNN(Casual-Convolutional Neural Network)에 어텐션 메커니즘을 도입하여 우울증 진단을 위한 새로운 모델인 Attention-C-CNN을 제안한다. 텍스트 특징추출을 위해서는 2018년 Google에서 제안한 사전 학습된 모델인 BERT를 활용한다. 이에 co-attention 변환 계층에서 다중 헤드 어텐션의 키-값 쌍을 교환하여 오디오 및 텍스트 특징을 융합할 수 있는 co-attention를 적용하였다. DAIC-WOZ 데이터 세트를 이용하여 실험한 결과 제안된 모델은 단일 모드 데이터 및 기존의 멀티모달 데이터 기반 방법보다 좋은 성능을 얻었다.

M. Niu[28]는 텍스트/오디오 양식에 GAT(Graph Attention Network)[29]를 적용하여 관계형 인터뷰 질문 사이의 컨텍스트 정보를 효과적으로 파악하고 통합할 수 있는 우울증 진단을 위한 HCAG(Hierarchical Context-Aware Graph) 기반 어텐션 모델을 제안한다. 구체적으로, 계층적 상황 인식 구조는 답변 내에서 중요한 정보를 파악할 수 있으며, GAT 네트워크는 인터뷰 질문 사이에 충분한 관계 및 논리적 정보를 추가로 수집할 수 있다. 결과는 제안된 HCAG가 더 강력하고 5가지 평가 메트릭 모두에서 기준 모델보다 성능이 우수함을 보여준다.

Y. Shen[30]의 연구에서는 우울증 진단 연구를 위해 162명의 자원봉사자의 인터뷰에서 추출한 오디오 및 텍스트 기록으로 구성된 공개적으로 사용 가능한 중국 우울증 데이터 세트인 EATD-Corpus와 인터뷰 형식의 임상 우울증 데이터인 DAIC-WoZ 데이터 세트를 이용하는 자동 우울증 진단을 위한 멀티모달 기반 딥러닝 모델을 제안한다. 제안된 모델은 GRU(Gate Recurrent Unit) 모델과 어텐션 계층이 있는 Bi-LSTM 모델을 사용하여 오디오 및 텍스트 특징을 융합한다. 텍스트 특징은 ELMo를 사용하여 문장을 고차원 문장 임베딩으로 투영하여 추출한다. 오디오 특징의 경우 오디오에서 Mel 스펙트로그램을 추출한다. 실험 결과는 제안된 방법이 매우 효과적임을 보여준다.

C. Lau[31]는 우울증 진단, 중증도 추정 및 증상 중증도 진단 측면에서 자동화된 우울증 평가 성능을 향상시키기 위한 바이모달 MTL(Multi-Task Learning) 패러다임을 제안한다. DAIC-WOZ 데이터를 이용해 딥러닝 모델 중 하나인 Bi-LSTM을 기반으로 하여 어텐션 메커니즘을 적용하여 우울증을 진단한다. 실험은 우울증 평가 결과의 공동 예측이 예측성능과 확률 보정 모두를 향상시키는 것으로 나타나 경쟁적인 결과를 얻을 수 있었다.

L. Yang[32]의 연구에서는 심층 합성곱 신경망(DCNN)과 심층 신경망(DNN) 모델로 구성된 멀티모달 융합구조를 가지는 모델을 제안한다. 또한, 텍스트 및 비디오에 대한 새로운 특징 설명자를 제안한다. 제안된 모델은 오디오, 비디오 및 텍스트 데이터를 고려한다. 각 데이터에 대해 hand-crafted 특징 설명자는 DCNN에 입력되어 컴팩트한 동적 정보로 높은 수준의 전역 특징을 학습한 다음 학습된 특징을 DNN에 공급하여 PHQ-8 점수를 얻는다. 다중 모드 융합의 경우 세 가지 모드에서 추정된 PHQ-8 점수가 DNN에 통합되어 최종 PHQ-8 점수를 얻는다. 텍스트 설명자의 경우 수면 장애, 단락 벡터(PV)를 사용하여 이러한 문장의 분산 표현을 학습한다. 비디오 설명자의 경우 변위와 속도를 측정하기 위해 얼굴 랜드마크에서 직

접 계산된 새로운 전역 설명자인 HDR(Histogram of Displacement Range)을 제안한다. AVEC 2017 우울증 데이터 세트에 대한 실험 결과는 DCNN과 DNN 모델을 융합하는 멀티모달 하이브리드 구조가 유망한 정확도를 나타냄을 보여준다.

T. Alhanai[33]의 연구에서는 오디오 및 텍스트 특징을 멀티모달 형태로 결합한 LSTM 모델을 제안한다. 오디오 및 텍스트 특징은 사람의 상태에 대한 차별적이고 시간적으로 변화하는 정보뿐만 아니라 보완적인 정보도 포함될 수 있다. 제안된 모델은 두 개의 LSTM 분기로 나누어져 있으며, 오디오 데이터는 COVAREP 특징을 이용하고, 이를 학습 시키기 위해 LSTM 계층을 쌓아 모델을 구성한다. 또한 텍스트 데이터는 Doc2Vec를 통해 특징을 추출하고 이를 학습 시키기 위해 2개의 LSTM 계층을 쌓아 모델을 구성한다. 마지막 출력은 최종 피드포워드 네트워크에 융합한다. 분기는 서로 다른 토폴로지로 구성되었으며 각 모달리티의 특성 및 정보 내용에 대해 최적화하였다. 본 연구에서는 시간 도메인 상에서 음성 및 텍스트 데이터의 시퀀스 특징만 사용하여 주파수 특성이 시간에 따라 달라지는 것을 고려하지 않았다. 이 연구는 여러 가지의 오디오 및 텍스트 특징을 이용하였지만 음성의 2차원 이미지 특징을 고려하지 않았고, 성능 평가 지표로 사용된 정밀도(Precision), 재현율(Recall), F1-Score가 각각 0.71, 0.83, 0.77으로 성능 개선에 관한 연구가 여전히 필요하다.

L. Lin[34]의 연구에서는 환자 인터뷰에서 오디오 및 텍스트 양식의 정보를 결합하여 새로운 자동 우울증 진단 방법을 제안한다. 구체적으로 텍스트 콘텐츠를 처리하는 어텐션 계층이 있는 Bi-LSTM 네트워크와 음성신호를 처리하는 1D-CNN 모델의 출력을 통합하여 두 개의 FC 네트워크에 공급되어 분류를 진행하는 네트워크 구조를 갖는다. 이 방법은 우울증의 존재를 진단하고 우울 증상의 중증도를 평가한다. 공개적으로 사용 가능한 두 가지 데이터 세트, 즉 DAIC-WOZ 데이터 세트와 AViD-Corpus(Audio-Visual Depressive Language Corpus) 데이터 세트를 이용하여 제안된 모델을 평가한 결과 우울증을 진단하는 작업에서 높은 성능을 달성한다. 이 연구에서는 음성신호의 2차원 시간-주파수 변환 기반 Mel 스펙트로그램 특징을 이용하고, 텍스트에서 중요하지 않은 메타 정보를 제거하고 각각 1-D CNN과 어텐션 계층이 있는 Bi-LSTM에 적용해 이를 앙상블 하는 방법을 제안했다. 하지만 음성의 스펙트럼에서 배음 구조를 유추할 수 있는 MFCC, GTCC 등의 특징을 고려하지 않았다. 또한 데이터 불균형 문제를 해결하기 위해 데이터 리샘플링 방법을 이용하여 균형 잡힌 DAIC-WoZ 데이터 세트를 구성했지만 이에 대한 정밀도

(Precision), 재현율(Recall), F1-Score가 각각 0.85, 0.79, 0.92로 여전히 성능 개선에 대한 연구가 필요하다.

G. Lam[35]의 연구에서는 우울증 진단을 위한 상황 인식 및 데이터 기반 접근 방식을 통합하는 새로운 방법을 제안한다. 주제 모델링을 기반으로 하는 데이터 확장 절차를 도입하고 오디오 및 텍스트 양식에 대한 딥러닝 모델 학습에 대한 효율성을 보여준다. 심층 1D CNN 및 Transformer 모델은 각각 오디오 및 텍스트 데이터를 학습시킨 결과 좋은 성능을 얻었고, 멀티모달 구조에서는 더욱 개선된 성능을 보였다. 이 연구에서는 주제 모델링을 기반으로 하는 데이터 확장을 훈련 데이터 세트에서 진행하고 음성 및 텍스트 데이터를 융합하여 2-stream 기반 우울증 진단모델을 제안했는데, 높은 성능을 가지지만 더욱 정확한 우울증 진단을 위해서는 성능 개선에 관한 연구가 필요하다.

이러한 연구들은 1차원 음성의 특징 또는 2차원 시간-주파수 변환 기반 스펙트로그램 특징 중 한 가지만 사용하여 음성의 스펙트럼에서 배음 구조를 유추할 수 있는 MFCC, GTCC 등의 특징과 비선형성 특성이 있는 음성의 시간에 따라 변하는 주파수 특징을 동시에 고려하지 못한다. 또한 텍스트 데이터는 주로 입력값을 순차적으로 처리하는 LSTM 모델에 적용되어 문장의 지역 정보를 보존함으로써 단어 및 표현의 등장순서를 학습에 반영하지 못해 우울증을 정확하게 진단하지 못하는 경향이 있다. 따라서 본 논문에서는 1차원 음성신호의 특징추출 방법인 MFCC와 GTCC를 동시에 사용하여 음성에 담긴 많은 정보를 얻는다. 또한 심리 음향학에서 사용되는 주파수 척도를 이용하여 사람 음성의 특징을 더 잘 드러낼 수 있는 스펙트로그램을 생성한다. 이 두 가지 특징을 모두 융합함으로써 스펙트럼의 배음 구조를 유추하고 시간에 따라 변하는 주파수 특징을 동시에 고려한다. 또한 텍스트 학습모델로 Bi-LSTM뿐만 아니라 단어의 순서를 고려하여 다음 단어를 예측하고, 오타를 발견할 수 있는 n-gram 개념을 기반으로 하는 CNN 모델도 사용하여 입력값을 순차적으로 처리할 뿐만 아니라 단어/표현의 등장순서도 학습에 반영한다. 이렇게 4개의 딥러닝 모델을 융합함으로써 학습이 무거워질 수 있지만 많은 정보를 동시에 고려할 수 있다. 데이터 전처리 부분에서는 음성에서 소음, 묵음, 다른 화자의 말소리 등을 제거하고 텍스트에서 불필요한 단어들을 정리함으로써 데이터 질을 향상시킨다. 또한 데이터양이 많은 클래스에 학습이 치중될 수 있는 클래스 불균형 문제를 해결하기 위해 우울증 데이터에 대한 확장을 진행한다. 이는 제안된 우울증 진단모델의 성능 향상에 긍정적인 영향을 준다.

제3장 음성 및 텍스트 데이터로부터 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단

본 장에서는 음성신호의 1차원 2차원 특징추출 방법과 이를 이용한 우울증 진단모델을 Bi-LSTM과 CNN 전이학습 모델을 이용해 설계하고, 텍스트 데이터의 워드 임베딩 방법에 관해 기술하고 이를 통해 우울증 진단모델을 Bi-LSTM과 n-gram 기반 CNN 모델을 이용해 설계한다. 마지막으로 멀티모달 데이터 중 음성 및 텍스트 데이터를 이용하여 Bi-LSTM과 CNN 모델을 late score fusion 하여 4-stream을 기반으로 우울증을 진단하는 방법을 제안한다. 제안하는 우울증 진단 모델의 흐름도는 그림 3.1에서 확인할 수 있다.

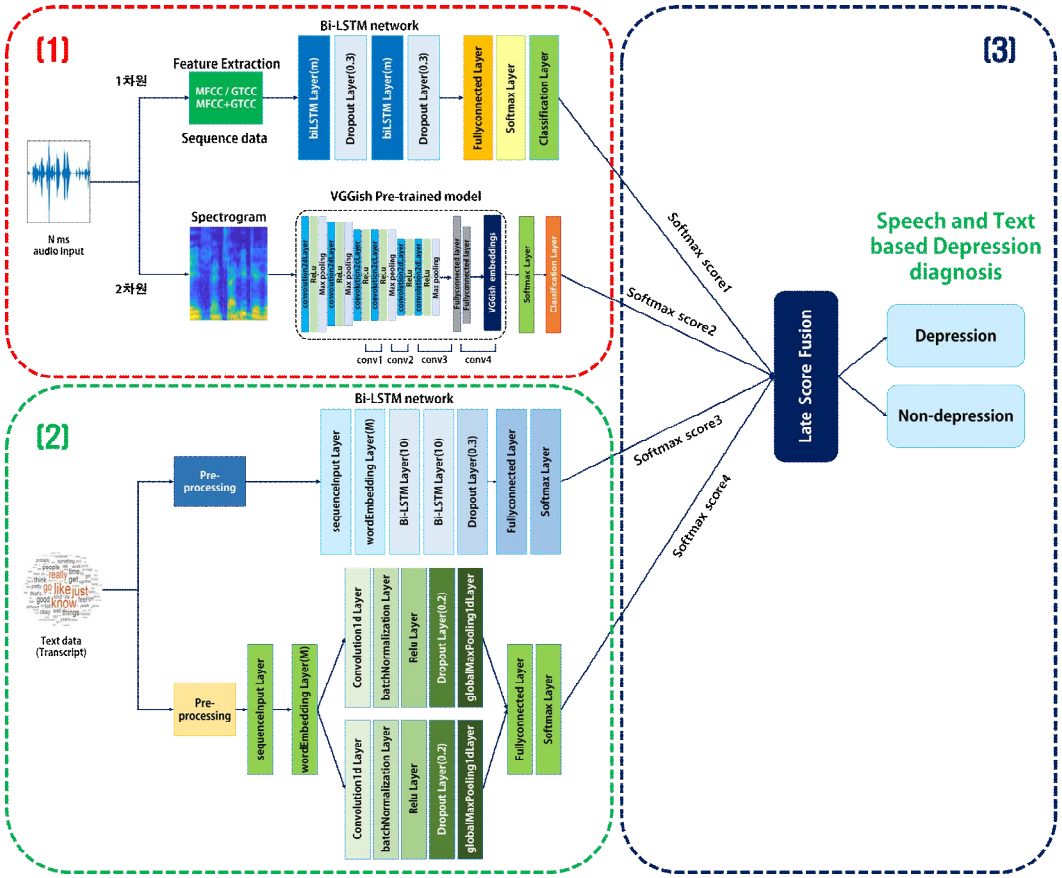


그림 3.1 제안된 우울증 진단모델의 흐름도

제1절 음성신호를 이용한 딥러닝 모델 설계

1. 1차원 음성신호 특징추출 방법

가. MFCC(Mel Frequency Cepstral Coefficient)

음성 기반 인터페이스를 개발하기 위해서 가장 중요한 기술은 음성 데이터에서 좋은 특징을 추출하는 것이다. 가장 많이 사용되는 특징추출 방법은 LPC(Linear Predictive Coding) cepstrum, PLP(Perceptual Linear Predictive) cepstrum, MFCC, 필터뱅크 에너지 등이 있다. 그중 MFCC는 음성신호에서 추출할 수 있는 특징으로 소리의 고유한 특징을 나타내는 수치를 말한다. 이는 인간의 청각 주파수의 특성을 반영했기 때문에 일반적으로 음성인식에서 가장 효과적인 방식으로 사용되고 있는 특징추출 방법의 하나이다[36]. MFCC의 계산 과정은 그림 3.2에서 간략하게 볼 수 있고 그 과정은 다음과 같다.

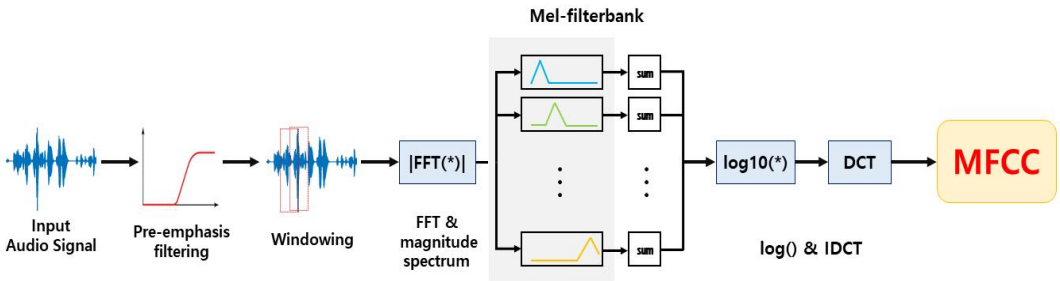


그림 3.2 MFCC 계산 과정

[과정1] Pre-emphasis filtering

Pre-emphasis filtering은 간단히 말해서 음성신호에 High-pass filter를 적용한 것이다. 사람의 음성은 고주파 부근에서 크기가 작아지는 특성이 있으므로 전처리를 통해 고주파 성분을 변조가 강하게 걸리도록 강조해줄 수 있다. 이를 통해서 주파수 스펙트럼의 밸런스를 맞출 수 있다. pre-emphasis filter 식은 수식 (3.1)과 같이 정의할 수 있다.

$$y(m) = s(m) - \gamma s(m-1) \tag{3.1}$$

여기서, 감마값은 pre-emphasis의 정도를 제어하는 값으로 필터의 차단 주파수를 정한다. 감마값은 0.95 또는 0.97로 지정할 수 있다. pre-emphasis는 cepstral mean normalization 과정으로 대체할 수 있어 생략할 수 있다.

[과정2] Windowing

음성신호에서 주파수 성분을 뽑아내기 위해서는 푸리에 변환(Fourier Transform)을 적용해야 한다. 하지만 음성은 비정상성(non-stationary) 데이터이기 때문에 전체 신호를 고속 푸리에 변환(Fast Fourier Transform, FFT)하지 않는다. 길이가 다른 음성은 학습시키기 어렵기 때문에 음성 데이터를 20~40ms로 쪼개 프레임 단위로 분할한다. 이때 프레임을 50% overlap 하여 프레임끼리의 연속성을 만들어준다. 여기서 각각의 프레임들에 대해서 윈도우를 적용하게 되는데, 그 이유는 A 프레임과 B 프레임이 서로 연속되지 않는다면 프레임이 접합하는 부분에서의 주파수 성분이 무한대가 되어버리기 때문이다. 이러한 문제를 방지하고 프레임의 시작점과 끝점을 똑같이 유지해주기 위해 대표적으로 해밍윈도우(hamming window)라는 함수를 많이 사용한다. 해밍윈도우의 식은 수식 (3.2)와 같다.

$$w[m] = 0.54 - 0.46\cos\left(\frac{2\pi m}{M-1}\right) \qquad 0 \le m \le M-1 \tag{3.2}$$

여기서 m은 해밍윈도우 값의 인덱스를 나타내고, M은 윈도우의 길이를 뜻한다. 본 연구에서는 MFCC 특징추출을 위해 해밍윈도우를 사용하였으며 윈도우 길이는 480으로 설정하였다.

[과정3] FFT & magnitude spectrum

각각의 프레임들에 대해 푸리에 변환을 적용하여 주파수 성분을 얻어낸다. 푸리에 변환이란 시간(time) 도메인의 음성신호를 주파수(frequency) 도메인으로 바꾸는 과정을 말한다. 푸리에 변환을 실제로 적용할 때는 고속 푸리에 변환이라는 기법을 사용하게 되는데, 이는 기존 푸리에 변환에서 중복된 계산량을 줄이는 방법이다. 과정 3부분의 고속 푸리에 변환까지만 적용하더라도 충분히 학습 가능한 특징을 뽑을 수 있다. 하지만 사람 몸의 특성을 고려한 멜 척도(mel-scale)를 적용한 특징이 보통 더 나은 성능을 보이기 때문에 다음 과정을 진행한다.

[과정4] Mel-filter bank

각각의 프레임에 대해 얻어낸 주파수들 성분들에서 멜 값을 얻어내기 위한 필터를 적용한다. 이는 그림 3.2.에서 볼 수 있듯이 달팽이관의 특성을 고려하여 낮은 주파수에서는 작은 삼각형 필터를, 고주파 대역으로 갈수록 넓은 삼각형 필터를 적용한다. 주파수가 8kHz일 때 주로 멜 필터의 개수 R=24인 멜 필터뱅크(mel-filter bank)를 이용할 수 있다. 그림 3.3과 같은 삼각형 필터 N개를 모두 적용한 필터를 멜 필터뱅크라고 한다. 앞서 푸리에 변환을 적용한 신호를 멜 필터뱅크에 통과시키게 되면 멜 스펙트럼 특징을 추출할 수 있다.

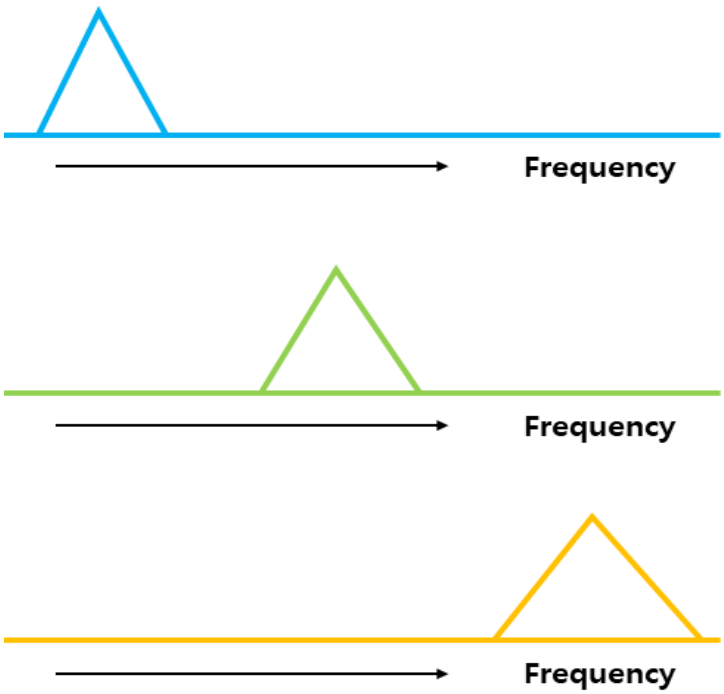


그림 3.3 Mel-filter bank 개념

[과정5] log() & DCT

앞서 나온 멜 스펙트럼 특징에 대해 log를 취하고 행렬을 압축해서 표현해주는 이산 코사인 변환(Discrete Cosine Transform, DCT) 연산을 수행한다. 또한 이는 앞의 멜 스펙트로그램은 주파수끼리 상관관계(correlation)가 형성되어 있는데, 이를 de-correlate 해주는 역할을 수행한다. 이 과정까지 모두 수행하게 되면 출력으로 MFCC 특징을 얻을 수 있다.

나. GTCC(Gammatone Cepstral Coefficients)

GTCC는 음성 인식 시스템의 또 다른 FFT 기반의 음성 특징추출 방법이다. 이는 각 지점에서 청각 필터의 폭을 측정하는 등가 직사각형 대역폭(Equivalent Rectangular Bandwidth, ERB) 대역의 감마톤 필터 뱅크를 기반으로 한다. 감마톤 필터 뱅크의 임펄스 응답은 인간의 청각 필터의 크기 특성과 매우 유사하다[37]. GTCC의 계산 과정은 그림 3.4와 같다.

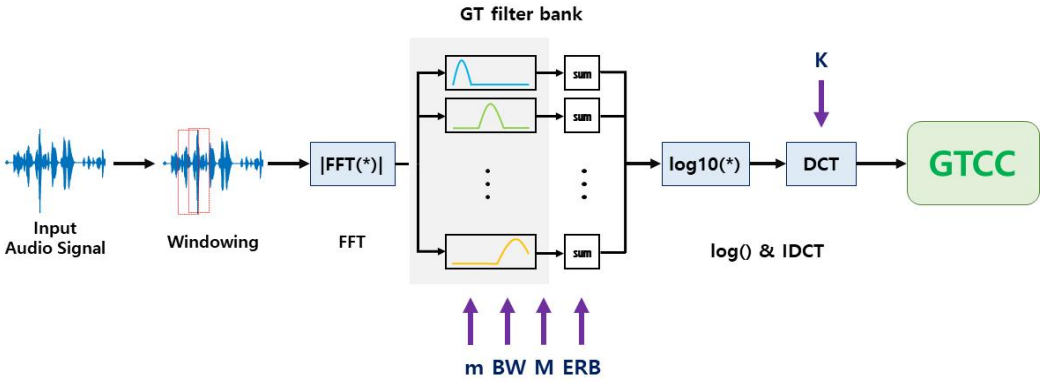


그림 3.4 GTCC 계산 과정

GTCC의 계산 과정은 MFCC 특징추출 방법과 유사하다. 먼저 음성신호를 10~50ms의 짧은 프레임으로 나눈다. 그 후, 음성신호에 FFT 연산을 수행하여 시간영역에서 주파수 영역으로 변환하고, 이는 감마톤 필터뱅크를 통과하여 시각적으로 의미있는 음성신호의 주파수를 강조한다. 마지막으로 로그 함수와 이산 코사인 변환을 적용하여 로그 압축 필터 출력을 역상관하고 더 나은 에너지 압축을 생성하여 GTCC 특징을 얻는다. GTCC를 구하는 식은 수식 (3.3)과 같다.

$$GTCC_s = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \log(F_m) \cos\left[\frac{\pi m}{M}\left(s - \frac{1}{2}\right)\right] \qquad 1 \le s \le K \tag{3.3}$$

여기서 F_m은 m번째 스펙트럼 대역에서 신호의 에너지를, M은 감마톤 필터의

수, K는 GTCC의 개수를 나타낸다.

2. 2차원 시간-주파수 변환 기반 특징추출 방법

가. 바크 스펙트로그램(Bark Spectrogram)

바크 주파수 척도(Bark Frequency Scale)는 1961년 독일의 음향 과학자 Eberhard Zwicker가 제안한 심리학 음향 척도이고, 소리의 크기에 대한 주관적인 측정을 최초로 제안한 Heinrich Barkhausen의 이름을 따서 명명되었다[38]. 인간은 소리의 크기와 높이, 길이, 음색 등의 특징을 청각기관을 이용하여 구분할 수 있는데, 소리의 특징을 구체적으로 나타내기 위해서는 서로 다른 소리를 구별하는 척도가 필요하다. 따라서 수많은 심리 음향 실험의 결과에 기초하여, 인간의 청각 임계 대역이 각각 하나의 Bark 너비를 갖도록 바크 척도를 정의하였다. 척도에는 1에서 24 사이의 값이 존재하고, 이 24개의 값은 24개의 임계 청각 대역과 관련이 있다. 이러한 척도는 음성 데이터에서 특정 대역에 존재하는 중요 요소를 나타내는 데 있어 효과적이다.

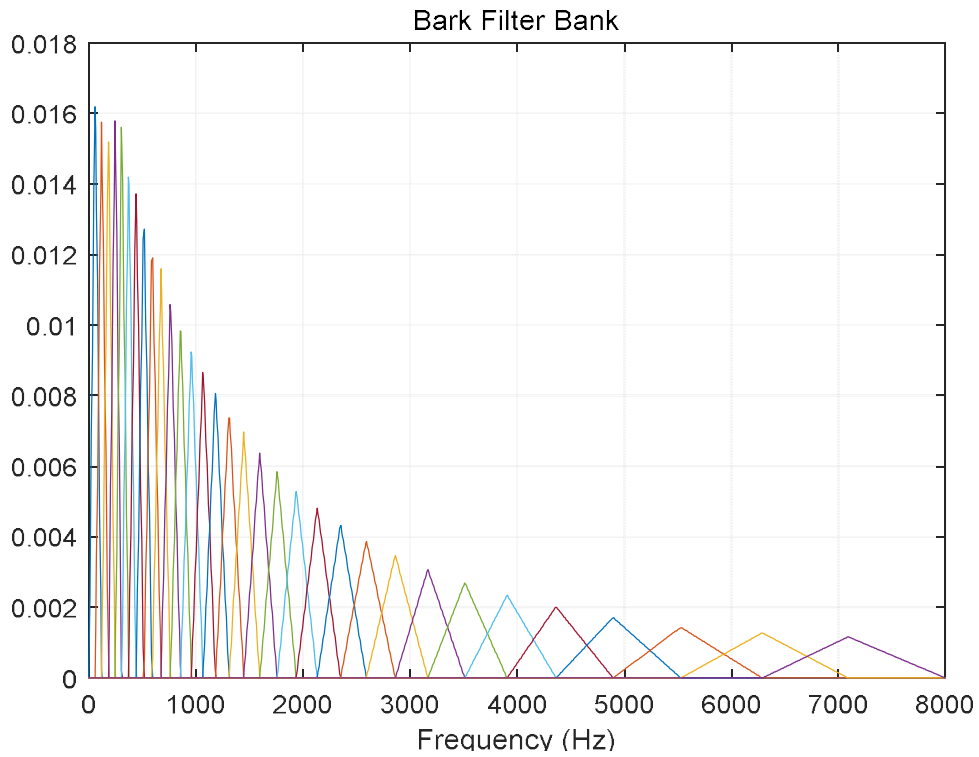
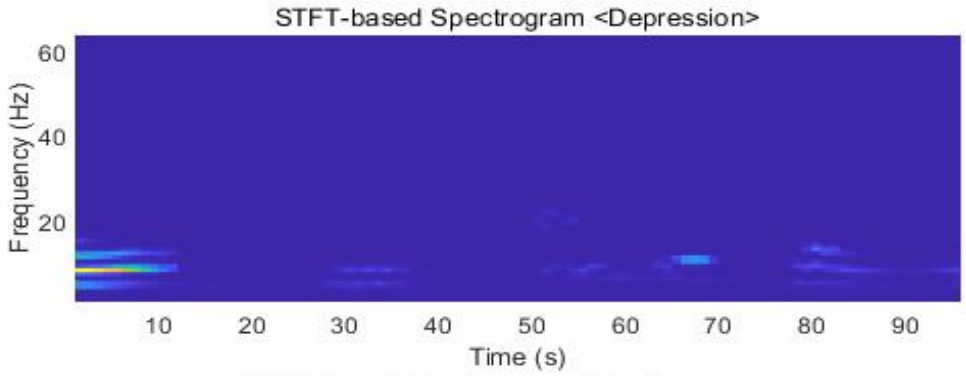
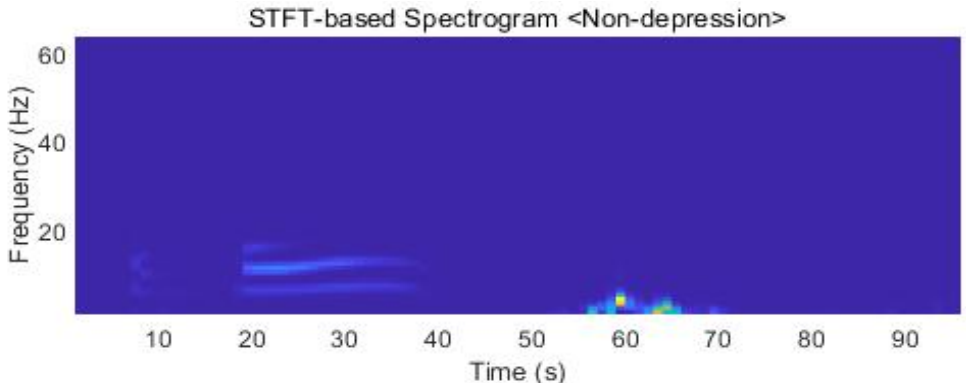


그림 3.5 바크 기반 청각 필터 뱅크 시각화

바크 주파수 척도를 기반으로 하여 청각 필터 뱅크를 설계할 주파수 범위, DFT를 계산하는 데 사용되는 점의 개수인 FFT 길이, 대역통과 필터의 수 등을 지정하여 그림 3.5와 같은 바크 필터 뱅크를 디자인할 수 있다. 오디오 신호에 단시간 푸리에 변환(Short Time Fourier Transform, STFT) 연산을 적용하여 스펙트로그램으로 변환한 뒤, 앞서 디자인한 필터뱅크와 곱해주면 바크 스펙트로그램을 얻을 수 있다. 그림 3.6은 각각 우울증과 비우울증 데이터에 대해 STFT 기반 스펙트로그램을 시각화한 것이고 그림 3.7은 각각 우울증과 비우울증 데이터에 대해 바크 척도 기반 스펙트로그램을 시각화한 것이다. 그림에서 확인할 수 있듯이 필터뱅크를 사용하지 않고 STFT만 적용한 스펙트로그램보다 바크 필터 뱅크를 사용한 스펙트로그램의 특징이 더 두드러지게 나타난다.

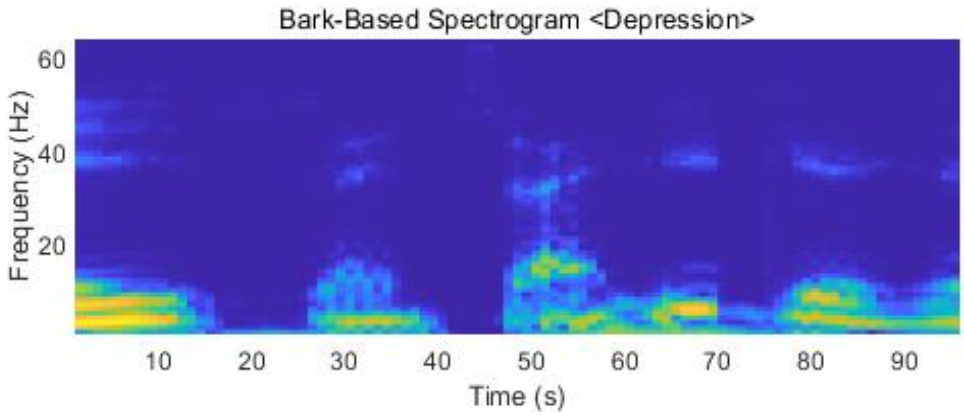


(a) 우울증 음성 데이터

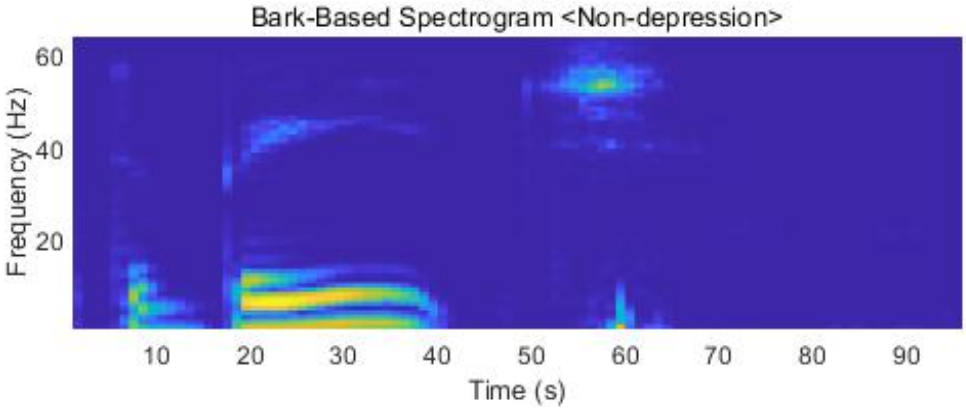


(b) 비우울증 음성 데이터

그림 3.6 STFT 기반 스펙트로그램 시각화



(a) 우울증 음성 데이터



(b) 비우울증 음성 데이터

그림 3.7 바크 척도 기반 스펙트로그램 시각화

나. ERB 스펙트로그램(ERB Spectrogram)

ERB 척도는 심리 음향학에서 사용되는 척도로, 사람의 청각에 있는 필터 대역폭에 대해 근사치를 제공한다. 이는 특정 필터의 최대 투과율과 동일한 통과 대역의 투과율을 가지며, 이론적으로 백색 소음 입력 시 실제 필터와 동일한 전력을 전달하고, 현실적이지 않지만 매우 유용한 직사각형 임계 대역으로 정의된다[39]. ERB의 정의는 수식 (3.4)와 같은 수학적 공식으로 정의된다. 여기서 $|G(f)|$는 필터 전송 함수를 나타내고, 이의 최댓값은 1이다. B. R. Glasberg[40]의 연구에서는 ERB 값에 대해 생리학적으로 동기가 부여된 몇 가지 공식이 제안되었고, 중심 주파수에서 ERB는 수식 (3.5)와 같은 공식을 따른다.

$$ERB = \int_0^\infty |G(f)|^2 df \tag{3.4}$$

$$ERB(f_r) = 24.7 + 0.108 f_r \tag{3.5}$$

ERB 개념을 기반으로 하여 그림 3.8과 같은 ERB 청각 필터 뱅크를 디자인하고, 이 필터 뱅크와 스펙트로그램의 곱 연산을 하게 된다. 그 후 본 논문에서 사용하고자 하는 ERB 스펙트로그램 특징을 추출할 수 있다. 그림 3.9에서는 우울증과 비우울증 음성 데이터에 대한 ERB 스펙트로그램 특징을 확인할 수 있다.

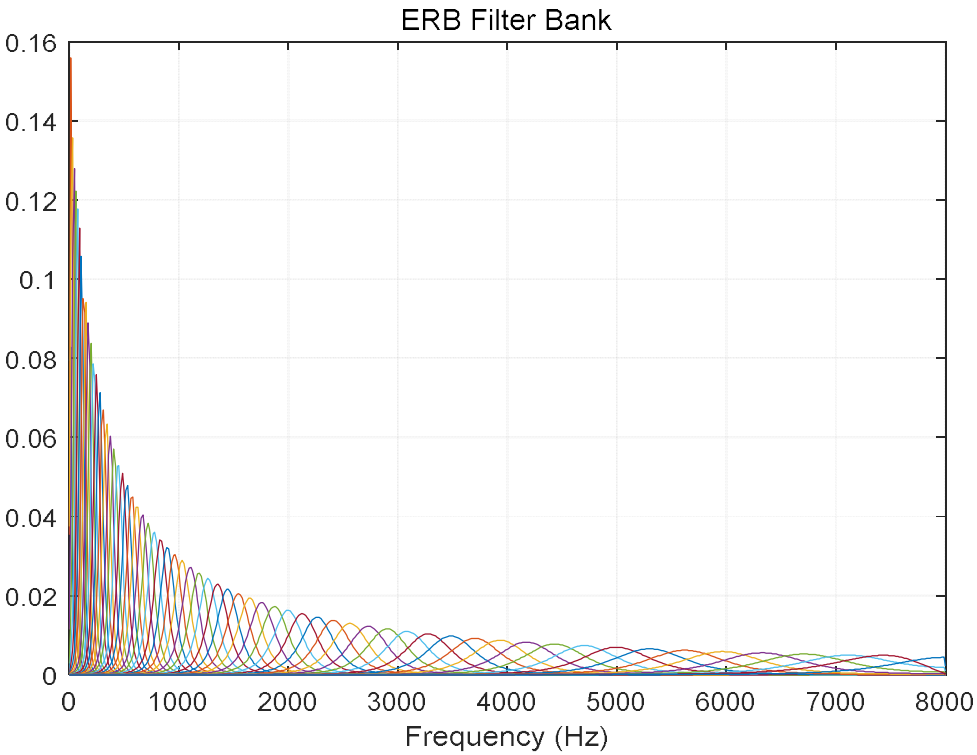
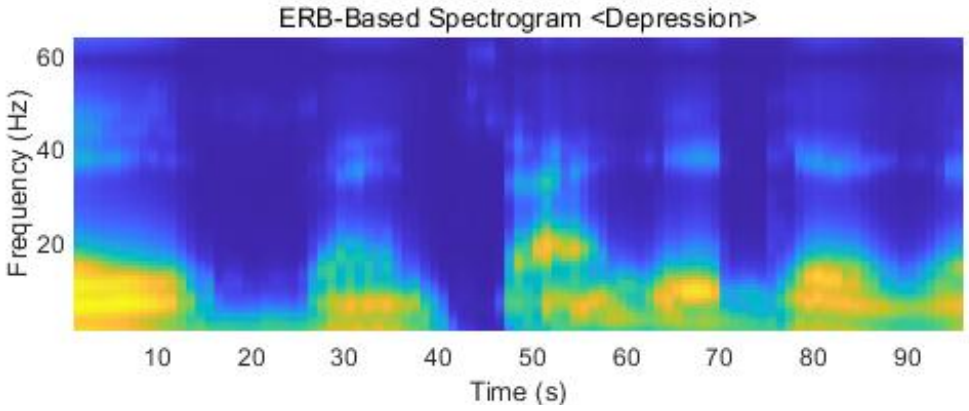
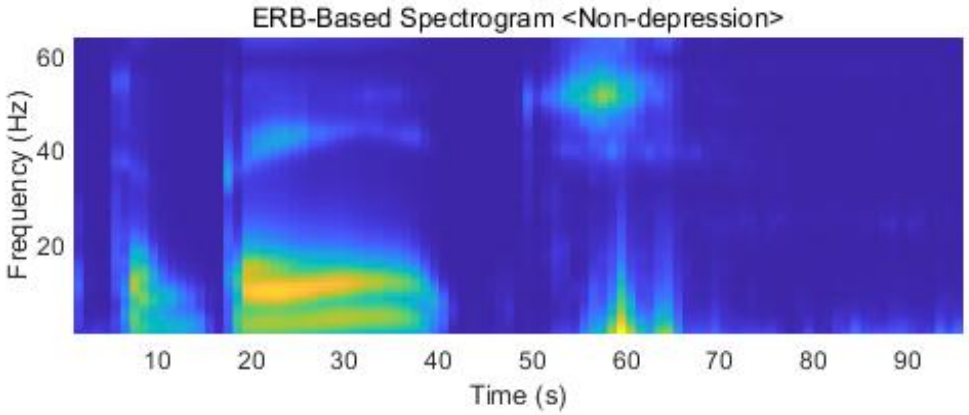


그림 3.8 ERB 기반 청각 필터 뱅크 시각화



(a) 우울증 음성 데이터



(b) 비우울증 음성 데이터

그림 3.9 ERB 척도 기반 스펙트로그램 시각화

다. 로그 멜 스펙트로그램(Log Mel Spectrogram)

사람이 음성신호를 인식할 때는 주파수에 따라 선형적으로 인식하는 것이 아닌, 멜 척도로 인식한다. 멜 척도는 사람의 청각 구조를 기반으로 하여 실제 주파수 정보를 수학적 형식으로 변환하는 방법이다. 이는 주파수 정보를 중요도에 따라 다르게 사용할 수 있어 음성처리 분야에서 다양하게 사용되고 있다.

로그 멜 스펙트로그램은 스펙트로그램 특징을 기반으로 만들어진다. 스펙트로그램은 음성신호를 프레임 단위로 나누어 시간 도메인에서 주파수 도메인으로 축변환시킨 다음, 이를 수평으로 쌓아 얻은 그래프를 뜻한다. 음성신호를 스펙트로그램으로 변환하면 매우 복잡한 오디오 신호를 각각의 주파수에서 해석할 수 있다. 따라서 로그 멜 스펙트로그램은 음성신호를 멜 척도로 변환하고 단시간 푸리에 변환을 적용하여 멜 스펙트로그램 특징을 추출한 뒤 이에 log 변환을 취하는 방법을 통해 얻어진다. log 변환을 취해 얻어지는 스펙트로그램 특징은 음성신호의 특징을 잘 나타낼 수 있어 음성처리 분야에서 다양하게 사용되고 있고 높은 성능을 나타낸다[41].

이러한 로그 멜 스펙트로그램 특징은 앞서 설명한 Bark, ERB 스펙트로그램과 마찬가지로 그림 3.10과 같은 멜 필터뱅크를 디자인한 뒤 단시간 푸리에 변환을 적용해 스펙트로그램을 얻어 멜 필터뱅크와 곱하고 log 변환을 취하면 추출할 수 있다. 그림 3.11에서는 각각 우울증과 비우울증에 대한 로그 멜 스펙트로그램 특징을 확인할 수 있다.

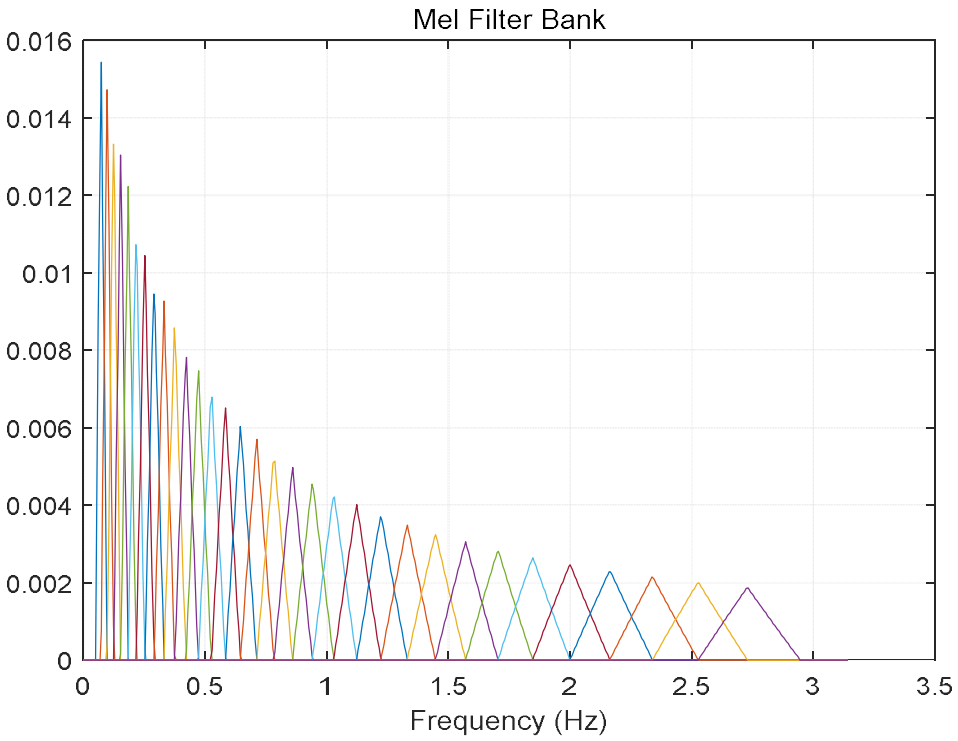
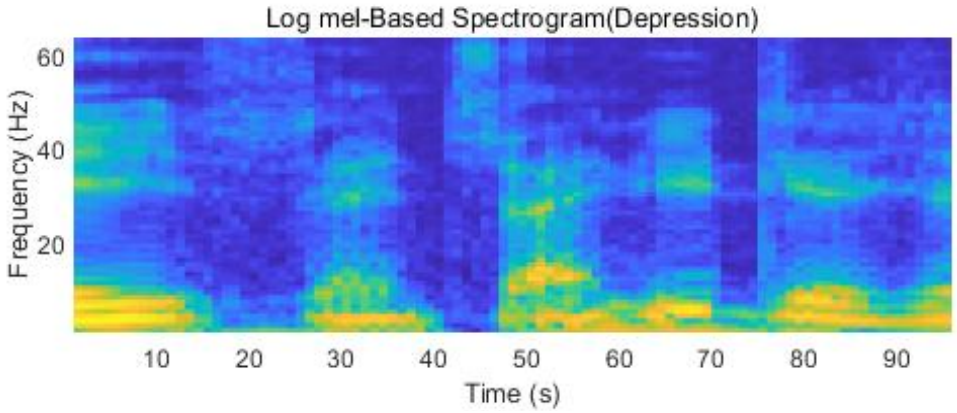
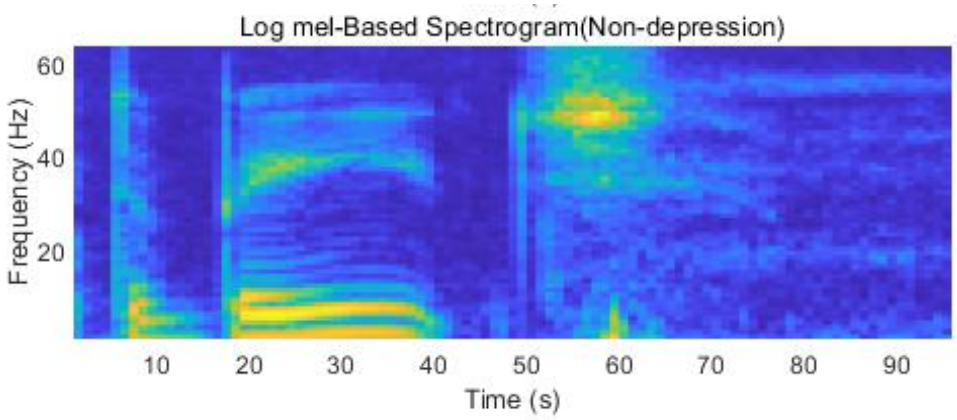


그림 3.10 멜 기반 청각 필터 뱅크 시각화



(a) 우울증 음성 데이터



(b) 비우울증 음성 데이터

그림 3.11 로그 멜 스펙트로그램 시각화

3. 1차원 음성신호 기반 Bi-LSTM 모델

가. LSTM(Long Short-Term Memory)

장단기 메모리(Long Short-Term Memory, LSTM)는 스스로 반복하면서 이전 단계에서 얻은 정보를 지속하는 순환 신경망(Recurrent Neural Network, RNN)의 한 종류이다. 이는 RNN의 장기의존성(long-term dependencies) 문제를 해결하기 위해 명시적으로 설계된 딥러닝 모델이고, 긴 의존 기간을 요구하는 학습을 수행할 수 있다. LSTM은 S. Hochreiter and J. Schmidhuber [42]의 연구에서 소개되었고, 그 후에 여러 추후 연구로 계속 발전하고 유명해지면서 여러 분야에서 널리 사용되고 있다.

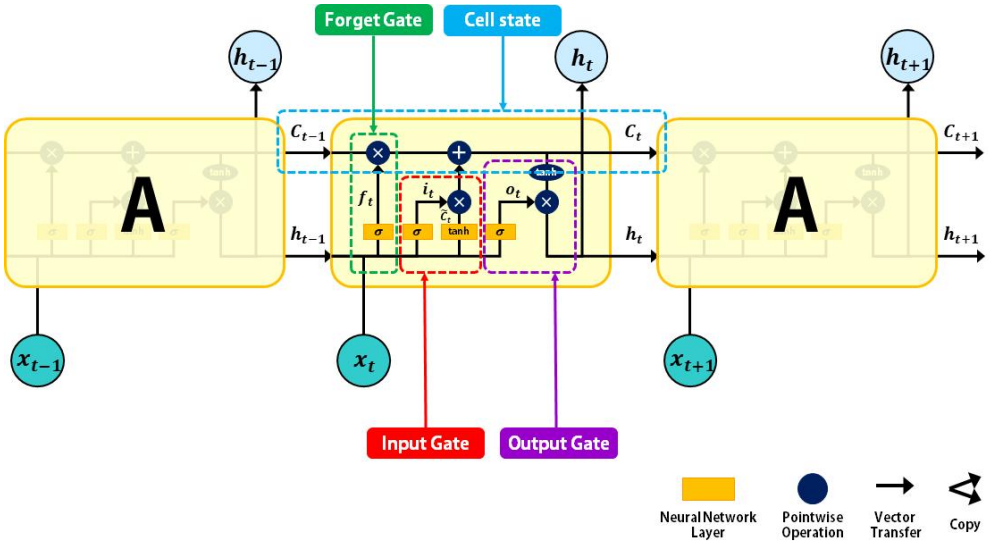


그림 3.12 LSTM 구조

모든 RNN은 neural network 모듈을 반복시키는 체인과 같은 형태를 지니고 반복 모듈을 단순한 neural network layer 한 층 구조로 구성되어있다. 그림 3.12에서 보면 LSTM도 똑같이 체인과 같은 구조로 되어 있지만, 각 반복 모듈은 4개의 layer가 특정한 방법으로 서로 정보를 주고받게 되어 있다. 이의 핵심적인 아이디어는 cell state를 기반으로 만들어졌다는 것이다. cell state는 컨베이어 벨트

역할을 해, 작은 선형 상호작용만을 적용해 전체 체인을 계속 작동시킨다. 이는 정보가 전혀 바뀌지 않고 그대로 흐르게 할 수 있다는 장점이 있다. 또한 LSTM은 cell state에 어떤 요소를 더하거나 없앨 수 있는 능력이 있는데, 이 능력은 gate라고 불리는 구조에 의해 제어된다. Gate는 정보가 전달될 수 있는 추가적인 방법으로, 시그모이드 계층과 pointwise 곱셈으로 이루어져 있다. LSTM의 단계는 총 4단계로 다음과 같다.

[단계 1] 망각 게이트 계층(Forget Gate Layer)

LSTM의 첫 단계는 cell state로부터 과거의 정보를 버릴 것인지 유지할 것인지 결정하는 것으로 망각 게이트 계층(Forget Gate Layer)이라고 한다. 수식 (3.6)은 시그모이드 함수를 적용한 망각 게이트 함수를 나타내고, 그림 3.13은 망각 게이트의 구조를 보여준다. 이는 현시점의 정보과 과거의 은닉층의 값에 각각 가중치를 곱하여 더한 후 시그모이드 함수를 적용하여, 그 출력값을 직전 시점의 cell에 곱해준다. 이 단계에서는 h_{t-1}와 x_t를 입력으로 받아 시그모이드 함수의 값인 (0, 1) 사이의 값을 C_{t-1} 에 보내주게 되는데, 그 값이 0이면 모든 정보를 버리는 것이고, 1이면 유지하는 것이다.

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{3.6}$$

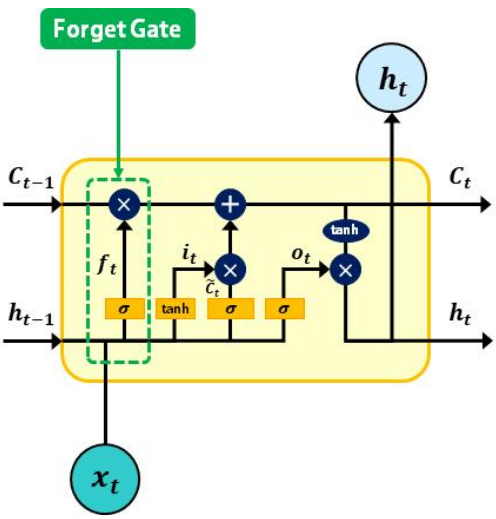


그림 3.13 LSTM의 망각 게이트 계층

[단계 2] 입력 게이트 계층(Input Gate Layer)

두 번째 단계는 현재 cell state 값에 어떤 정보를 저장할지를 결정하는 것으로 이는 입력 게이트 계층(Input Gate Layer)이라고 부른다. 그림 3.14에서 확인할 수 있는 입력 게이트는 현재 정보를 기억하기 위한 게이트를 뜻한다. 현시점에서 실제로 가지고 있는 정보가 얼마나 중요한지를 반영하여 셀에 기록하게 된다. tanh layer가 새로운 후보 값들인 $\tilde{C}_t$라는 벡터를 만들고, cell state에 더할 준비를 한다. 이렇게 나온 정보를 합쳐서 cell state 업데이트를 위한 요소를 준비한다. 이를 수식으로 정리하면 수식 (3.7), 수식 (3.8)와 같다.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.7}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.8}$$

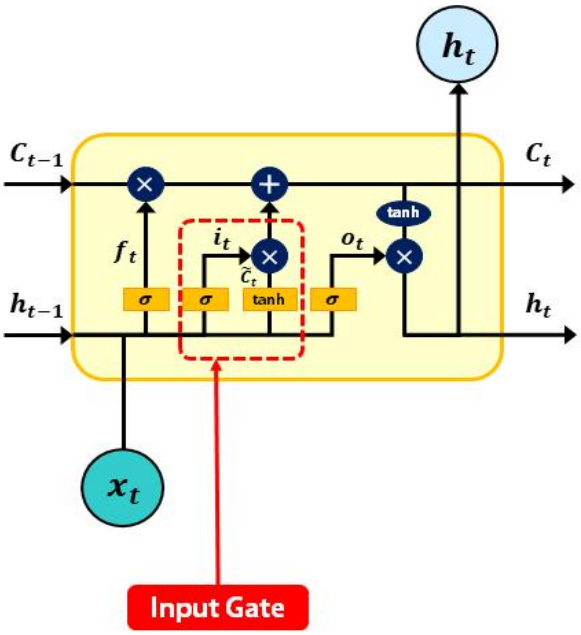


그림 3.14 LSTM의 입력 게이트 계층

[단계 3] cell state 업데이트

다음 단계는 앞에서 계산한 망각 게이트, 입력 게이트, 입력 후보를 이용하여 메모리 셀에 저장하여 cell state를 업데이트하는 것이다. 먼저 과거의 정보를 망각 게이트에서 계산된 만큼 잊고, 현시점의 정보 후보에 입력 게이트의 중요도를 곱해준 것을 더해 현시점을 기분으로 메모리 셀을 계산하여 cell state를 업데이트 한다. 이를 수식으로 나타내면 수식 (3.9)와 같고, 그림 3.15에서는 LSTM의 cell state update하는 과정을 간략하게 볼 수 있다. 본 수식에서 *는 pointwise 계산을 나타낸다.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3.9}$$

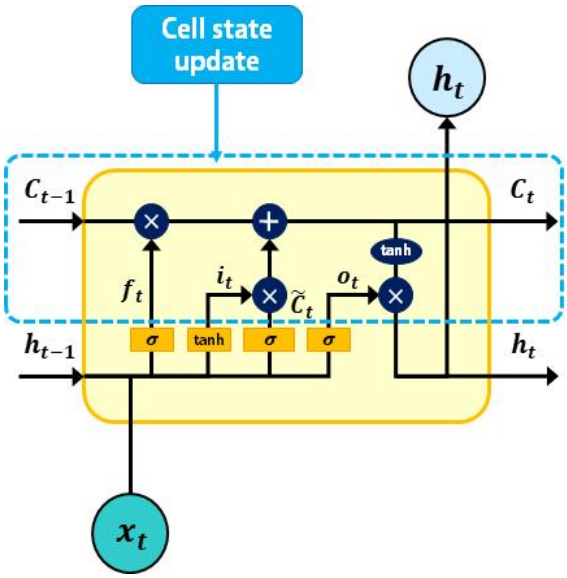


그림 3.15 LSTM의 cell state update

[단계 4] Output Gate Layer

마지막 단계는 어떤 출력값을 출력할지 결정하는 과정이다. 앞서 계산된 현시점의 메모리 셀을 현시점의 은닉층 값으로 출력할 양을 출력 게이트를 통해 결정한다. 먼저, 입력 데이터를 시그모이드 함수에 적용하여 어떤 부분의 cell state 출력으로 내보낼지 결정한다. 그리고 cell state를 tanh layer에 적용해 -1과 1 사이의 값을 받은 뒤에 앞서 계산된 sigmoid gate의 출력과 곱해준다. 그렇게 하면 사용자가 보내고자 하는 부분만 출력으로 내보낼 수 있게 된다. 수식으로 나타내면 수식 (3.10), 수식 (3.11)과 같다. 그림 3.16은 LSTM의 출력 게이트 계층 간단하게 도식화한 그림이다.

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{3.10}$$

$$h_t = o_t * \tanh(C_t) \tag{3.11}$$

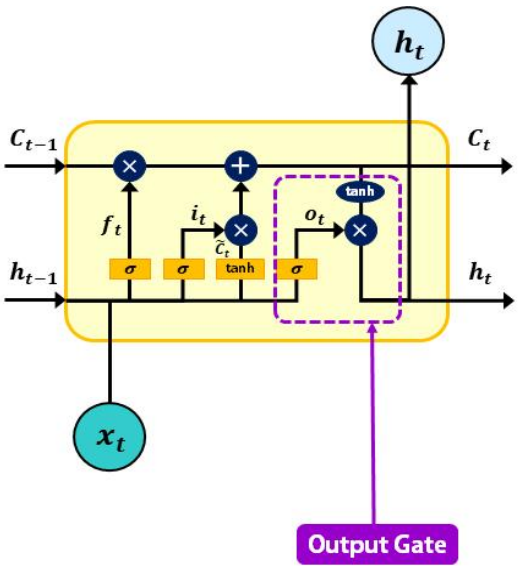


그림 3.16 LSTM의 출력 게이트 계층

나. Bi-LSTM(Bidirectional Long Short-Term Memory)

Bi-LSTM은 원래의 LSTM과 달리 입력이 양방향으로 들어가고, 양방향의 정보를 활용할 수 있는 LSTM이라고 할 수 있다. LSTM은 순서대로 학습을 진행하는 순방향 학습만 가능하고, Bi-LSTM 순방향 학습과 역방향 학습이 모두 가능하다.

그림 3.17에서 Bi-LSTM의 구조를 살펴보면, 노란색 블록의 순방향 LSTM이 있고 하늘색 블록의 역방향 LSTM이 나란히 있다. Bi-LSTM은 먼저 정보 흐름의 방향을 반대로 하는 LSTM 계층을 하나 더 추가한다. 간단히 말해서 입력 시퀀스가 추가 LSTM 계층에서 역방향으로 흐른다는 의미이다. 그런 다음 평균, 합, 곱셈 또는 연결과 같은 여러 방법으로 두 LSTM 계층의 출력을 결합한다. BI-LSTM은 일반적으로 작업에서 순서의 지정이 필요한 경우에 사용된다. 이러한 종류의 네트워크는 텍스트 분류, 음성인식 및 예측 모델에 사용할 수 있다.

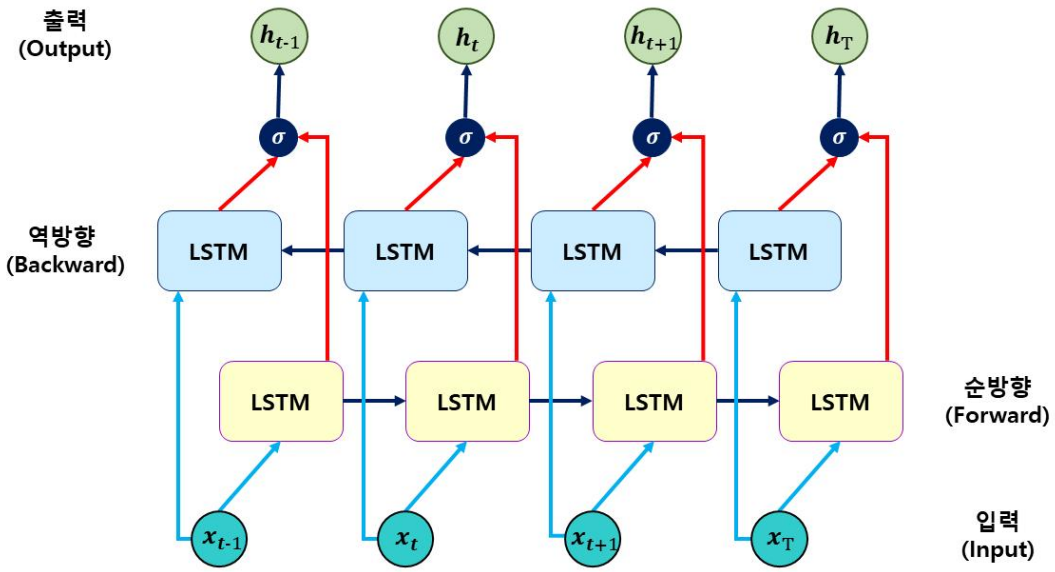


그림 3.17 Bi-LSTM 구조

본 연구에서는 1차원 음성신호의 특징을 추출하여 Bi-LSTM 모델을 기반으로 학습을 진행하고, 우울증 여부를 진단하는 모델을 제안한다. 제안된 Bi-LSTM 네트워크 구조는 그림 3.18에서 확인할 수 있다. 먼저 음성신호를 MFCC, GTCC 특징추출 방법을 이용하여 1차원적인 특징을 얻고, 시퀀스 데이터로 변환한다. 그 후

Bi-LSTM 계층을 2개로 쌓아 네트워크를 생성하고, 완전 연결 계층 소프트맥스 계층, 분류 계층을 추가하여 데이터 학습을 진행한다. 이때, 과적합(overfitting)을 방지하기 위해 Bi-LSTM 계층 사이에 30% 확률을 가지는 드롭아웃 계층을 추가해준다. 마지막으로 소프트맥스 계층에서 출력된 확률을 가지고 우울증인지 비우울증인지를 구분하는 우울증 여부를 진단한다.

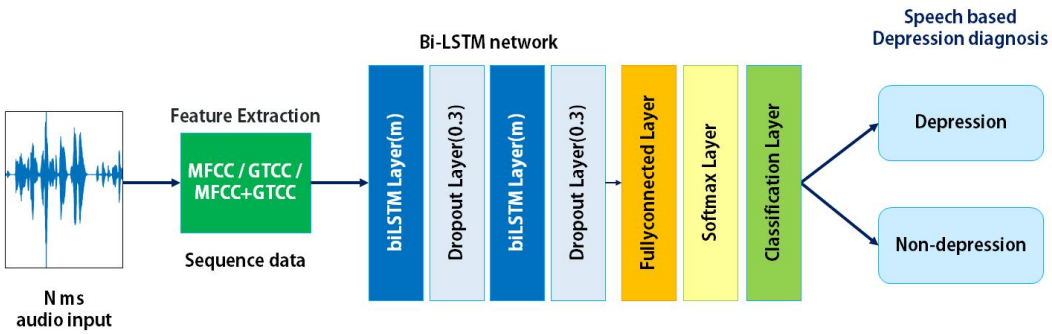


그림 3.18 음성신호를 이용한 제안된 Bi-LSTM 네트워크 구조

4. 2차원 시간-주파수 변환 기반 CNN 전이학습 모델

본 논문에서는 음성신호를 Bark, ERB, Log-Mel 스펙트로그램과 같은 2차원 시간-주파수 표현으로 변환한 이미지 특징을 CNN 기반 전이학습(transfer learning) 모델을 이용해 학습한다. 일반적으로 CNN 기반의 딥러닝 모델을 제대로 학습시키기 위해서는 많은 데이터가 필요하다. 하지만 많은 시간이 소요되고 큰 비용이 발생하기 때문에 많은 양의 데이터 세트를 구축하기는 쉽지 않다. 이러한 데이터 부족 문제를 해결할 방법의 하나가 바로 전이 학습이다. 전이 학습이란 대규모 데이터 세트, 일반적으로 대규모 이미지 분류 작업에서 학습되어 사전에 정의된 딥러닝 모델을 말한다. 이는 사전 학습된 모델을 그대로 사용하거나, 모델의 가중치를 불러와 사용자가 해결하고자 하는 과제에 맞게 재보정하여 사용한다. 전이학습 모델은 비교적 적은 수의 데이터를 가지고도 딥러닝 학습을 할 수 있어 효과적이며, 높은 정확도와 빠른 학습 속도를 제공한다.

가. VGGish

CNN 기반 전이학습 모델 중 하나인 VGGish은 S. Hershey[43]의 연구에서 유튜브에 있는 대규모 비디오 데이터베이스의 오디오 신호를 이용하여 신경망을 학습시켜 오디오 클래스를 분류하기 위해 제안된 딥러닝 신경망이다. 이는 527개의 오디오 클래스를 포함하고 있는 유튜브 동영상 200만 개 이상으로 구성된 오디오 콘텐츠를 사용하여 신경망이 학습되었다. 성인 남자와 성인 여자의 목소리, 아기의 옹알이, 동물 소리 등이 527개의 클래스에 포함된다[44]. VGGish 모델은 이미지 분류를 위해 컴퓨터 비전 분야에서 많이 사용되어왔던 VGG를 기반으로 네트워크 아키텍처가 형성되어 있다. 그림 3.19는 VGGish 네트워크 아키텍처를 간단하게 보여준다. VGGish는 오디오 클립의 96*64*1의 크기를 가지는 스펙트로그램 기반 특징을 입력으로 받고, 4개의 Convolution 블록으로 구성되어있다. 각각의 블록은 특징 추출기 역할을 하는 2d 기반 Convolution 계층, 활성화 함수인 ReLu, 이미지의 특징은 유지하고 차원을 줄이는 Max Pooling 계층을 포함한다. 다음으로 분류기 역할을 하는 2개의 완전 연결 계층, 임베딩 계층, 회귀 출력계층이 뒷부분에 포함된다.

본 연구에서는 음성신호를 시간-주파수 2차원 특징으로 변환해 CNN 기반 전이학습 모델을 이용하여 학습을 진행한 후 우울증 여부를 진단한다. 우울증의 여부를 진단하기 위해 2개의 클래스를 분류하는 모델을 구성해야 하므로 원래의 완전연결 계층 뒤에 새로운 완전 연결 계층을 추가하여 분류 클래스 수를 2로 변경하였다. 또한 회귀 출력계층을 분류 출력계층으로 교체하여 우울증 데이터를 분류할 수 있도록 분류 모델을 생성하였다.

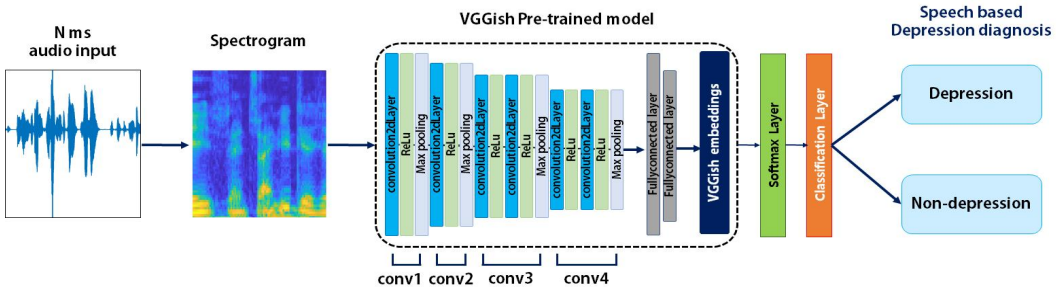


그림 3.19 VGGish 네트워크 아키텍처 구조

나. YAMNet

YAMNet 모델은 200만 개 이상의 유튜브 비디오 AudioSet에서 웃음소리, 강아지가 짖는 소리 또는 사이렌 소리 등과 같은 521개 클래스를 가진 오디오에 대해 훈련된 음향 감지 모델이다. YAMNet 모델은 Ellis와 Chowdhry에 의해 개발되었고, AudioSet 말뭉치에 대한 오디오 이벤트 분류를 위한 계산적으로 효율적인 모델이다. VGGish 모델에는 7,200만 개 이상의 매개변수를 사용하여 계산이 복잡하다는 단점이 있다. 그에 비해 YAMNet 모델은 470만 개의 매개변수만을 사용하기 때문에 VGGish 모델보다는 계산적인 부분에서 효율이 높다. 이는 깊이별로 분리할 수 있는 convolution 커널을 이용해 경량 모델을 만들어 컴퓨터 비전 분야에서 사용할 수 있도록 하였고, H. A. Andrew et al.[45]의 연구에서 제안된 MobileNet 아키텍처를 기반으로 생성되었다. YAMNet은 96*64*1의 크기를 가지는 스펙트로그램 이미지를 입력으로 받고, 14개의 convolution 계층 블록으로 구성되어있으며, 첫 번째 계층을 제외한 모든 계층은 깊이별 convolution 커널을 기반으로 한다. 본 연구에서는 마지막 convolution 계층 뒷부분에 완전 연결 계층을 추가해 클래스 개수를 다시 정해주고, 분류 계층을 교체함으로써 전이학습 모델을 보정해주었다.

다. OpenL3

다양한 유형의 전이학습 모델 중 사람의 목소리를 인식할 수 있도록 훈련된 VGGish 및 YAMNet 모델 이외에도, 음악과 환경 소리를 식별하도록 최적화된 OpenL3 모델을 음성 우울증 데이터 학습에 사용한다. 본 모델을 사용하는 이유는 음악과 환경 소리를 인식하도록 훈련된 모델이 우울증 데이터에 있는 음성 병렬 언어의 특성을 인식하는 데 사용될 수 있는지를 확인하기 위해서이다. J. Cramer[46]의 연구에서는 R. Arandjelovic and A. Zisserman[47]의 연구에서 제안된 L3(Look, Listen, Learn) 개념을 기반으로 하여 OpenL3 모델을 개발하였다. 추가로 단시간 푸리에 변환 기반 스펙트로그램 또는 멜 스케일 주파수 기반 스펙트로그램과 딥 어쿠스틱 임베딩을 위한 다른 크기 사이의 종류와 같은 다양한 네트워크 아키텍처 종류를 조사하였다. 여기서 OpenL3 모델을 훈련하는 데 사용된 비디오가 AudioSet 코퍼스 내에서 선별되었지만 VGGish 모델과 YAMNet 모델보다 분류 클래스가 훨씬 적다는 것이 중요하다. OpenL3 모델의 네트워크 아키텍처는 128*199*1의 크기를 가지는 스펙트로그램 입력을 받아 특징 추출기로 사용되는 4개의 Convolution 계층 블록으로 구성된다. MaxPooling 계층의 연산은 512에서 6,144의 크기를 가지는 임베딩을 생성하는 옵션을 사용하여 특징 추출기 출력에서 수행된다.

제2절 텍스트 데이터를 이용한 딥러닝 모델 설계

1. 워드 임베딩(word embedding)

텍스트를 컴퓨터가 이해하고, 효율적으로 처리하게 하기 위해서는 컴퓨터가 이해할 수 있도록 텍스트의 단어를 숫자로 적절하게 변환해야 한다. 현재는 각 단어를 신경망을 통해 벡터화하는 워드 임베딩(word embedding)이라는 방법이 가장 많이 사용되고 있다.

원-핫 인코딩을 통해서 나온 원-핫 벡터들은 표현하고자 하는 단어의 인덱스 값만 1이고, 나머지 인덱스에는 전부 0으로 표현되는 벡터이다. 이렇게 벡터 또는 행렬값 대부분이 0으로 표현되는 방법을 희소 표현(sparse representation)이라고 한다. 원-핫 벡터는 희소 벡터(sparse vector)라고도 한다. 이러한 희소 벡터의 문제점은 단어의 개수가 늘어나면 벡터의 차원이 계속해서 커진다는 점이다. 텍스트 데이터를 원-핫 벡터로 표현할 때 포함된 단어가 50,000개였다면 벡터의 차원도 똑같이 50,000이고, 그중 단어의 인덱스에 해당하는 부분만 1이고 나머지는 모두 0의 값을 가진다. 이는 공간적 낭비를 일으키고, 단어의 의미를 표현하지 못한다는 단점이 있다.

희소 표현과 반대되는 표현으로 밀집 표현(dense representation)이라는 개념이 있다. 이는 벡터의 차원을 단어 집합의 크기로 가정하지 않고, 사용자가 지정한 값으로 모든 단어의 벡터 표현의 차원을 맞춘다. 또한, 이러한 단어 벡터 표현은 0과 1만 가진 값이 아니라 실숫값을 가진다. 사용자가 밀집 표현의 차원을 256으로 설정한다면, 모든 단어의 벡터 표현의 차원은 256으로 바뀌면서 모든 값이 실수가 된다. 이 경우 벡터의 차원이 조밀해졌다고 하여 밀집 벡터(dense vector)라고 한다. 이렇게 단어를 밀집 벡터의 형태로 표현하는 방법을 워드 임베딩이라고 한다. 이 밀집 벡터를 워드 임베딩 과정을 통해 나온 결과라고 하여 임베딩 벡터(embedding vector)라고도 말한다. 워드 임베딩을 통해 벡터를 기반으로 하는 비정형 문장을 분석할 수 있고, 추론을 기반으로 최적의 문장을 선택할 수 있을뿐더러 의미를 기반으로 유사 단어를 지정할 수 있다. 이는 단어를 군집화하고 벡터 연산을 통해 단어 간의 관계를 파악하여 추론이 가능해짐에 따라 자연어 처리 모델링에 필수로 사용되고 있는 개념이다.

2. 텍스트 데이터를 이용한 딥러닝 모델

가. Bi-LSTM 기반 우울증 진단모델

텍스트를 분류를 위한 첫 번째 딥러닝 모델은 Bi-LSTM을 기반으로 한 우울증 진단모델이다. Bi-LSTM 신경망은 시퀀스 데이터의 시간 스텝 간의 장기적인 종속성을 학습할 수 있는 RNN의 일종이고, 텍스트 데이터는 본질적으로 순차적인 데이터이기 때문에 Bi-LSTM을 통해 처리할 수 있다. Bi-LSTM의 입력으로 텍스트 데이터를 넣기 위해서는 먼저 텍스트 데이터를 숫자형 시퀀스로 변환해야 한다. 따라서 단어 인코딩을 이용하여 텍스트를 숫자형 인덱스를 가지는 시퀀스로 매핑하였다. 단어를 표현하는 방법에 따라 신경망의 성능이 달라지는데, 본 논문에서는 더 좋은 결과를 얻기 위해 신경망에 워드 임베딩 계층을 포함시켜 단어 모음에 있는 모든 단어를 스칼라형 인덱스가 아닌 숫자형 벡터로 매핑하였다. 이는 비슷한 의미를 갖는 단어들이 비슷한 벡터를 갖도록 단어 의미 정보를 수집하게 되고, 벡터 연산을 통해 단어 사이의 관계를 모델링 할 수 있다.

제안된 텍스트 데이터를 이용한 Bi-LSTM 기반 우울증 진단모델의 구조는 그림 3.20에서 볼 수 있다. 먼저 텍스트 데이터를 가져오고 전처리를 통해 얻은 텍스트를 단어 인코딩을 통해 숫자형 시퀀스로 변환한다. 시퀀스 입력 계층 다음에 워드 임베딩 계층, 은닉유닛의 수가 10개인 BI-LSTM 계층을 2개 쌓고, 30%의 확률을 가지는 드롭아웃 계층을 배치하여 Bi-LSTM 신경망을 설계하고 완전 연결 계층, 소프트맥스 계층을 차례대로 배치하여 소프트맥스 계층에서 출력된 확률을 이용하여 우울증을 진단하게 된다.

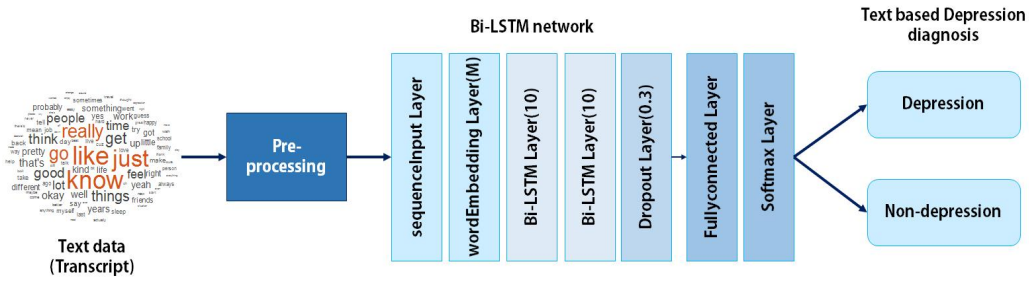


그림 3.20 텍스트 데이터를 이용한 Bi-LSTM 기반 우울증 진단모델 구조

나. CNN 기반 우울증 진단모델

텍스트를 분류를 위한 두 번째 딥러닝 모델은 합성곱 신경망(Convolutional Neural Networks, CNN)을 기반으로 한 모델이다. CNN을 이용하여 텍스트 데이터를 분류하려면 입력 데이터의 시간 차원에 대해 합성곱 연산을 취하는 1차원 합성곱 계층을 사용해야 한다. 제안된 모델은 다양한 너비를 가지는 1차원 합성곱 필터를 사용하여 네트워크를 학습시킨다. 합성곱 계층의 필터 너비는 각각 필터가 볼 수 있는 단어의 수(n-gram 길이)에 해당한다. n-gram이란 연속적인 n개의 단어로 구성된 것으로, 횟수를 사용하여 단어를 벡터로 표현하는 방법이다. 이는 한 단어 이상의 단어 시퀀스를 분석 대상으로 하고 n-gram 앞에 있는 n의 수에 따라 단어 시퀀스를 몇 개의 단어로 구성할지 결정한다. n=1, 2, 3인 경우를 각각 Uni-gram, Bi-gram, Tri-gram이라 부르며 그림 3.21을 보면 이를 통해 단어 시퀀스가 어떻게 구성되는지 확인할 수 있다. n-gram은 다음 단어를 예측하고, 오타를 발견할 수 있고, 단어의 순서를 무시하는 bag of words 방법의 단점을 보완할 수 있는 장점이 있다.

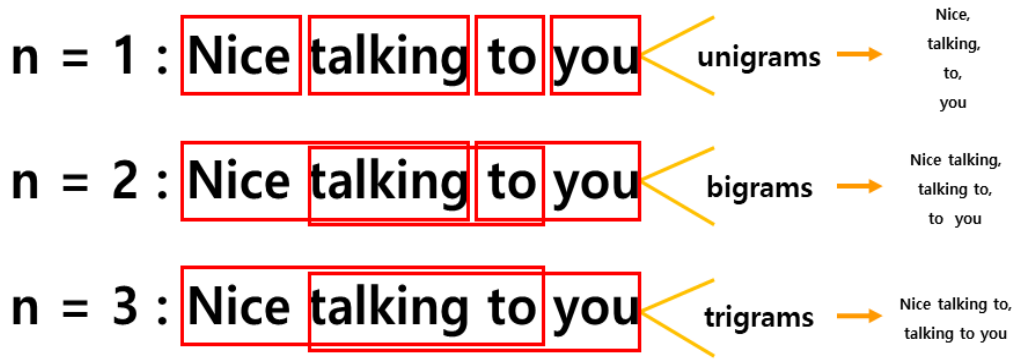


그림 3.21 n-gram을 이용한 단어 시퀀스 구성

이러한 n-gram 방법을 이용하여 네트워크를 구성했고, 제안된 CNN 기반 텍스트 분류 모델의 네트워크는 그림 3.22에서 볼 수 있다. 먼저 텍스트를 가져오고 전처리를 통해 얻은 텍스트 단어를 숫자형 시퀀스로 변환하고 시퀀스 입력 계층과 워드 임베딩 계층의 입력으로 넣는다. 그 후 합성곱 네트워크가 2개로 나뉘어 있고, 각각 2와 3의 값을 가지는 n-gram의 길이를 사용한다. 각각의 합성곱 네트워크는

합성곱 계층, 배치정규화 계층, 활성화 함수인 ReLu 계층, 20%의 확률을 가지는 드롭아웃 계층, 전역최대풀링 계층이 차례로 배치되어 있다. 마지막으로 완전 연결 계층과 소프트맥스 계층을 이용하여 우울증 여부를 분류하게 된다.

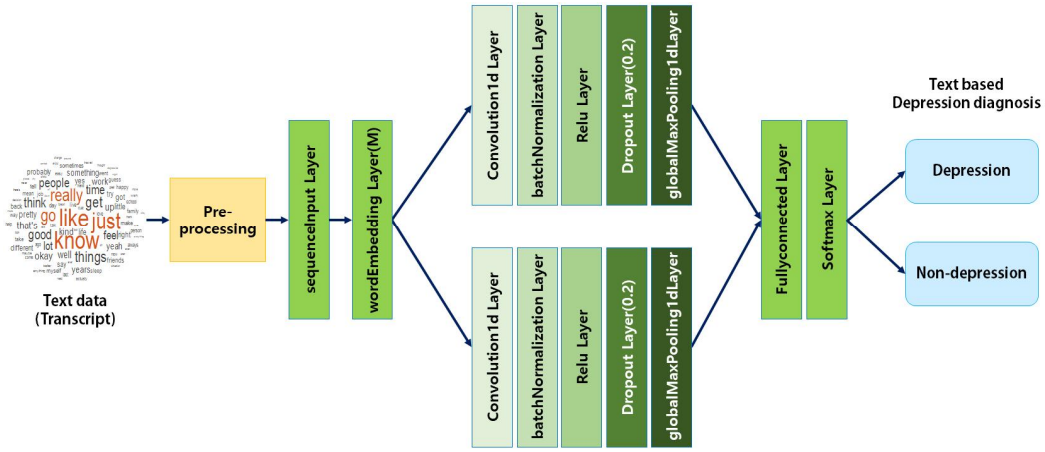


그림 3.22 텍스트 데이터를 이용한 CNN 기반 우울증 진단모델 구조

제3절 음성 및 텍스트 데이터를 이용한 4-stream 기반 딥러닝 모델 설계 및 우울증 진단

1. 멀티모달 및 Late score fusion 방법

가. 멀티모달(Multi Modal)

멀티모달이란 여러 가지 형태와 의미로 컴퓨터와 대화 환경을 의미한다. 멀티모달에서 모달은 모달리티(modality)를 의미하는데 모달리티는 양방향 통신 과정에서 사용되는 의사소통 채널을 말한다. 멀티 모달 인터페이스는 전통적으로 텍스트 외에 음성, 제스처, 시선, 표정, 생체신호 등 여러 입력 방식을 융합하여 인간과 컴퓨터 사이에 자연스러운 의사소통이 가능한 사용자 친화형 기술이다. 과거에는 기계가 이해하기 쉬운 형태로 입력을 줬다면 최근에는 사용자가 이해하기 쉬운 형태로 컴퓨터에게 입력을 전달하는 형태로 발전한 것이다. 이러한 멀티모달을 이용하면 사람의 여러 신체 부위에 컴퓨터와 소통할 수 있는 모달 장치를 부착하고 해당 장치들을 통해 행동 분석, 감정 분석 등을 할 수 있다. 즉, 사람과 컴퓨터를 연결하여 데이터를 수집하고 분석할 수 있는 것이다. 이러한 데이터 수집을 기반으로 불규칙적인 사람의 여러 감정과 행동에 대한 데이터를 수집할 수 있는 모델을 구현할 수 있다.

나. Late score fusion 방법

기계학습/딥러닝에서 late score fusion은 가장 일반적으로 간단하게 사용할 수 있는 융합 방법이다. 이 접근 방식은 독립적으로 훈련할 수 있는 딥러닝 모델들을 이용하여 서로 다른 양식에서 출력된 공통된 값을 결합하여 최종값을 얻을 수 있다[48]. 이는 서로 다른 데이터를 이용하여 딥러닝 모델에 각각 학습시킨 후 마지막 계층인 소프트맥스의 점수를 하나의 분류 값으로 융합하는 방법이다. 소프트맥스 계층의 점수를 더하거나 곱하여 그 값의 최대 또는 평균 점수를 취하여 최종 출력값을 얻고 이를 클래스를 분류하는 데 사용하게 된다. late score fusion 방법은 그림 3.23에서 간략하게 살펴볼 수 있다.

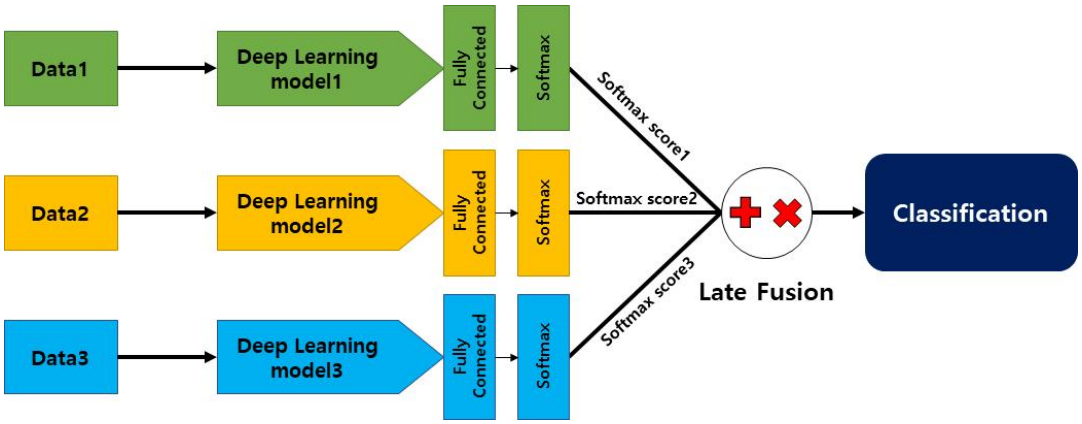


그림 3.23 late score fusion 방법

2. 음성 및 텍스트 데이터로부터 우울증 진단을 위한 4-stream 기반 딥러닝 모델

본 논문에서는 멀티모달 데이터 중 음성 및 텍스트를 이용하여 Bi-LSTM과 CNN을 통해 학습시킨 후 그에 대한 소프트맥스 점수를 late fusion 하여 우울증을 진단하는 4-stream 기반 딥러닝 모델을 설계한다. 1차원 음성 데이터는 스펙트럼의 배음 구조를 유추할 수 있도록 MFCC, GTCC와 같은 특징을 추출해 Bi-LSTM 모델을 통해 학습시킨다. 또한 시간에 따라 변하는 주파수 특성을 고려하기 위해 Bark, ERB, Log-Mel 스펙트로그램과 같은 2차원 특징을 추출하여 VGGish, YAMNet, OpenL3와 같은 CNN 기반 전이학습 모델을 통해 학습시킨다. 그 후 각각의 소프트맥스 점수를 얻는다. 텍스트 데이터는 Bi-LSTM 신경망에 텍스트를 숫자형 시퀀스로 변환한 후 단어를 벡터로 매핑하는 wordEmbedding 계층을 추가하고, CNN 모델은 시퀀스 벡터를 입력으로 받는 1차원 기반 CNN 신경망을 생성하여 학습을 진행하여 각각의 소프트맥스 점수를 얻는다. 마지막으로 우울증 진단을 위해 음성 및 텍스트 데이터를 4개의 딥러닝 모델에 학습시켜 나온 각각의 소프트맥스 값을 late score fusion 방법을 이용해 모두 합하거나 곱하여 최댓값을 취한다. 그 후 최종 출력값과 분류 성능을 얻고 우울증과 비우울증으로 이진 분류하여 우울증 여부에 대해 진단한다. 그림 3.24는 제안된 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단모델의 구조를 보여준다.

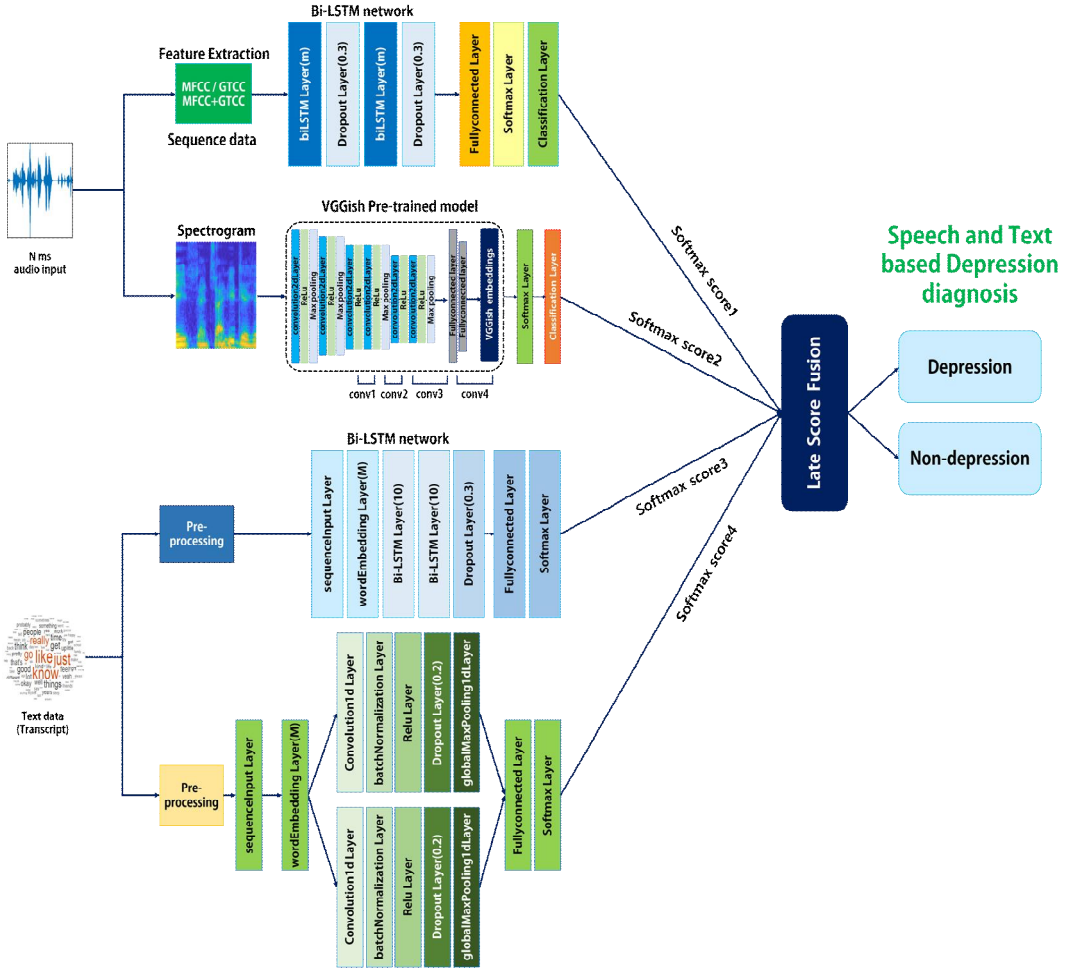


그림 3.24 제안된 Bi-LSTM과 CNN의 4-stream 기반 우울증 진단모델의 구조

.

제4장 실험 및 결과분석

제1절 Extended DAIC-WOZ 우울증 데이터베이스

EDAIC-WOZ(Extended Distress Analysis Interview Corpus Wizard of Oz) 우울증 데이터 세트는 DAIC-WOZ(Distress Analysis Interview Corpus Wizard of Oz) 데이터 세트의 확장된 버전이다. 본 데이터는 AVEC 2019(The 2019 Audio/Visual Emotional Challenge)에서 사용된 우울증 데이터 세트이다[49]. DAIC-WOZ 데이터 세트는 더 큰 데이터 세트인 DAIC(Distress Analysis Interview Corpus)의 일부로, 불안, 우울증, 외상 후 스트레스 장애(Post-Traumatic Stress Disorder, PTSD)와 같은 사람들의 심리적인 고통 상태를 진단하는 데 도움을 주기 위해 설계되었다. 본 데이터 세트는 참가자들과의 임상 인터뷰를 통해 정신질환의 언어적 및 비언어적인 지표를 식별하는 컴퓨터 에이전트를 만들기 위한 더 큰 노력의 일환으로 수집되었다[49-50]. 임상 인터뷰 속에는 참가자들의 음성, 비디오, 텍스트(스크립트) 파일이 포함되어있다.

그림 4.1 가상 인터뷰진행자 'Ellie'와의 인터뷰

DAIC-WOZ 데이터 세트에서 임상 인터뷰는 그림 4.1에서와 같이 참가자와 애니메이션 캐릭터로 만들어진 'Ellie'라는 가상 인터뷰진행자, 이를 제어하는 사람을 통해 진행된다. 이때 참가자는 우울증 증상을 가지고 있는 사람들과 가지고 있지 않은 정상적인 사람이 모두 포함된다. EDAIC-WOZ 데이터 세트에는 DAIC-WOZ에서와 같이 가상 인터뷰진행자인 'Ellie'와의 인터뷰를 포함할 뿐만 아니라 AI 제어 에이전트를 사용하여 수집된다. 이 에이전트는 다양하고 자동화된 인식 및 행동 생성 모듈을 사용하여 완전히 자율적으로 동작한다. [P300-P492] 범위의 ID를 가진 세션의 189명의 사람들은 가상 인터뷰진행자인 'Ellie'와의 인터뷰를 통해 수집되었고 [P600-P718] 범위의 ID를 가진 세션의 86명의 사람들은 AI 제어 에이전트를 통해 수집되었다[51].

그림 4.2와 그림 4.3에서는 가상 인터뷰진행자 'Ellie'와 AI 제어 에이전트를 이용한 EDAIC-WOZ 데이터 세트 수집 방법을 간단하게 보여주고 있다. 참가자들은 데이터 수집 전 우울증 설문조사 중 하나인 PHQ-8(Patient Health Questionnaire depression scale) 설문지를 통해 사전 설문조사를 진행하여 우울증 여부를 판단하였다. PHQ-8 응답 설문지는 설문지를 작성하는 사람이 지난 2주 동안 9개의 DSM(Diagnostic and Statistical Manual of Mental Disorders) 기준 증상 중 8개를 경험한 일수를 기준으로 점수를 매겨 우울증의 척도를 표준화하였다[52]. 이는 각 항목의 점수를 합산하여 0에서 24점 사이의 총점을 산출한다. 표 4.1을 보면 총점이 0~4점은 우울 증상이 없음, 5~9점은 가벼운 우울 증상이 있음, 10~14점은 중등도의 우울 증상이 있음, 15~19점은 중증의 심각한 우울 증상이 있음, 20~24점은 심각한 우울 증상이 있음을 나타낸다. 본 연구에서는 우울증 여부를 진단하기 위해 PHQ-8의 점수가 10점 미만이면 증상이 없는 정상적인 사람으로, 10점 이상이면 우울증 증상이 있는 사람으로 구분하여 실험을 진행하였다. [P300-P492] 범위의 ID를 가진 189명의 사람들은 그림 4.2와 같이 PHQ-8 설문조사가 끝난 참가자들은 가상 인터뷰 진행자 'Ellie'와 인터뷰를 진행하게 된다. 'Ellie'는 제 3자에 의해 제어되고 인터뷰가 진행되는 동안 참가자들은 카메라를 통해 자세와 얼굴 데이터를 마이크를 통해 음성과 텍스트 데이터를 수집하게 된다. [P600-P718] 범위의 ID를 가진 86명의 사람들은 그림 4.3과 같이 AI 제어 에이전트와의 인터뷰를 통해 자세, 얼굴, 음성, 텍스트 데이터를 수집하게 된다.

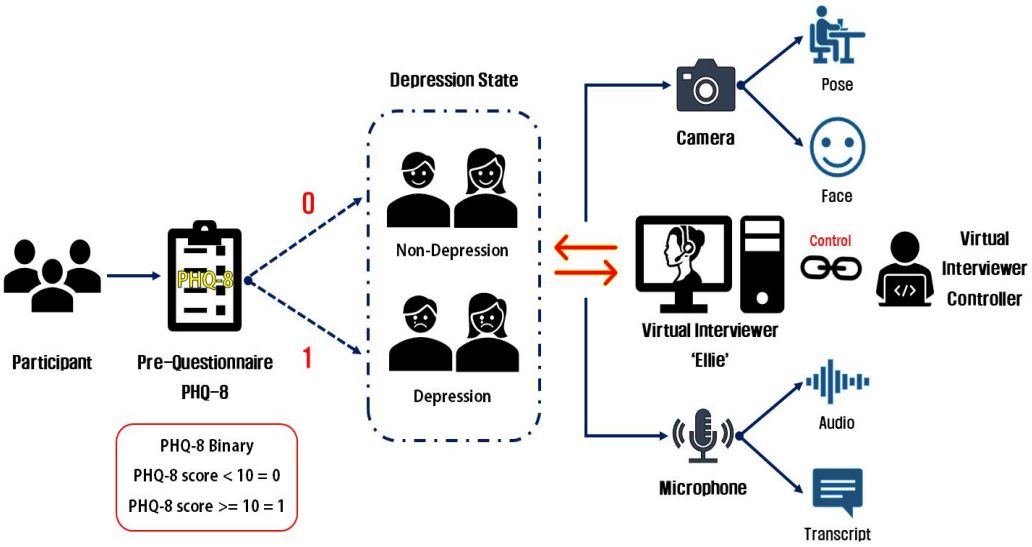


그림 4.2 가상 인터뷰진행자 'Ellie'를 통한 데이터 수집

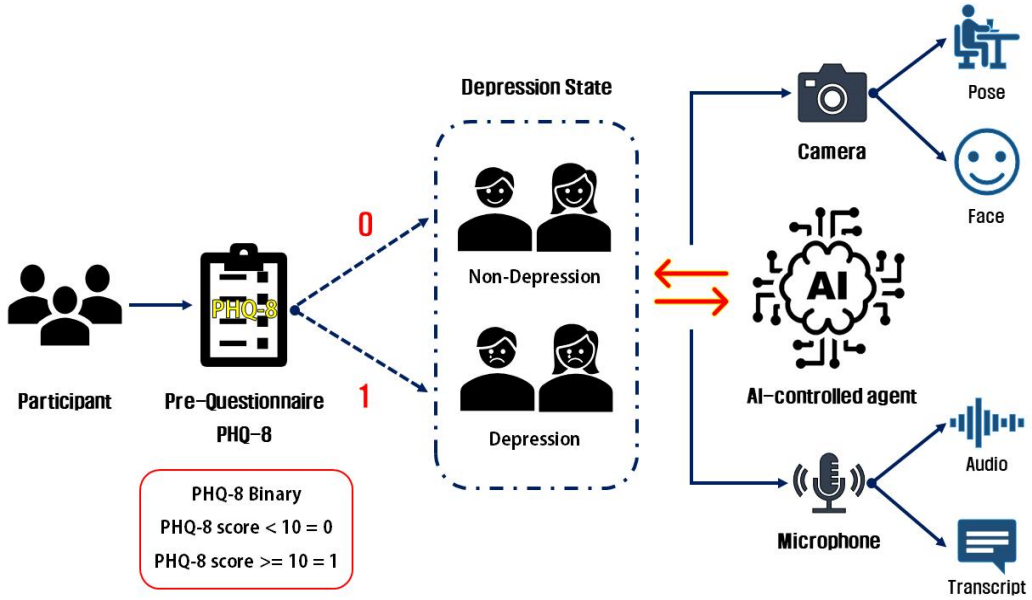
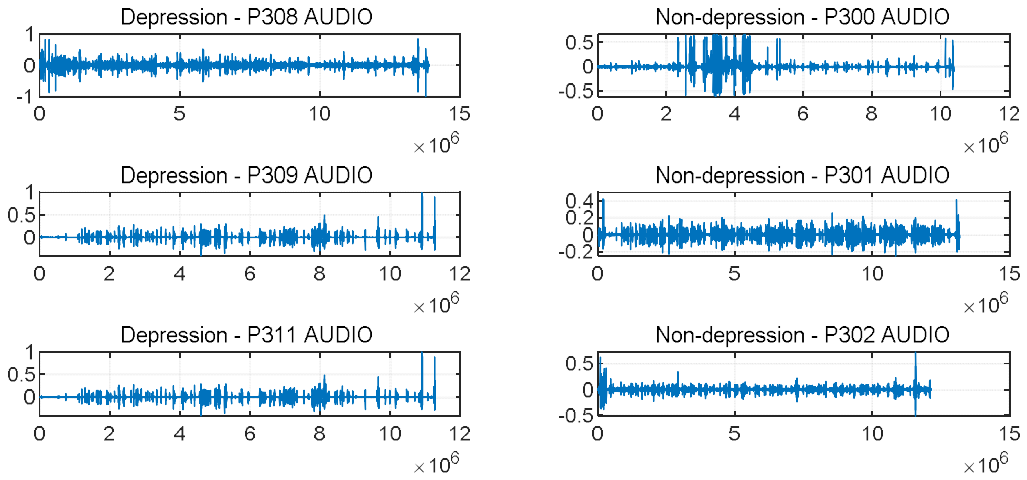


그림 4.3 AI 제어 에이전트를 통한 데이터 수집

표 4.1 PHQ-8 점수에 따른 우울증 증상

PHQ-8 점수	분류
0 ~ 4	우울 증상 없음
5 ~ 9	가벼운 우울 증상
10 ~ 14	중증의 우울 증상
15 ~ 19	중증의 심각한 우울 증상
20 ~ 24	아주 심각한 우울 증상

본 데이터 세트를 구축하기 위해 총 275명의 사람들이 참가했고, 인터뷰 시간은 한 참가자당 7분에서 33분 사이로 평균적으로 16분 정도 인터뷰를 진행하였다. 본 연구에서는 참가자들의 음성 파일과 인터뷰를 기반으로 작성된 텍스트(스크립트) 파일을 사용하였고, 음성 파일은 16kHz로 녹음되었고 wav 형식으로 포함된다. 텍스트(스크립트)는 인터뷰가 진행되는 동안의 대화들이 엑셀 파일 형식으로 저장되어있다. 데이터의 형태는 그림 4.4와 4.5에서 확인할 수 있다.



(a) 우울증 (b) 비우울증

그림 4.4 EDAIC-WOZ 음성 데이터 시각화

Start_Time	End_Time	Text	Confidence
0	9.6	okay perfe	0.927091
11.7	13.6	all right r	0.758649
17.4	18	okay	0.901707
20	21	good to g	0.755682
24.9	31.1	when she	0.891598
33.5	52.4	hi I know	0.94458
56.3	57.3	where are	0.984761
59	60.2	Los Ange	0.973976
66.1	75.9	the South	0.959529
77.3	80.7	what are	0.903911

(a) 우울증

Start_Time	End_Time	Text	Confidence
14.3	15.1	so I'm goi	0.93421
20.3	21.1	interview	0.60847
23.9	24.3	okay	0.690606
62.1	62.7	good	0.951897
68.8	69.8	Atlanta G	0.987629
74.8	77.1	my paren	0.983949
83.4	84.3	I love it	0.969752
88.1	92.9	I like the	0.974442
104.2	105.3	at the mi	0.718575
107.5	108.4	someone	0.722719
113.8	115.1	congestic	0.919954

(b) 비우울증

그림 4.5 EDAIC-WOZ 텍스트 데이터 시각화

본 데이터 세트는 연령, 성별, PHQ-8 점수 등을 고려하여 학습, 검증, 테스트 데이터로 분할된다. 표 4.2에서는 성별에 따른 EDAIC-WOZ 데이터 수를 살펴볼 수 있다. 남성 참가자 중 우울증 증상을 가지고 있는 사람은 35명, 증상이 없는 사람은 135명으로 총 170명이 참가하였다. 또한 여성 참가자 중 우울증 증상을 가지고 있는 사람은 31명, 증상이 없는 사람이 74명으로 총 105명이 참가하였고 여성 참가자보다 남성 참가자들이 더 높은 비율을 차지하고 있다. 이는 우울 증상을 가진 참가자 66명, 우울 증상이 없는 참가자 209명 총 275명의 데이터를 포함한다. 표 4.3에서는 학습, 검증, 테스트 데이터에 포함된 데이터의 수를 확인할 수 있다. 학습데이터는 우울증 37명, 비우울증 126명 총 163명의 데이터가, 검증데이터에는 우울증 12명, 비우울증 44명 총 56명의 데이터가, 테스트 데이터에는 우울증 17명, 비우울증 39명 총 56명의 데이터를 포함한다. 학습 및 검증데이터에는 'Ellie'와의 인터뷰와 AI 제어 에이전트와의 인터뷰가 혼합되어 있지만, 테스트 데이터는 AI 제어 에이전트와의 인터뷰만으로 구성된다. 본 논문에서는 본래 나뉘어 있는 데이터 세트를 사용하지 않고, 전체 데이터의 80%인 220명의 데이터(우울증:53명, 비우울증:167명)는 학습 데이터로 나머지 20%인 55명의 데이터(우울증:13명, 비우울증:42명)는 검증 데이터로 사용하였다.

표 4.2 성별에 따른 EDAIC-WOZ 데이터 수

성별	참가자	우울증	비우울증
남성	170명	35명	135명
여성	105명	31명	74명
합계	275명	66명	209명

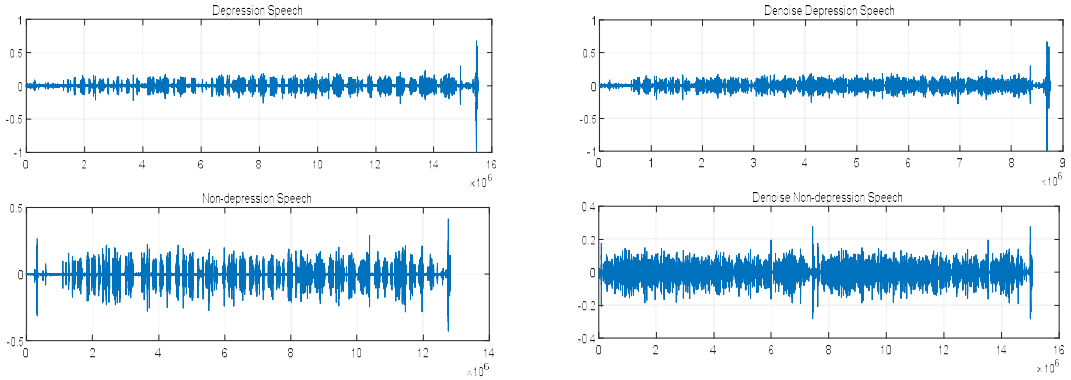
표 4.3 EDAIC-WOZ 데이터 분할

분류 클래스	학습	검증	테스트
우울증	37명 (남성:18명, 여성:19명)	12명 (남성:8명, 여성:4명)	17명 (남성:9명, 여성:8명)
비우울증	126명 (남성:75명, 여성:51명)	44명 (남성:26명, 여성:18명)	39명 (남성:34명, 여성:5명)
합계	163명	56명	56명

제2절 데이터 전처리 방법

1. 음성 데이터 전처리 및 확장

음성 데이터는 주변 환경의 영향으로 녹음을 진행할 때 소음, 묵음 등과 같은 다양한 잡음들이 포함될 수 있다. 데이터의 품질은 모델 성능에 영향을 미칠 수 있어 이러한 잡음 제거하는 과정이 필요하다. EDAIC-WOZ 데이터 구축에 참여한 사람들은 인터뷰 진행 시 최대한 잡음이 발생하지 않은 환경에서 좋은 품질을 가지고 있는 근접 마이크를 사용했지만 그래도 발생하는 잡음들이 조금씩 있다. 이 데이터에는 잡음뿐만 아니라 가상 인터뷰진행자인 'Ellie'의 말도 함께 녹음되어 있어 이러한 잡음들과 'Ellie'의 말소리를 제거해야 한다. 이는 파이썬의 음성 라이브러리 중 하나인 pyAudioAnalysis의 세분화 모듈을 통해 제거되었다. pyAudioAnalysis는 광범위한 음성 분석 절차를 제공하는 파이썬의 공개 소스 라이브러리로 음성의 특징을 추출하고, 분류 모델의 학습, 지도 및 비지도 학습을 사용하여 음성 스트림을 분할하고, 콘텐츠 관계를 시각화하는 데 사용되는 모듈이다. 이를 사용하면 알 수 없는 음성 세그먼트를 미리 정의된 클래스로 분류하고, 음성 녹음 데이터를 분할하고, 동종 세그먼트를 분류하고, 음성 녹음에서 묵음 영역을 제거하고, 음성 세그먼트의 감정을 추정하고, 음악 트랙에서 음성 썸네일을 추출하는 등의 작업을 수행할 수 있다[53]. 본 논문에서는 semi-supervised 침묵 제거 방법과 화자 분할 방법을 이용하여 음성 데이터에 포함된 소음, 묵음, 'Ellie'의 말소리를 제거하였다. 그림 4.6의 (a)는 잡음 제거 전, (b)는 잡음 제거 후의 각각 우울증, 비우울증 음성신호를 시각화한 것이다. 잡음 제거 전 말소리 사이사이 묵음들이 있는데, 잡음 제거 후에는 묵음이 확연히 줄어든 것을 시각적으로 확인할 수 있다. 음성 데이터의 잡음을 제거하는 전처리는 웹 브라우저에서 텍스트와 프로그램 코드를 자유롭게 작성할 수 있는 구글 코랩을 통해 파이썬 언어를 사용하여 진행하였다.



(a) 잡음 제거 전 (b) 잡음 제거 후

그림 4.6 pyaudioAnalysis를 이용한 잡음 제거 음성신호 시각화

일반적으로 질병이 있는 사람의 데이터가 질병이 없는 사람의 데이터에 비해 양적으로 훨씬 적다. EDAIC-WOZ 데이터는 우울증 데이터가 66개, 비우울증 데이터가 209개로 데이터의 양에 차이가 약 3배 정도 있어 클래스 불균형 문제(class imbalance) 문제가 발생한다. 데이터 불균형은 과적합 문제를 일으킬 수 있고, 많은 양의 데이터가 있는 클래스에 모델이 가중치를 많이 두어 가중치가 높은 클래스를 더 예측하려고 한다. 이는 정확도는 높아질 수 있지만, 데이터의 양이 적은 클래스에 대한 정밀도와 재현율이 낮아지는 문제가 발생할 수 있다. 따라서 이러한 문제를 해결하기 위해 우울증 데이터를 확장하는 방법을 통해 각 클래스에 있는 데이터의 양을 맞춰주었다. 데이터 확장은 음성의 피치를 좌우로 이동하는 pitch shifting, 음성의 시간을 좌우로 이동하는 time shifting, 음성신호에 백색 잡음을 추가하는 adding noise 등의 방법을 활용하여 우울증 데이터의 각 파일을 3개씩 확장하였다. 데이터를 확장한 뒤 우울증 데이터는 199개, 비우울증 데이터는 209개로 각 클래스에 포함되는 데이터의 양을 비슷하게 맞춰줌으로써 클래스 불균형 문제를 해결하였다.

2. 텍스트 데이터 전처리 및 확장

자연어 처리에서 분석하고자 하는 데이터를 용도에 맞게 토큰화, 대/소문자 변경, 특수문자 삭제 등과 같은 클렌징 작업이 필요하고, 이는 텍스트 분석에서 매우 중요하다. 따라서 EDAIC-WOZ 데이터 세트에 포함된 텍스트 스크립트 파일을 불러와 전처리를 진행하였다. 먼저 스크립트 파일은 인터뷰 진행 중에 가상 인터뷰 진행자와 참가자가 대화한 내용이 모두 엑셀 파일로 저장되어있어, 가상 인터뷰진행자의 텍스트를 모두 제거하고 참가자의 텍스트만 데이터로 사용하였다. 다음은 텍스트 데이터를 딥러닝 모델에 학습시키고 다양한 성능 비교를 위해 3가지 방법으로 나누어 전처리를 진행하였다. 전처리 방법은 다음과 같다.

[method 1] 첫 번째 방법은 토큰화(tokenization), 텍스트 소문자 변환, 문장 부호 지우기 순으로 사전 처리하였다. 토큰화는 텍스트 데이터에서 토큰(token)이라고 불리는 단위로 나누는 작업이다. 토큰의 단위는 상황에 따라 다르지만, 보통 의미 있는 단위로 토큰을 정의한다. 이러한 방법을 통해 토큰화를 진행한 뒤 영어로 이루어진 데이터에서 대문자를 소문자로 변환하는 작업을 진행한다. 마지막으로 데이터에 포함된 아무런 의미도 갖지 않는 문장 부호와 기호를 지워준다. 그림 4.7은 첫 번째 방법으로 데이터를 전처리한 전후의 워드 클라우드를 보여준다.

그림 4.7 method 1 방법을 이용한 데이터 전처리 전후 워드 클라우드

[method 2] 두 번째 방법은 텍스트 데이터를 토큰화 작업만을 통해 처리하는 것이다. 그림 4.8은 두 번째 방법으로 데이터를 전처리한 전후의 워드 클라우드를 보여준다.

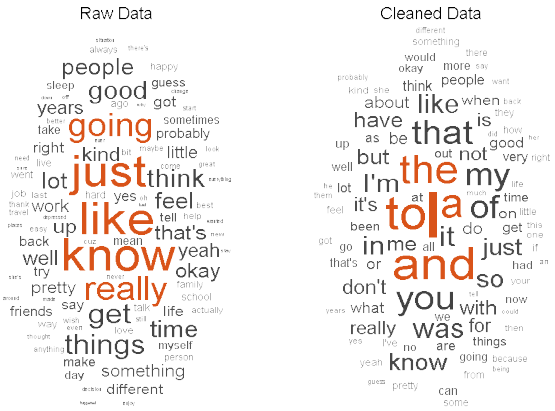


그림 4.8 method 2 방법을 이용한 데이터 전처리 전후 워드 클라우드

[method 3] 세 번째 방법은 토큰화, 표제어 추출(lemmatization), 문장 부호 지우기, 불필요한 단어 제거(removing unnecessary words), 길이가 짧거나 긴 문장 제거 순으로 사전 처리하였다. 텍스트 데이터를 불러와 토큰화 작업을 진행하고,표제어 추출을 진행한다. 그 후 다음 문장 부호를 제거하고 텍스트 데이터에서 잡음이라고 할 수 있는 불용어를 제거한다. 마지막으로 두 단어 미만이거나 열다섯 단어 초과인 문장을 제거한다. 그림 4.9는 세 번째 방법으로 데이터를 전처리한 전후의 워드 클라우드를 보여준다.

그림 4.9 method 3 방법을 이용한 데이터 전처리 전후 워드 클라우드

다음은 음성 데이터와 마찬가지로 텍스트 스크립트도 우울증 데이터 66개, 비우울증 데이터 209개로 클래스 불균형 문제가 있다. 이를 해결하기 위해 EDA(Easy Data Augmentation) 개념을 이용하여 우울증 환자들의 텍스트 데이터를 확장하였다. EDA는 J. Wei[54]가 제안한 텍스트 데이터 확장 기술이고 이는 텍스트 분류 작업의 성능을 향상하는 데 도움을 줄 수 있다. EDA는 동의어 대체(Synonym Replacement, SR), 랜덤 삽입(Random Insertion, RI), 랜덤 스왑(Random Swap, RS) 및 랜덤 삭제(Random Deletion, RD)와 같은 작업으로 구성된다. SR은 문장에서 중단된 단어가 아닌 n개의 단어를 무작위로 선택하고, 이 단어들을 무작위로 선택한 동의어로 교체하는 것이다. RI는 n번 동안 문장에서 중단된 단어가 아닌 임의 단어의 임의 동의어를 찾고, 그 동의어를 문장의 랜덤한 위치에 삽입하는 것이다. RS는 문장에서 n번 동안 무작위로 두 단어를 선택하고 위치를 바꾸는 것이다. RD는 문장의 각 단어에 대해 p의 확률로 무작위 제거하는 것이다. 이러한 방법들의 값을 바꿔가며 우울증 데이터를 각 파일마다 3개씩 확장하였다. 이를 통해 우울증 데이터는 199개, 비우울증 데이터는 209개로 클래스 불균형 문제를 해결해 주었다.

제3절 실험 및 결과분석

본 절에서 음성 및 텍스트 신호로부터 Bi-LSTM과 CNN을 이용한 4-stream 기반 딥러닝 모델의 우울증 진단 성능을 비교 분석한다. 본 연구의 실험 순서는 다음과 같고 실험에 사용한 환경은 표 4.4에서 확인 할 수 있다.

- 음성 데이터를 전처리 전, 잡음 제거 및 데이터 확장 후의 경우로 나누어 Bi-LSTM 및 CNN 기반 전이학습 모델을 이용한 우울증 진단모델의 성능을 비교 분석한다.

- 텍스트 데이터를 데이터 증대 전과 후로 나누어 Bi-LSTM 및 CNN 모델의 우울증 진단 성능을 비교 분석한다.

- 음성 및 텍스트로부터 딥러닝 4-stream 기반 우울증 진단모델의 성능을 확인하고, 기존 2-stream 모델의 성능과 비교 분석한다.

표 4.4 실험 환경

구분		사용
하드웨어	CPU	Intel Core i9 10900K @ 3.70GHz
	GPU	NVIDIA GeForce RTX 2080 SUPER
	RAM	128GB
소프트웨어	OS	Windows10
	프로그램 언어	Python, Matlab

1. 음성 데이터를 이용한 우울증 진단 실험 및 결과

가. Bi-LSTM 기반 우울증 진단

첫 번째 실험은 EDAIC-WOZ 데이터 중 음성 데이터를 이용한 Bi-LSTM 기반 우울증 진단의 실험 결과를 보여준다. 실험은 데이터 잡음 제거 및 확장 전후로 나누어 1차원 음성신호의 특징추출 방법을 MFCC, GTCC, MFCC+GTCC 세 가지 경우로 나누고 Bi-LSTM 계층의 은닉유닛 개수를 10, 50, 100으로 바꿔가며 진행하였다. 표 4.5에서는 Bi-LSTM 모델의 학습 파라미터값을 보여준다. 표 4.5에서 보는 바와 같은 값으로 파라미터를 고정하여 학습을 진행하였다.

표 4.5 Bi-LSTM 모델의 학습 파라미터(음성)

학습 옵션 / 파라미터값	최적화 함수	기울기 이동평균 감쇠율	미니배치 사이즈	최대 학습 횟수
파라미터값	Adam	2	512	10

그림 4.10에서는 데이터 잡음 제거 및 확장 전 Bi-LSTM 모델의 분류 성능에 대한 그래프를 보여준다. 그림 4.10에서 보는 바와 같이 음성 데이터에 아무런 전처리를 하지 않았을 때는 대체적으로 비슷한 성능을 보인다. 여기서 MFCC와 GTCC의 특징을 함께 추출하고 은닉유닛의 개수가 100개일 때 Bi-LSTM의 정확도가 80%로 가장 높은 성능을 보인다.

그림 4.11에서는 데이터 잡음 제거 및 데이터 확장 후 Bi-LSTM 모델의 분류 성능에 대한 그래프를 보여준다. 4.11에서 보는 바와 같이 데이터 전처리 전의 실험에서보다 전체적으로 성능이 개선되었다. 여기서 Bi-LSTM 계층의 은닉유닛의 수가 10이고 MFCC와 GTCC 특징을 같이 추출했을 때 96.34%로 가장 높은 성능을 보여준다. 이는 음성 데이터를 이용한 Bi-LSTM 기반 우울증 진단에서 가장 높은 성능을 보였다. 같은 조건에서 데이터 전처리 전의 성능과 비교했을 때, 성능이 약 18.16% 정도 개선되었다.

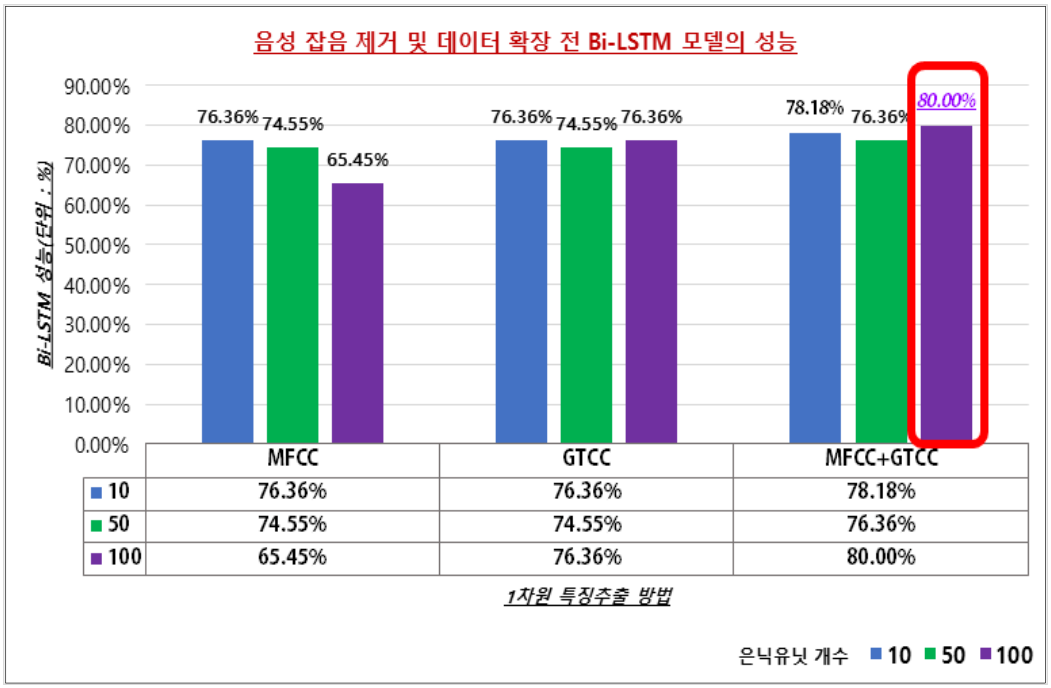


그림 4.10 음성 잡음 제거 및 데이터 확장 전 Bi-LSTM 성능 그래프

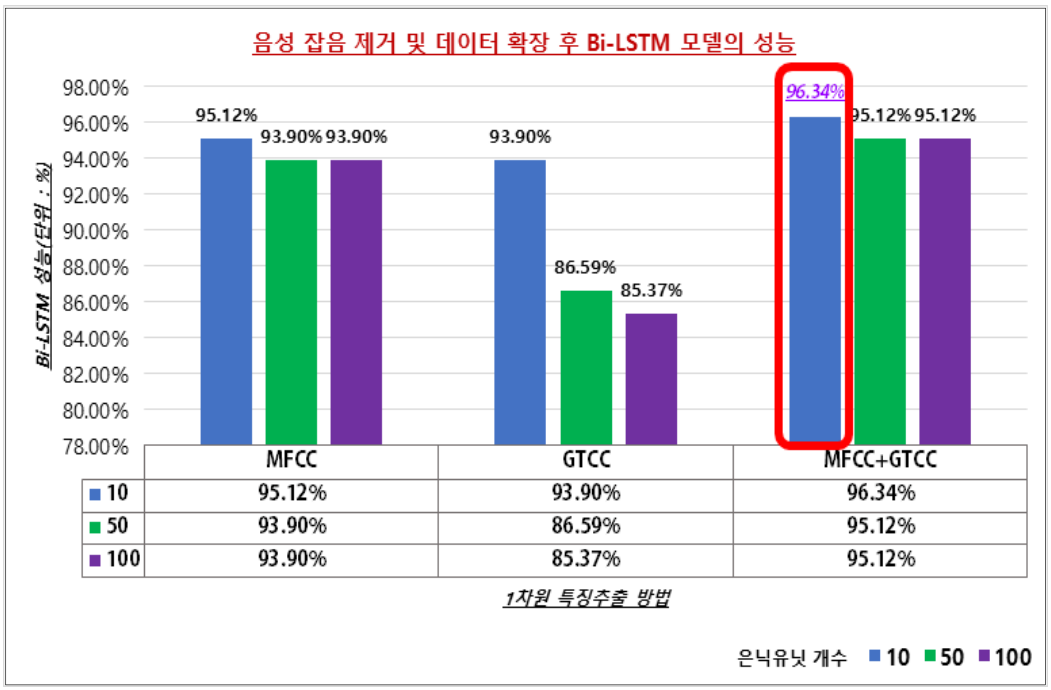


그림 4.11 음성 잡음 제거 및 데이터 확장 후 Bi-LSTM 성능 그래프

나. CNN 전이학습 모델 기반 우울증 진단

다음은 음성 데이터를 이용한 CNN 기반 전이학습 모델을 통한 우울증 진단의 실험 결과를 보여준다. 딥러닝 모델은 음성 기반 전이학습 모델 중 VGGish, YAMNet, OpenL3를 이용하였다. 음성신호의 2차원 시간-주파수 표현변환에 의한 Bark Spectrogram, ERB Spectrogram, Log-Mel Spectrogram 특징을 이용하였다. 이때, 시간-주파수 표현의 이미지 특징은 흑백과 RGB 두 가지 경우로 나누어 실험을 진행하였다. 실험에서 사용된 전이학습 모델은 흑백 이미지를 입력으로 받는다. 따라서 RGB 이미지를 입력으로 넣을 때는 모델의 입력단과 첫 번째 합성곱 계층을 입력의 크기에 맞게 변경해주었다. 각 전이학습 모델의 입력 크기는 표 4.6에서 확인할 수 있다. 표 4.7에서는 전이학습 모델의 학습 파라미터값을 보여준다. 표 4.7에서 보는 바와 같은 값으로 파라미터를 고정하여 학습을 진행하였다.

표 4.6 CNN 기반 전이학습 모델별 이미지 입력 크기

전이학습 모델 / 이미지 입력	VGGish	YAMNet	OpenL3
흑백 이미지	96*64*1	96*64*1	128*199*1
RGB 이미지	96*64*3	96*64*3	128*199*3

표 4.7 CNN 기반 전이학습 모델의 학습 파라미터(음성)

학습 옵션 / 파라미터값	최적화 함수	기울기 이동평균 감쇠율	미니배치 사이즈	최대 학습 횟수
파라미터값	Adam	2	256	5

그림 4.12와 그림 4.13은 각각 데이터 잡음 제거 및 확장 전 흑백 이미지와 RGB 이미지를 입력으로 넣었을 때 특징추출 방법에 따른 전이학습 모델의 분류 성능을 보여준다. 그림 4.12에서 보는 바와 같이 전처리 전, 음성신호의 2차원 시간-주파수 변환 기반 특징의 흑백 이미지를 전이학습 모델의 입력으로 넣었을 때는 대체적으로 성능이 비슷하다. 그중 모든 특징에서 가장 높은 성능을 보이는 모델은 76.36%의 정확도를 가지는 VGGish 모델이다.

마찬가지로 그림 4.13에서 확인할 수 있듯이 RGB 이미지 특징을 입력으로 넣었을 경우에도 비슷한 성능을 보인다. 결과적으로 전처리 전 CNN 기반 전이학습 모델은 데이터양이 많은 비우울증 클래스에 대해 더 집중적으로 학습이 되어 우울증 데이터를 제대로 분류하지 못하는 모습을 보였다.

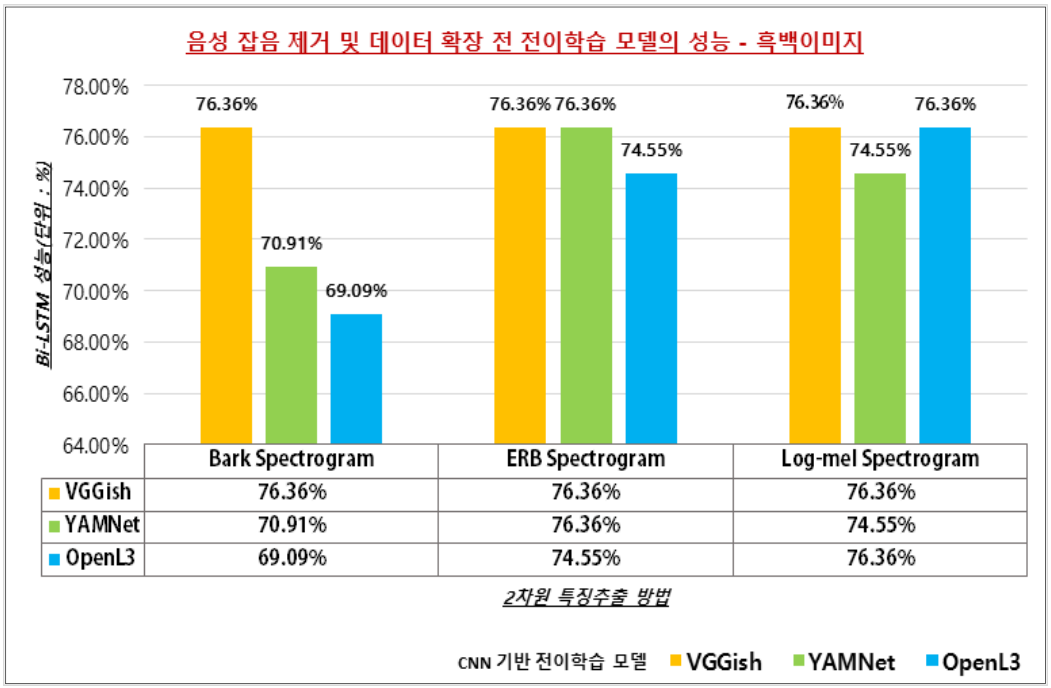


그림 4.12 음성 잡음 제거 및 데이터 확장 전 전이학습 모델 성능 그래프 - 흑백 이미지

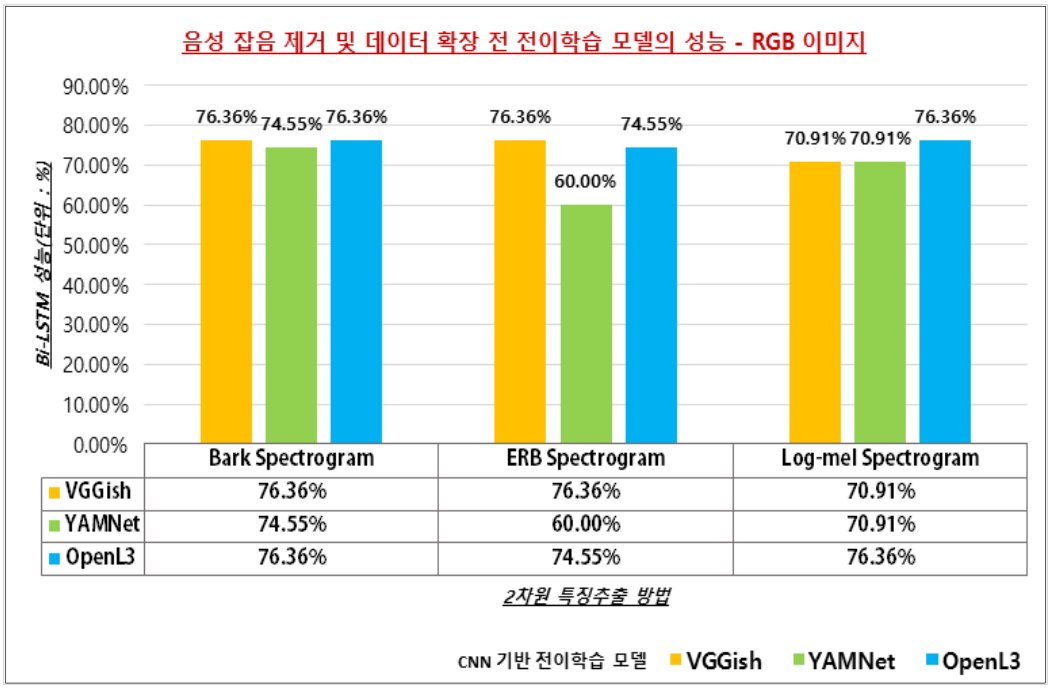


그림 4.13 음성 잡음 제거 및 데이터 확장 전 전이학습 모델 성능 그래프 - RGB 이미지

그림 4.14와 그림 4.15에서는 각각 흑백 및 RGB 이미지 특징을 입력으로 넣었을 때 데이터 잡음 제거 및 확장 후 특징추출 방법에 따른 전이학습 모델의 분류 성능 그래프를 보여준다. 그림 4.14에서 보면 Bark 및 Log-Mel spectrogram 특징을 이용해 학습을 진행했을 때는 대체적으로 좋은 성능을 보였다. 반면 ERB spectrogram은 흑백 이미지에서 특징이 잘 드러나지 않아 데이터를 제대로 분류하지 못하는 모습을 보였다. 여기서 Log-Mel spectrogram 특징을 이용하여 OpenL3 모델에 학습시켰을 때 95.12%로 가장 높은 분류 성능을 보였다.

그림 4.15에서는 흑백 이미지를 입력으로 넣었을 때보다 전체적으로 성능이 향상했다는 것을 확인할 수 있다. ERB spectrogram은 흑백 이미지를 넣었을 경우보다 RGB 이미지를 전이학습 모델의 입력으로 넣었을 경우, 음성의 2차원적인 특징이 잘 드러나 클래스별로 잘 분류하는 모습을 보였다. 여기서 각 전이학습 모델은 Log-Mel spectrogram을 입력으로 넣었을 때 가장 좋은 성능을 보였다. 특히 OpenL3 모델의 분류 성능이 96.34%로 가장 높은 성능을 보였다.

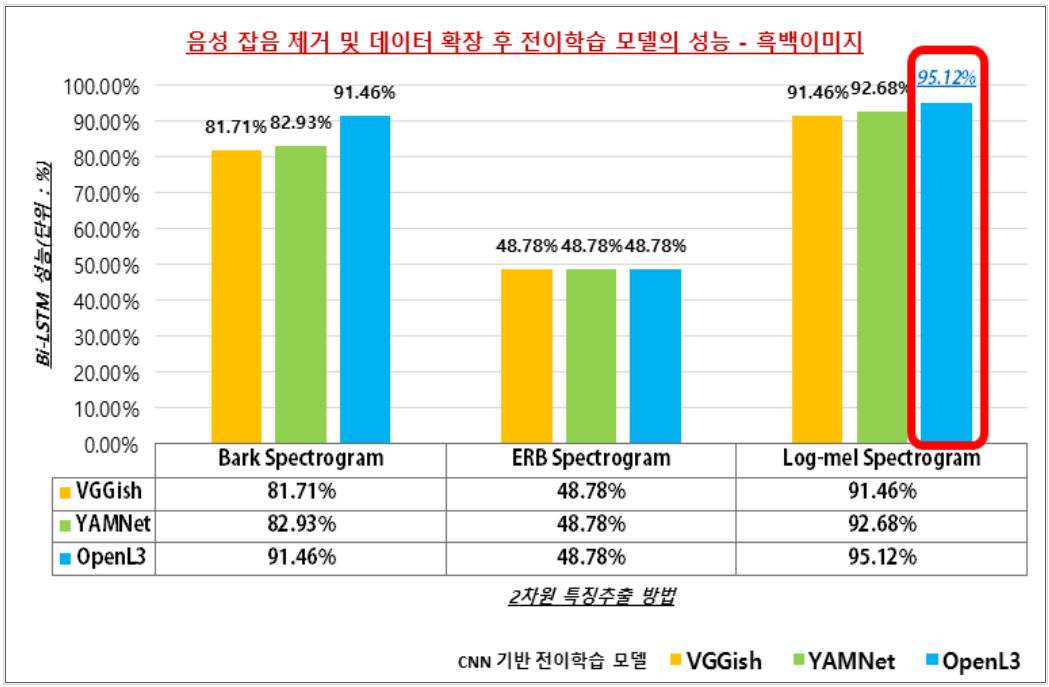


그림 4.14 음성 잡음 제거 및 데이터 확장 후 전이학습 모델 성능 그래프 - 흑백 이미지

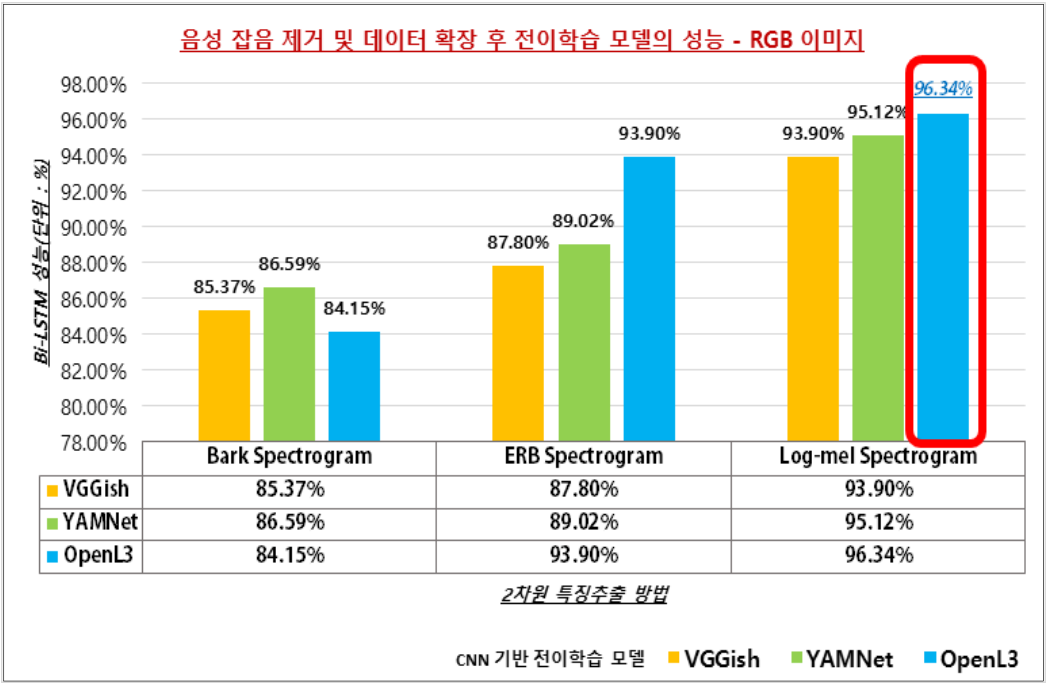


그림 4.15 음성 잡음 제거 및 데이터 확장 후 전이학습 모델 성능 그래프 - RGB 이미지

결과적으로 첫 번째 실험에서는 음성 데이터의 잡음 제거 및 우울증 데이터 확장 후 데이터 질이 향상되고 클래스 불균형 문제가 해결됨으로써 Bi-LSTM 및 CNN 기반 전이학습 모델의 분류 성능이 개선됨을 볼 수 있다. 이때 BI-LSTM 모델의 분류 성능은 Bi-LSTM 계층의 은닉유닛의 수가 100이고 MFCC와 GTCC 특징을 같이 추출했을 때 96.34%로 가장 높은 성능을 보인다. CNN 기반 전이학습 모델의 성능은 log-mel spectrogram 특징을 RGB 이미지 입력으로 넣었을 때 OpenL3 전이학습 모델의 분류 정확도가 96.34%로 우울증을 진단하는데 가장 높은 성능을 보였다.

2. 텍스트 데이터를 이용한 우울증 진단 실험 및 결과

두 번째 실험은 EDAIC-WOZ 데이터 중 텍스트 스크립트 데이터를 이용한 Bi-LSTM 및 CNN 기반 우울증 진단 실험의 결과를 보여준다. 실험은 우울증 데이터 확장 전후의 경우로 나누어 표 4.8에서 보는 바와 같이 텍스트 전처리 방법을 3가지로 나누고 임베딩 계층의 차원을 100, 200, 500으로 바꿔가며 진행하였다. Bi-LSTM 및 CNN 모델의 학습 옵션은 표 4.9와 같이 파라미터값을 지정하여 모델 학습에 사용하였다.

표 4.8 텍스트 전처리 방법

방법	전처리
method 1	토큰화 -> 텍스트 소문자 변환 -> 문장 부호 지우기
method 2	토큰화
method 3	토큰화 -> 표제어 추출 -> 문장 부호 지우기 -> 불필요한 단어 제거 -> 길이가 짧거나 긴 문장 제거

표 4.9 Bi-LSTM 및 CNN 모델의 학습 파라미터(텍스트)

학습 옵션 / 파라미터값	최적화 함수	기울기 이동평균 감쇠율	미니배치 사이즈	최대 학습 횟수
파라미터값	Adam	2	512	10

가. Bi-LSTM 기반 우울증 진단

먼저 텍스트 데이터를 이용한 BI-LSTM 기반 전이학습 모델을 통한 우울증 진단의 실험 결과를 보여준다. 표 4.10에서는 텍스트 데이터 확장 전후의 Bi-LSTM의 분류 성능을 보여준다. 표에서 보면 데이터 확장 전의 경우, 임베딩 차원의 값과 전처리 방법에 상관없이 전체적으로 비슷한 성능을 가진다. 여기서 가장 높은 성능은 78.18%로 임베딩 차원이 100일 때 텍스트 데이터를 토큰화 전처리만 했을 경우이다.

표를 살펴보면 데이터 확장 전에는 우울증 데이터를 제대로 분류하지 못하는 모습을 보였다. 하지만 EAD를 이용하여 우울증 텍스트 데이터를 확장한 후 클래스 불균형이 해결됨에 따라 성능이 높지는 않지만, 우울증과 비우울증을 골고루 분류하는 모습을 보였다. 본 실험에서는 데이터 증대 후 전처리 방법이 method 2(텍스트 데이터 토큰화)일 때 임베딩 계층의 차원을 200으로 지정하여 학습했을 경우 모델의 정확도가 76.83%로 가장 좋은 성능을 보인다.

표 4.10 텍스트 데이터를 이용한 Bi-LSTM 모델 성능

	임베딩 차원 / 전처리 방법	100	200	500
데이터 확장 전	method 1	76.36%	76.36%	70.91%
	method 2	78.18%	76.36%	76.36%
	method 3	70.91%	69.09%	76.36%
데이터 확장 후	method 1	70.73%	64.63%	74.39%
	method 2	70.73%	**76.83%**	64.63%
	method 3	73.17%	70.73%	69.51%

나. CNN 모델 기반 우울증 진단

다음은 텍스트 데이터를 이용한 CNN 기반 전이학습 모델을 통한 우울증 진단의 실험 결과를 보여준다. 표 4.11에서는 텍스트 데이터 확장 전후의 CNN 모델의 분류 성능을 보여준다. 표를 보면 데이터 확장 전의 경우에는 전체적으로 성능이 비슷함을 확인할 수 있다. 그중 첫 번째 전처리 방법(method 1)을 이용하고 임베딩 차원이 100일 때 78.18%로 가장 좋은 성능을 보인다. 데이터 확장 후의 경우에는 전처리 방법이 method 2(텍스트 데이터 토큰화)일 때 CNN 모델이 전체적으로 좋은 성능을 보인다. 특히 임베딩 계층의 차원을 100으로 지정하여 학습을 진행했을 때 모델의 정확도가 81.71%로 가장 좋은 성능을 보였다. 이 결과는 Bi-LSTM 모델보다 4.88% 정도 성능이 개선되었음을 보인다.

표 4.11 텍스트 데이터를 이용한 CNN 모델 성능

	임베딩 차원 / 전처리 방법	100	200	500
데이터 확장 전	method 1	**78.18%**	63.64%	72.73%
	method 2	74.55%	70.91%	76.36%
	method 3	74.55%	69.09%	76.36%
데이터 확장 후	method 1	75.61%	71.95%	79.27%
	method 2	**81.71%**	76.83%	79.27%
	method 3	74.39%	70.73%	68.29%

결과적으로 두 번째 실험에서는 데이터 확장 후 딥러닝 모델의 분류 성능이 눈에 띄게 개선되지는 않았다. 하지만 클래스 불균형 문제를 해결함으로써 데이터양이 많은 클래스에 학습이 치중되는 문제를 막을 수 있었다. Bi-LSTM 모델은 임베딩 차원이 200이고 데이터를 토큰화만 했을 때 76.83%로 가장 높은 성능을 보인다. 또한 CNN 모델은 임베딩 차원이 100이고 데이터를 토큰화만 했을 때 81.71%로 텍스트 데이터로부터 우울증을 진단하기 위한 모델로 가장 적합함을 보였다.

3. 음성 및 텍스트 데이터를 이용한 4-stream 모델 기반 우울증 진단 실험 및 결과

가. Bi-LSTM과 CNN의 4-stream 모델 기반 우울증 진단

마지막 실험에서는 본 논문에서 제안한 멀티모달 데이터를 이용하여 late score fusion 방법에 따른 4-stream 딥러닝 모델의 우울증 진단 성능을 보여준다. 표 4.12에서는 late score fusion을 위해 첫 번째, 두 번째 실험에서 가장 높은 성능을 가지는 음성 및 텍스트 데이터의 특징추출 및 전처리 방법의 세 가지 경우를 보여준다. 이는 4-stream 딥러닝 모델의 우울증 진단 성능을 확인하기 위해 각 case마다 고정된 값으로 사용된다. 최종적으로 음성 데이터 잡음 제거 및 데이터 확장 후 CNN 기반 전이학습 모델의 성능을 기준으로 표 4.12에 정리된 고정된 값들과 late score fusion 하여 4-stream 모델의 분류 성능을 확인한다.

표 4.12 late score fusion을 위한 데이터의 특징추출 및 전처리 방법과 정확도

	데이터	특징추출 및 전처리 방법	정확도(%)
case 1	음성	1차원 : 은닉유닛 10, MFCC + GTCC	96.34
		2차원 : Bark Spectrogram(RGB)	85.37/86.59/84.15
	텍스트	Bi-LSTM : 임베딩 계층 200, 토큰화	76.83
		CNN : 임베딩 계층 100, 토큰화	81.71
case 2	음성	1차원 : 은닉유닛 10, MFCC + GTCC	96.34
		2차원 : ERB Spectrogram(RGB)	87.80/89.02/93.90
	텍스트	Bi-LSTM : 임베딩 계층 200, 토큰화	76.83
		CNN : 임베딩 계층 100, 토큰화	81.71
case 3	음성	1차원 : 은닉유닛 10, MFCC + GTCC	96.34
		2차원 : Log-mel Spectrogram(RGB)	96.90/95.12/96.34
	텍스트	Bi-LSTM : 임베딩 계층 200, 토큰화	76.83
		CNN : 임베딩 계층 100, 토큰화	81.71

그림 4.16에서는 late score sum 방법을 이용하여 4개의 딥러닝 모델 소프트맥스 값을 모두 더했을 때 각 case 별 4-stream 모델의 성능을 그래프 형태로 보여준다. 그림 4.16에서 보는 바와 같이 case 2에서 VGGish와 OpenL3 모델을 이용했을 때의 정확도가 97.56%로 표 4.12에서 case 2의 가장 높은 단일 모델의 성능보다 1.22%가 향상되었음을 보여준다.

그림 4.17은 late score sum 방법을 이용한 4-stream 기반 딥러닝 모델 중 가장 높은 성능을 가지는 OpneL3_4-stream 모델에 대한 오차 행렬을 보여준다. 그림 4.17의 (a)와 (b)는 음성 데이터를 이용한 우울증 진단모델의 오차 행렬을 출력한 그림이다. 이를 살펴보면 높은 확률로 우울증과 비우울증을 고르게 잘 분류하고 있는 것을 확인할 수 있다. (c)와 (d)는 텍스트 데이터를 이용한 우울증 진단모델의 오차 행렬을 출력한 그림이다. 이는 높은 확률은 아니지만 한 클래스에만 치중되어 분류하는 것이 아니라 두 클래스 모두 고르게 분류하고 있는 것을 확인할 수 있다. (e)는 late score sum 기반 4-steram 모델에 대한 오차 행렬을 확인할 수 있다. 이를 살펴보면 각 클래스당 1개의 데이터를 분류하지 못했지만 비우울증과 우울증 모두 잘 분류한 모습을 보인다.

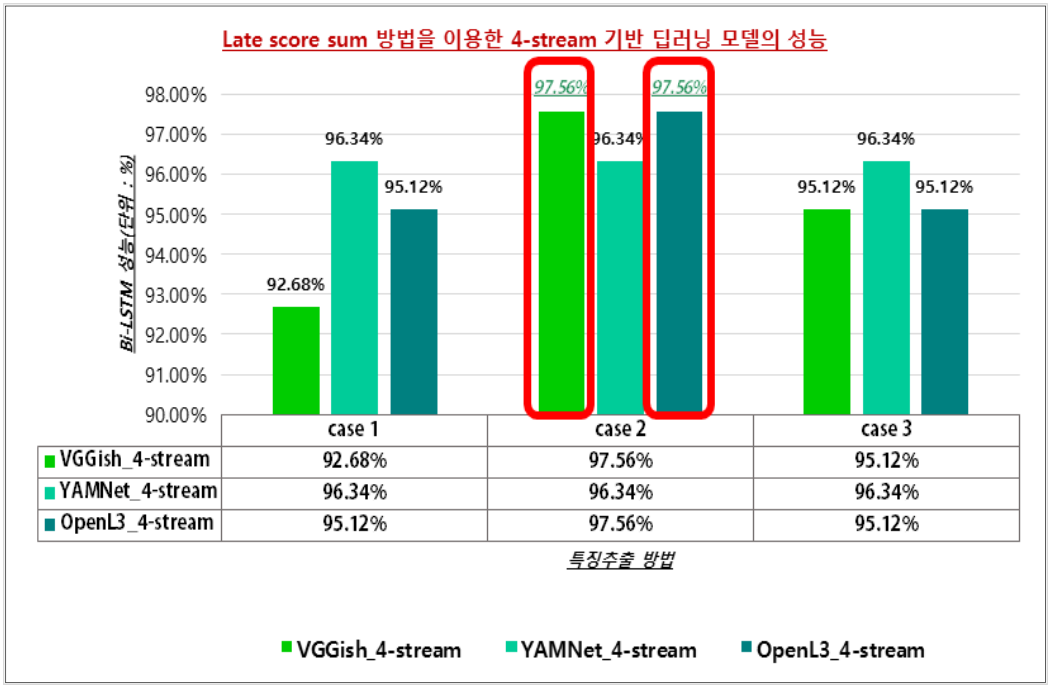
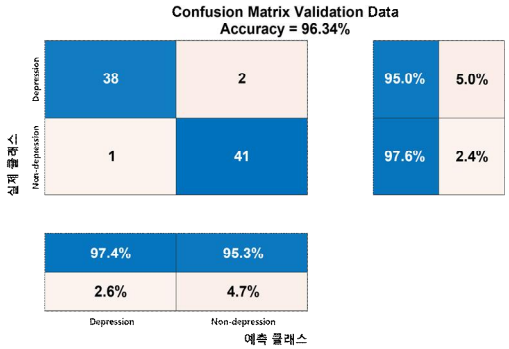
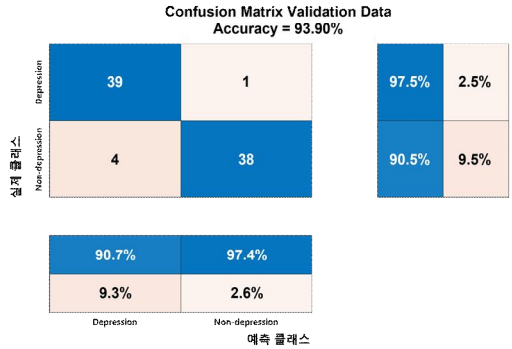


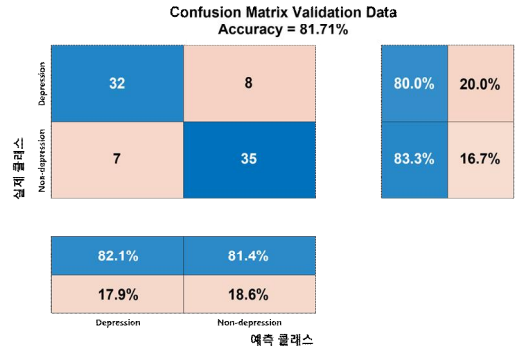
그림 4.16 late score sum 방법 4-stream 기반 딥러닝 모델의 성능 그래프



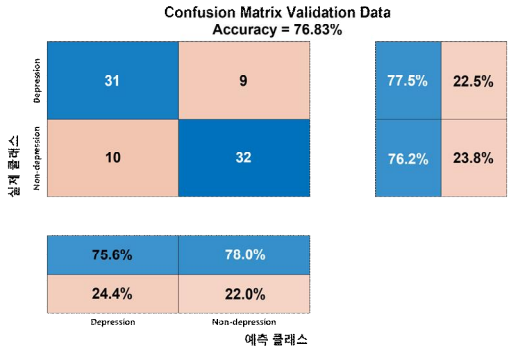
(a) 음성 Bi-LSTM 모델 - MFCC+GTCC



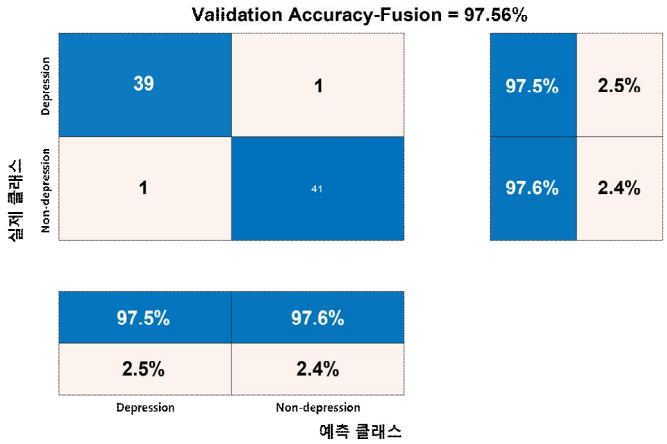
(b) 음성 OpenL3 모델 - ERB



(c) 텍스트 Bi-LSTM 모델



(d) 텍스트 CNN 모델



(e) late score sum 기반 4-steram 모델

그림 4.17 late score sum 방법 4-stream 기반 딥러닝 모델의 오차 행렬(case2, OpenL3_4-stream 모델)

그림 4.18에서는 late score product 방법을 이용하여 4개의 딥러닝 모델 소프트맥스 값을 모두 곱했을 때 각 case 별 4-stream 모델의 성능을 그래프 형태로 보여준다. 그림 4.18에서 보는 바와 같이 case 1에서는 YAMNet 모델을 이용했을 때의 정확도가 97.56%로 표 6-22에서 case 1의 가장 높은 단일 모델의 성능보다 1.22%가 향상되었음을 보여준다. 또한 case 2에서는 OpenL3 모델을 이용했을 때의 정확도가 98.78%로 표 4.12에서 case 2의 가장 높은 단일 모델의 성능보다 2.44%가 개선되었다. 마지막으로 case 3에서는 모든 모델의 성능이 표 4.12에서 case 3의 가장 높은 단일 모델의 성능보다 개선되었음을 확인할 수 있다.

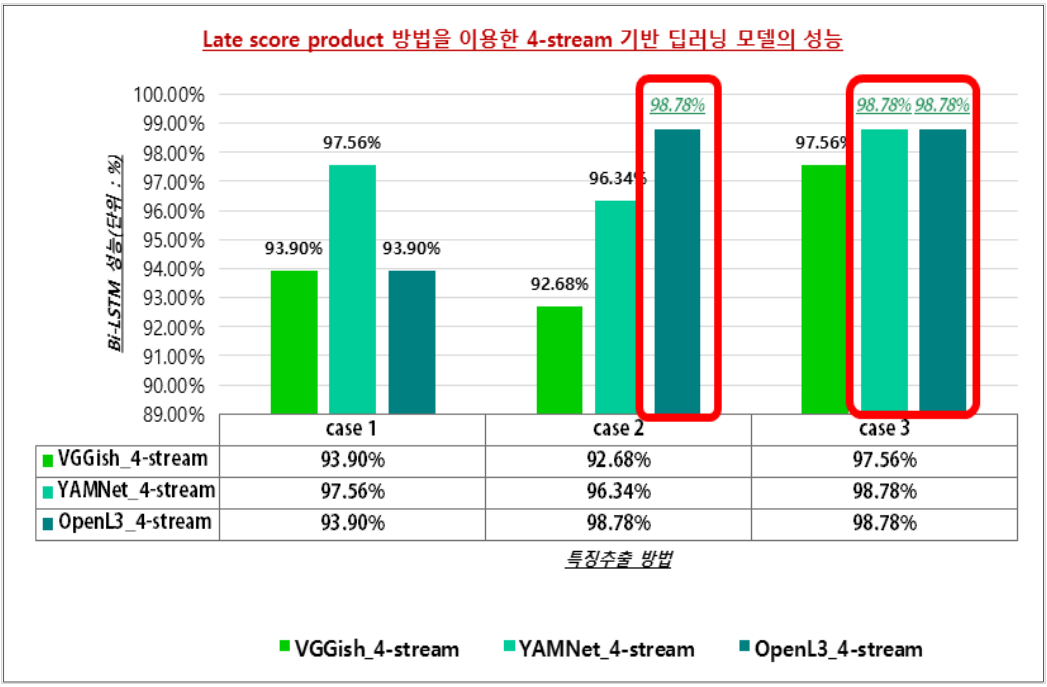
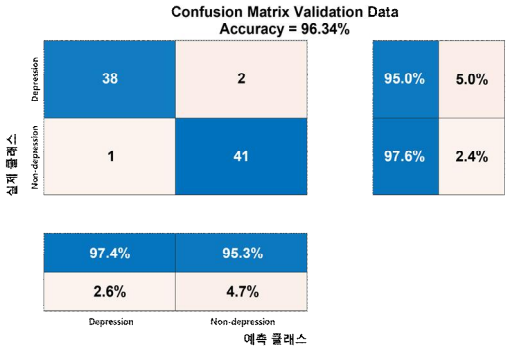


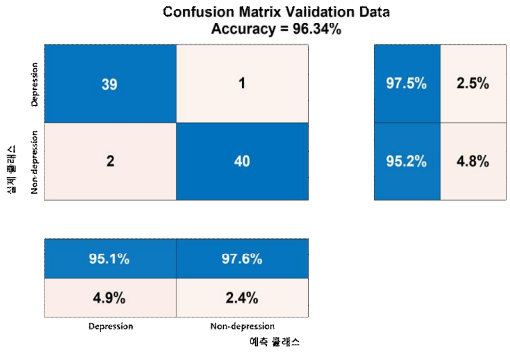
그림 4.18 late score product 방법 4-stream 기반 딥러닝 모델의 성능 그래프

그림 4.19는 late score product 방법을 이용한 4-stream 기반 딥러닝 모델 중 가장 높은 성능을 가지는 OpneL3_4-stream 모델에 대한 오차 행렬을 보여준다. 그림 4.19 (a), (c), (d)의 오차 행렬은 앞서 그림 4.17의 (a), (c), (d)에서 보여준 오차 행렬과 같은 값을 보여준다. 그림 4.19의 (b)는 Log-Mel 스펙트로그램 특징을 RGB 이미지 형태로 OpenL3 모델에 적용했을 때의 오차 행렬을 보여준다. 이를 살펴보면 높은 확률로 우울증과 비우울증을 고르게 잘 분류하고 있는 것을 확인할 수 있다. 그림 4.19의 (e)는 late score product 기반 4-steram 모델에 대한 오차 행렬을 확인할 수 있다. 이를 살펴보면 우울증 데이터 1개를 제외한 나머지 데이터를 모두 잘 분류한 모습을 보인다.

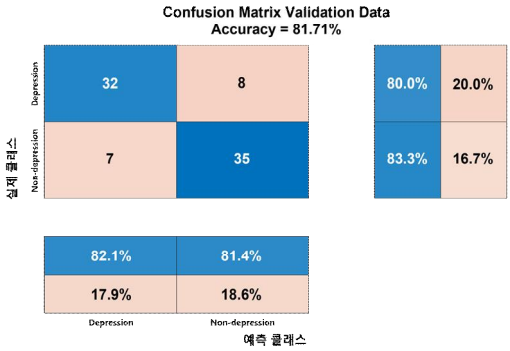
결과적으로 소프트맥스 확률값을 모두 곱했을 때 case 2에서 OpenL3 및 case 3에서 YAMNet과 OpenL3이 포함된 4-stream 모델이 98.78%로 가장 높은 분류 정확도를 가진다. 따라서 단일 데이터만 사용했을 때보다 멀티모달 데이터를 사용한 4-stream 모델의 정확도가 2.44% 향상하여, 우울증 진단을 위한 딥러닝 모델의 성능이 개선되었음을 증명하였다.



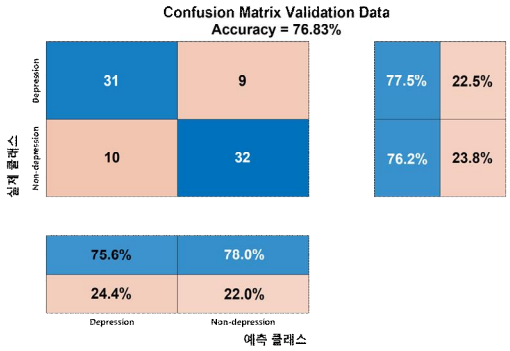
(a) 음성 Bi-LSTM 모델 - MFCC+GTCC



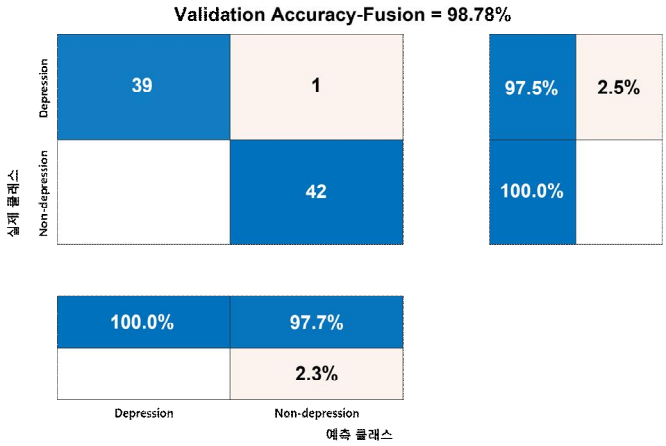
(b) 음성 OpenL3 모델 - Log-Mel



(c) 텍스트 Bi-LSTM 모델



(d) 텍스트 CNN 모델



(e) late score product 기반 4-steram 모델

그림 4.19 late score sum 방법 4-stream 기반 딥러닝 모델의 오차 행렬(case3, OpenL3_4-stream 모델)

나. 기존 연구와의 성능 비교 및 분석

본 논문에서는 제안된 4-stream 기반 모델이 우울증 진단에 효과적임을 증명하기 위해 기존 연구와 제안된 모델의 성능을 비교 분석했다. 기존 연구에서 사용된 데이터베이스는 DAIC-WOZ 우울증 데이터베이스지만 실험에 사용된 데이터베이스는 EDAIC-WOZ 우울증 데이터베이스다. 따라서 동일한 조건에서 성능을 비교하기 위해 DAIC-WOZ 데이터를 이용해 가장 좋은 성능을 가지는 모델을 기반으로 하여 실험을 진행했다. 표 4.13은 DAIC-WOZ 데이터를 이용한 제안된 모델의 성능을 보여준다. 표에서 보는 바와 같이 DAIC-WOZ 데이터도 단일 데이터를 사용하는 것보다 멀티모달 데이터를 이용하여 late product 했을 때 96.67%로 가장 좋은 성능을 보인다.

표 4.13 DAIC-WOZ 데이터를 이용한 제안된 모델의 성능

사용 데이터	특징추출방법 및 전처리 방법	딥러닝 모델	정확도
음성	MFCC+GTCC	Bi-LSTM	90.00%
	log-mel spectrogram	OpenL3	93.33%
텍스트	토큰화	Bi-LSTM	71.67%
	토큰화	CNN	65.00%
멀티모달	더하기	4-stream 기반 모델	91.67%
	곱하기		**96.67%**

기존 연구와의 성능 비교를 위해 정밀도(Precision), 재현율(Recall), F1-Score이라는 성능지표를 사용하였다. 정밀도는 모델이 실제로 예측한 데이터 중 실제로 True인 데이터의 비율이고 재현율은 실제로 True인 데이터를 모델이 실제라고 인식한 데이터의 비율, F1-Score는 정밀도와 재현율의 조화평균을 뜻한다. 정밀도, 재현율, F1-Score는 모두 0과 1 사이의 값을 가지고, 이 값이 1에 가까울수록 성능이 개선된다.

이를 사용해 성능을 비교해본 결과, 표 4.14에서 확인할 수 있듯 DAIC-WOZ 데이터베이스를 이용했을 때 제안된 모델의 정밀도는 0.97, 재현율은 0.97, F1-score는 0.97의 값을 가진다. 이는 같은 조건 속에서도 기존 연구들에 비해 제안된 모델의 성능이 더 높은 것을 확인할 수 있다. 또한 EDAIC-WOZ 데이터베이스를 이용했을 때는 정밀도는 1.00, 재현율은 0.98, F1-socre은 0.99의 값을 가진다. 이는 그림 4.20의 그래프를 통해 더 쉽게 확인할 수 있다. 이 결과는 기존 연구들과 비교했을 때 가장 좋은 성능을 가지는 것을 확인할 수 있고 최종적으로 제안된 음성 및 텍스트 데이터를 이용한 Bi-LSTM과 CNN 모델의 4-stream 기반의 딥러닝 모델이 우울증 진단에 효과적임을 증명하였다.

표 4.14 기존 연구와 제안된 4-stream 모델의 성능 비교

성능 비교 기준 / 기존 연구 및 제안된 모델	사용 데이터	F1-score	Precision	Recall
T. Alhanai[33]	DAIC-WOZ	0.77	0.71	0.83
L. Lin[34]	DAIC-WOZ	0.85	0.79	0.92
G. Lam[35]	DAIC-WOZ	0.87	0.91	0.83
제안된 4-stream 모델 (비교실험)	**DAIC-WOZ**	**0.97**	**0.97**	**0.97**
제안된 4-stream 모델	**EDAIC-WOZ**	**0.99**	**1.00**	**0.98**

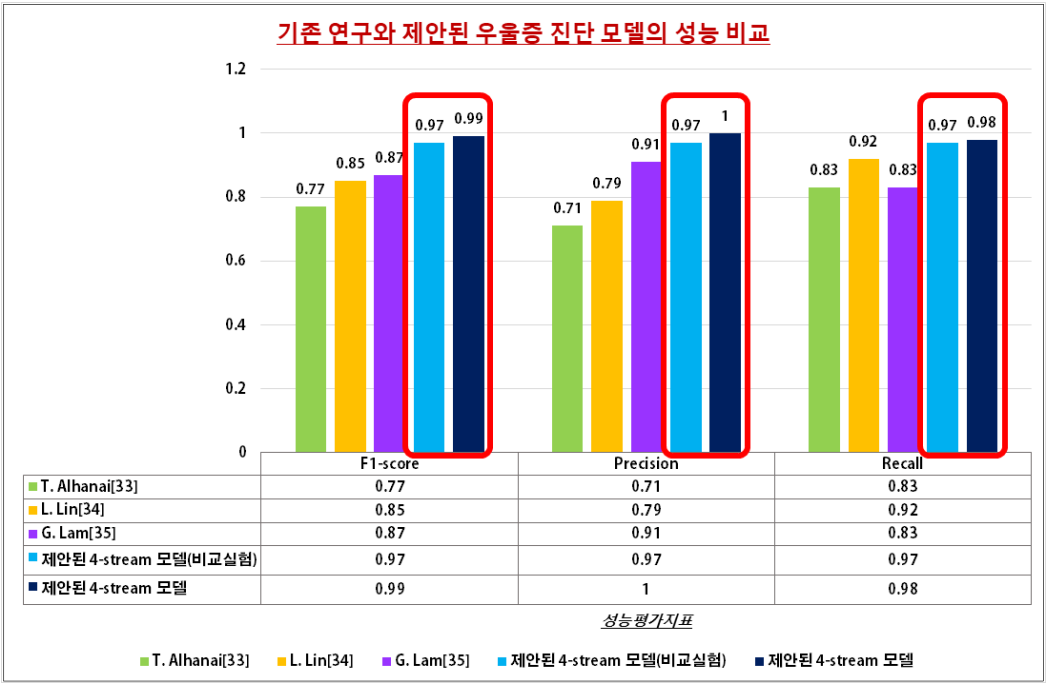


그림 4.20 기존 연구와 제안된 4-stream 모델의 성능 비교 그래프

제5장 결론

본 논문에서는 음성 및 텍스트 멀티모달 데이터로부터 Bi-LSTM과 CNN 모델의 late score fusion 방법을 이용한 4-stream 기반 우울증 진단모델을 설계하고 이에 대한 성능을 비교 분석했다. 우울증은 전 세계 모든 사람에게 흔히 발생할 수 있는 정신질환 중 하나이고 이를 방치하면 심각한 문제를 초래할 수 있다. 멀티모달 데이터를 이용한 우울증 진단은 복합적인 정신질환인 우울증에 대한 부족한 정보를 얻을 수 있어 진단에 더 효과적이다. 그중 음성과 텍스트와 같은 언어적 특성에는 우울 증상을 가지고 있는 사람들의 특징이 확연히 드러난다. 우울 증상을 가지고 있는 사람들은 말하는 강도가 보통 사람들보다 낮고, 음역대가 감소하고 말하는 속도가 느리다는 확연한 특징이 있어 우울증 진단 시 많이 사용되고 있다.

음성신호는 신호를 취득하는 환경이나 장비에 따라서 잡음이 포함되는 경우가 있다. 데이터의 품질은 모델의 성능에 영향을 미칠 수 있어 pyAudioAnalysis를 이용하여 음성신호에 포함된 잡음을 제거하였다. 본 논문에서 사용된 데이터베이스는 EDAIC-WOZ 우울증 데이터베이스로 우울 증상을 가지고 있는 사람 66명 정상인 사람 209명으로 총 275명의 참가자의 음성 및 텍스트 데이터가 포함되어있다. 본 데이터베이스는 비우울증 데이터가 우울증 데이터에 비해 3배 정도 많아 클래스 불균형 문제가 발생한다. 이를 해결하기 위해 우울증 데이터를 각각 3개씩 확장하여 모델의 학습에 사용하였다. 이렇게 전처리 된 음성신호는 1차원과 2차원으로 나누어 특징을 추출하고 각각 Bi-LSTM과 CNN 기반 전이학습 모델에 적용하였다.

텍스트 데이터는 토큰화, 대/소문자 변경, 특수문자 삭제 등과 같은 클렌징 작업을 통한 전처리 과정이 필요하다. 또한 이를 딥러닝 모델에 입력하기 위해서는 텍스트를 숫자형 시퀀스로 변환해야 한다. 음성 데이터와 마찬가지로 클래스 불균형 문제를 해결하기 위해 EAD(Easy Data Augmentation) 방법을 이용해 우울증 데이터를 3개씩 확장하였다. 전처리 된 텍스트 데이터는 워드 임베딩 계층이 포함된 Bi-LSTM 모델 및 n-gram 개념을 이용한 CNN 모델에 적용하였다.

실험은 전체 데이터의 80%는 학습 데이터로 나머지 20%는 검증데이터로 이용하여 진행되었다. 음성신호는 MFCC, GTCC 특징추출 방법을 이용해 1차원 특징을 추출하고 2차원 시간-주파수 변환 기반 Bark, ERB, Log-Mel 스펙트로그램 특징을 추출한다. 1차원 특징을 Bi-LSTM 모델에 적용한 결과 음성 데이터의 잡음을 제거 및

확장 후 성능이 가장 좋은 방법을 확인하였다. 특히 은닉유닛 개수가 10이고 MFCC 및 GTCC 특징을 같이 사용할 때 96.34% 가장 높은 성능을 보였다. 2차원 특징을 CNN 기반 전이학습 모델인 VGGish, YAMNet, OpenL3에 적용한 결과 음성 데이터의 잡음을 제거 및 확장 후 성능이 가장 좋은 것을 확인하였다. 특히 Log-mel 스펙트로그램 특징의 RGB 이미지를 입력으로 넣었을 때 OpenL3 모델의 성능이 96.34% 가장 높은 성능을 보였다. 텍스트 데이터를 이용한 실험에서 BI-LSTM은 데이터 확장 후 전처리를 토큰화만 진행하고 임베딩 차원이 200일 때 76.83%의 성능을 확인할 수 있었다. CNN 모델은 데이터 확장 후 토큰화만 진행하고 임베딩 차원이 100일 때 81.71%로 가장 높은 성능을 보였다.

음성 및 텍스트 데이터로부터 late score fusion 방법을 이용한 4-stream 기반 딥러닝 모델의 성능은 98.78%로 단일 모델의 성능보다 1.22%에서 2.44% 정도 더 개선됨을 확인할 수 있었다. 기존 연구와 제안된 4-stream 기반 딥러닝 모델의 성능을 같은 조건에서 비교하기 위해 DAIC-WOZ 데이터베이스를 이용한 실험을 진행하였다. 실험은 가장 좋은 성능을 가지는 방법 및 딥러닝 모델을 이용하여 진행하였고 late fusion 방법을 곱하기로 하였을 때 96.67%로 가장 높은 성능을 보여준다. 이는 기존 2-stream 기반 우울증 진단모델들과 비교했을 때 가장 좋은 성능을 가지는 것을 확인할 수 있었고 결과적으로 제안된 4-stream 기반의 딥러닝 모델이 우울증 진단에 효과적임을 증명하였다. 향후에는 음성 및 텍스트뿐만 아니라 뇌전도(EEG), 표정 등 다양한 멀티모달 데이터를 활용한 우울증 진단을 연구할 계획이다. 또한 우울증 여부를 진단하는 데 그치지 않고 우울증을 진단하고 심각성을 예측하는 방법에 대해 연구할 예정이다.

참고문헌

[1] J. M. Cénat, C. Blais-Rochette, C. K. Kokou-Kpolou, P. G. Noorishad, J. N. Mukunzi, S. E. McIntee, R. D. Dalexis, M. A. Goulet, P. R. Labelle, “Prevalence of symptoms of depression anxiety insomnia posttraumatic stress disorder and psychological distress among populations affected by the COVID-19 pandemic: A systematic review and meta-analysis”, Psychiatry Research, vol. 295, pp. 113599, 2020.

[2] M. T. Jordan, D. M. Kazemi, “COVID-19's impact on the mental health of older adults: Increase in isolation depression and suicide risk. An urgent call for action”, Public Health Nursing, Vol. 37, No. 5, pp. 637-638, 2020.

[3] M. Marcus, M. T. Yasamy, M. V. Ommeren, D. Chisholm, “Depression: A global public health issues”, APA PsycNet, pp.6-8, 2012.

[4] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, M. Breakspear, "Head pose and movement analysis as an indicator of depression", 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 283-288, 2013.

[5] J. K. Darby, N. Simmons, P. A. Berger, "Speech and voice parameters of depression: A pilot study", J Commun Disord, vol. 17, no. 2, pp. 75-85, 1984.

[6] S. M. Lamers, K. P. Truong, B. Steunenberg, F. de Jong, G. J. Westerhof, "Applying prosodic speech features in mental health care: An exploratory study in a life-review intervention for depression", Proc. NAACL, pp. 61-68, 2014.

[7] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, A. Othmani, “MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech”, Preprint submitted to Signal Processing: Image Communication, pp. 1-14, 2019.

[8] Y. Zhao, Z. Liang, J. Du, L. Zhang, C. Liu, L. Zhao, “Multi-head attention based Long Short-Term Memory for depression detection from speech”, Front. Neurorobot., Vol.15, No.684037, pp.1-11, 2021.

[9] D. Sztahó, K. Gábor, T. M. Gábriel, “Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning”, In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, Vol. 4, pp. 135-141, 2021.

[10] L. He, C. Cao, “Automated depression analysis using convolutional neural networks from speech”, ELSEIVER Journal of Biomedical Informatics, Vol. 83, pp. 103-111, 2018.

[11] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, “DepAudioNet: An efficient deep model for audio based depression classification”, AVEC'16: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 35-42, 2016.

[12] A. V. Romero, A. G. Antolín, “Automatic detection of depression in speech using ensemble Convolutional Neural Networks”, Entropy 22, Vol. 22, No.6, pp. 1-17, 2020.

[13] L. Yang, D. Jiang, H. Sahli, “Feature augmenting networks for improving depression severity estimation from speech signals”, IEEE Access, Vol. 8, pp. 24033-24045, 2020.

[14] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, H. Sahli, "Multimodal measurement of depression using deep learning models", AVEC '17: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 53-59, 2017.

[15] L. Yang, D. Jiang and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis hybrid architectures", IEEE Transactions on Affective Computing, Vol. 12, No. 1, pp. 239-253, 2021.

[16] M. Muzammela, H. Salamb, Y. Hoffmannc, M. Chetouanid, A. Othmania, “AudVowelConsNet: A phoneme-level based deep CNN architecture for

clinical depression diagnosis", Machine Learning with Applications, Vol. 2, pp. 1-12, 2020.

[17] R. Flores, M. Tlachac, E. Toto, E. A. Rundensteiner, "Depression screening using deep learning on follow-up questions in clinical interviews", 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 595-600, 2021.

[18] V. Ravi, J. Wang, J. Flint, A. Alwan, "Fraug: A frame rate based data augmentation method for depression detection from speech signals", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6267-6271, 2022.

[19] M. Rambocas, J. Gama, "Marketing Research: The role of sentiment analysis", The 5th SNA-KDD Workshop11. University of Porto, 2013.

[20] S. J. Park, S. B. Lee, W. J. Kim, M. Song, "A deep learning-based depression trend analysis of korean on social media", 정보관리학회지, Vol. 39, No. 1, pp. 91-117, 2022.

[21] A. H. Orabi, P. Buddhitha, M. H. Orabi, D. Inkpen, "Deep learning for depression detection of twitter users", Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 88-97, 2018.

[22] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek, R. Shil, Md. H. Kabir, "Early depression detection from social network using deep learning techniques", 2020 IEEE Region 10 Symposium(TENSYMP), pp. 823-826, 2020.

[23] D. E. Losada, F. Crestani, J. Parapar, "eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations", International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 346-360, 2017.

[24] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, M. Uddin, "Deep learning for depression detection from textual data", Electronics, Vol. 11, No. 5, pp.1-13, 2022.

[25] A. Shankdhar, R. Mishra, N. Shukla, “An application of deep learning in identification of depression among twitter users”, International Conference on Innovative Computing and Communications, Vol. 3, pp.661-669, 2022.

[26] J. H, Park, N. M. Moon, “Design and implementation of attention depression detection model based on multimodal analysis”, Sustainability, Vol. 14, No. 6, pp. 1-15, 2022.

[27] J. Xiao, Y. Huang, G. Zhang, W. Liu, “A deep learning method on audio and text sequences for automatic depression detection”, 2021 3rd International Conference on Applied Machine Learning(ICAML), pp. 388-392, 2021.

[28] M. Niu, K. Chen, Q. Chen, L. Yang, “HCAG: A Hierarchical Context-Aware Graph attention model for depression detection”, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 4235-4239, 2021.

[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y Bengio, "Graph attention networks", International Conference on Learning Representation, pp. 1-12, 2018.

[30] Y. Shen, H. Yang, L. Lin, “Automatic depression detection: an emotional audio-textual corpus and a Gru/Bilstm-based model”, IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 6247-6251, 2022.

[31] C. Lau, W. Y. Chan, X. Zhu, “Improving depression assessment with multi-task learning from speech and text information”, 2021 55th Asilomar Conference on Signals, Systems, and Computers, pp. 449-453, 2021.

[32] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, H. Sahli, “Multimodal measurement of depression using deep learning models”, AVEC '17: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 52-59, 2017.

[33] T. Alhanai, M. Ghassemi, J. Glass, “Detecting depression with audio/text sequence modeling of interviews”, Interspeech 2018, pp. 1716-1720, 2018.

[34] L. Lin, X. Chen, Y. Shen, L. Zhang, “Towards automatic depression detection: A BiLSTM/1D CNN-based model”, Applied Sciences, Vol. 10, No. 23, pp. 1-20, 2020.

[35] G. Lam, H. Dongyan, W. Lin, “Context-aware deep learning for multi-modal depression detection”, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 3946-3950, 2019.

[36] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time-frequency audio features”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, No. 6, pp.1142-1158, 2009.

[37] J. M. Liu, M. You, G. Z. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, S. Chen, “Cough signal recognition with Gammatone Cepstral Coefficients”, 2013 IEEE China Summit and International Conference on Signal and Information Processing, pp. 160-164, 2013.

[38] E. Zwicker, “Subdivision of the audible frequency range into critical bands”, The Journal of the Acoustical Society of America, Vol. 33, No. 2, pp. 248, 1961.

[39] P. Torben, “Acoustic Communication. Hearing and Speech. Version 2.0”, Online Research Database In Technology, pp.1-94, 2005.

[40] B. R. Glasberg, B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data”, Hearing Research, Vol.47, No.1-2, pp.103-138, 1990.

[41] 신동현, “3D 로그 멜-스펙트로그램에 dilated CNN과 attention based sliding LSTM을 적용한 음성 감정 인식”, 국내석사학위논문 한양대학교 대학원, 2020.

[42] S. Hochreiter, J. Schmidhuber, “Long Short-term Memory”, Neural Computation, Vol. 9, No. 8, pp.1735-1780, 1997.

[43] S. Hershey, S. Chaudhuri, Daniel P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. Wilson, "CNN architectures for large-scale audio classification", IEEE International Conference on Acoustics Speech and Signal Processing, pp. 131-135, 2017.

[44] Z. S. Syed, S. A. Memon, A. L. Memon, "Deep acoustic embeddings for identifying parkinsonian speech", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 11, No. 10, pp. 726-734, 2020.

[45] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv preprint arXiv:1704.04861, pp. 1-9, 2017.

[46] J. Cramer, H. H. Wu, J. Salamon, J. P. Bello, "Look, Listen, and Learn more: Design choices for deep audio embeddings", 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2019.

[47] R. Arandjelovic, A. Zisserman, "Look, Listen and Learn", 2017 IEEE International Conference on Computer Vision(ICCV), pp. 609-617, 2017.

[48] A. Coifman, P. Rohoska, M. S. Kristoffersen, S. E. Shepstone, Z. H. Tan, "Subjective annotations for vision-based attention level estimation", In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(VISIGRAPP 2019), Vol. 5, pp. 249-256, 2019.

[49] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. M. Messner, S. Song, S. Liu, Z.Zhao, A. M. Ragolta, Z. Ren, M. Soleymani, M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition", In Proceedings of the 9th International on Audio/Visual Emotion

Challenge and Workshop, pp. 3-12, 2019.

[50] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L. P. Morency, "The distress analysis interview corpus of human and computer interviews", In Proceedings of LREC 2014, pp. 3123-3128, 2014.

[51] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, L. P. Morency, "SimSensei kiosk: A virtual human interviewer for healthcare decision support", In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems(AAMAS' 14), pp. 1061-1068, 2014.

[52] S. S. Dhingra, K. Kroenke, M. M. Zack, T. W. Strine, L. S. Balluz, "PHQ-8 Days: a measurement option for DSM-5 Major Depressive Disorder(MDD) severity", Population health metrics, Vol. 9, No. 11, 2011.

[53] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis", PLOS ONE, pp. 1-17, 2015.

[54] J. Wei, K. Zou, "EDA: Easy Data Augmentation techniques for boosting performance on text classification tasks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6382-6388, 2019.