creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학박사 학위논문

# 음성과 텍스트를 이용하여 우울증 및 자살 위험을 평가하는 인공지능 기반 임상의사결정지원시스템에 관한 연구

A study on artificial intelligence-based clinical decision support system to evaluate depression and suicide risk using voice and text

2022년 2월

서울대학교 대학원

의학과 정신과학 전공

신 다 운

# 음성과 텍스트를 이용하여 우울증 및 자살 위험을 평가하는 인공지능 기반 임상의사결정지원시스템에 관한 연구

지도 교수 안 용 민

이 논문을 의학박사 학위논문으로 제출함

2021년 10월


서울대학교 대학원

의학과 정신과학 전공

신 다 운


신다운의 의학박사 학위논문을 인준함

2022년 1월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

# 초 록

**서론:** 전 세계적으로 우울증과 자살은 발병률이 지속적으로 증가하고 있으며, 그로 인한 사회경제적 손실이 막대하다. 하지만 이를 진단하는 방법은 환자의 주관적인 대답을 바탕으로 진단하는 방법 뿐이다. 그로 인해 환자가 증상을 축소하여 보고하거나 심층적인 면담을 진행하기 어려운 상황에서는 정확한 진단이 어렵고, 효율적인 개입을 하기 어렵다. 따라서 우울증과 자살을 진단할 수 있는 객관적인 마커들에 대한 다양한 연구들이 진행되고 있다. 그 중 인터뷰 시의 목소리와, 사용하는 단어 등은 임상의가 축적된 임상경험을 바탕으로 내리는 임상적 판단에 많이 활용되었던 지표들이다. 최근 목소리의 다양한 지표를 추출하고 발화 단어를 추출하는 기술들이 개발됨에 따라서 우울증과 자살 위험에 따른 목소리의 차이와 발화 단어들의 차이들이 밝혀지고 있다. 또한 인공지능이 의료계에 접목됨에 따라서 이러한 미세한 차이들을 감지하고 임상적 의사 결정에 도움을 줄 수 있게 되었다. 하지만 국내외의 선행 연구들에서는 아직 피험자의 수가 충분하지 않고, 약물 등의 다양한 임상적 상황들에 대한 고려가 부족하다는 한계가 있다. 따라서 본 연구에서는 기존의 한계들을 극복하여, 인터뷰 중의 피험자 목소리와 사용된 단어를 기반으로 우울증과 자살 위험군을 평가하는 인공지능 기반 임상의사결정지원시스템을 구축해보고자 한다.

**방법:** 우울한 증상을 호소하는 환자군과 정상대조군을 각각 모집하였고, 모든 피험자에게 간이 국제 신경 정신 인터뷰(Mini International Neuropsychiatric Interview)를 시행하여 해당 인터뷰를 녹음하였다.

녹음된 인터뷰 파일 중 피험자가 발화한 구간만을 추출한 뒤, 각 구간에서 목소리의 다양한 지표들과 텍스트 데이터를 추출하였다. 연구 I에서는 2차 년도까지의 초기 데이터를 이용하여 정상군과 경도 우울증군, 주요 우울증군을 비교하여 분석하였다. 그리고 연구 II와 III을 통해 최종적으로 모집된 초기 평가 데이터를 활용하여 우울증과 자살 위험을 평가하는 인공지능 기반 임상의사결정시스템을 구축해보고자 하였다. 자살 위험은 빈도가 작기 때문에, 자살 위험을 정확하게 평가하기 위하여 연구 II를 통해 음성과 텍스트를 이용하여 정상과 우울증을 구별할 수 있는 평가 알고리즘을 구축하였다. 그 후 2단계로서 연구 III을 통해 우울증 군 내에서 자살 저위험군과 자살 고위험군을 구분하여 자살 위험을 평가하는 알고리즘을 구축하여 그 성능을 확인하였다.

**결과:** 2차 년도까지 모집된 정상군 33명과 경도 우울증군 26명, 주요 우울증군 34명을 대상으로 분석한 연구 I의 결과, 세 군을 구별할 수 있는 음성 지표 7개를 발견하였다. 또한 음성 지표를 기반으로 세 군을 평가하는 진단 알고리즘을 다양한 모델을 통해 구축하였고, 다층 퍼셉트론의 성능이 가장 우수함을 확인하였다. 3차 년도까지 최종적으로 모집된 환자는 총 85명이었고, 그 중 자가보고설문지를 누락한 2명은 분석에서 제외하였다. 또한 정상대조군은 학내 게시판과 온라인 광고를 통해 총 105명이 모집되었으며 정신건강의학과 병력이 있는 22명은 분석에서 제외되었다. 이를 기반으로 진행된 연구 II에서는 음성 지표와 텍스트데이터를 분석하여 정상군과 우울증군을 구분할 수 있는 인공지능 기반 알고리즘을 구축하였고, 음성을 기반으로 한 인공지능 알고리즘의 곡선하면적은 0.806, 텍스트데이터를 기반으로 한 인공지능 알고리즘의 곡선하면적은 0.905로 우수한 성능을 확인하였다. 이어 연구 III의

우울증군 83명을 대상으로 자살 고위험군을 구분하는 인공지능 임상의사결정지원시스템을 구축하였다. 자살 고위험군을 구분하는 방법으로 각각 벡 자살사고 척도 기준의 방법과, 간이 국제 신경정신 인터뷰를 통해 분류하는 방법을 통해 구분하는 방법을 모두 적용하였다. 그 결과 음성지표를 기반으로 자살 고위험군을 평가하는 알고리즘의 경우 민감도 0.535가 최대값이었으며, 가장 우수한 성능은 로지스틱 회귀공식을 이용한 모델에서 보인 평균 정확도 0.495였다. 또한 텍스트를 기반으로 자살 위험을 평가한 알고리즘 역시 곡선하면적 0.632에 불과하여, 목소리와 텍스트를 기반으로 한 자살 고위험군 진단 알고리즘은 임상적으로 활용하기는 어려운 수준이었다. 하지만 텍스트 데이터와 사회인구학적 정보를 통합하여 구축한 앙상블 모델의 곡선하면적은 0.800으로 성능이 향상됨을 확인하였다.

**결론:** 본 연구는 구조화된 면담에서 피험자의 발화 구간을 통해 음성과 텍스트 데이터를 추출하여, 이를 기반으로 우울증과 자살 위험성을 평가하는 임상의사결정지원 알고리즘을 국내에서는 최초로 구축하였다. 우울증을 진단함에 있어서는 두 데이터 모두 우수한 성능을 보였으나, 자살 위험을 진단하기에는 부족하였다. 텍스트 데이터의 경우 구조화된 인터뷰를 통해 얻어진 텍스트라는 한계가 있음에도 불구하고 우울증의 진단함에 있어 임상의사결정지원시스템의 가능성을 확인하였다. 더욱이 자살 위험 진단에 있어서도 사회인구학적 정보와 통합된 앙상블모델의 우수한 성능을 확인함으로써 향후 임상적 활용가치를 보여주었다. 본 연구는 목소리, 텍스트 데이터가 가지는 객관적 진단 가치를 국내 최초로 입증한 연구로서, 디지털 진단 도구라는 정신건강의학과적으로는 새로운 분야로의 도전이다. 향후 해당 분야의 더 다양한 지역, 다양한 환경에서의 데이터에 대해서 추가적인 연구가 필요하다.

* 본 내용의 일부는

Shin, D., Cho, W. I., Park, C., Rhee, S. J., Kim, M. J., Lee, H., ... & Ahn, Y. M. (2021). Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning. Journal of clinical medicine, 10(14), 3046.

에 출판 완료된 내용이며, 그 외 내용들은 현재 출판 준비 중임.

**주요어** : 우울증, 자살 위험, 목소리, 텍스트, 기계학습, 임상의사결정지원시스템

**학 번** : 2018-25300

# 목 차

초록 ..... i

표 목차 ..... vi

그림 목차 ..... ix

약어 목록 ..... x

제 1 장 서 론 ..... 1

제 2 장 목소리를 통해 경도 우울증과 주요 우울증을 구별하는 인공지능 알고리즘의 개발 (연구 I) ..... 10

제 1 절 연구대상 및 방법 ..... 11

제 2 절 결과 ..... 22

제 3 절 고찰 ..... 26

제 3 장 목소리와 텍스트 데이터를 이용하여 우울증 진단 인공지능 알고리즘의 개발 (연구 II) ..... 31

제 1 절 연구대상 및 방법 ..... 33

제 2 절 결과 ..... 38

제 3 절 고찰 ..... 43

제 4 장 목소리와 텍스트 데이터를 통해 자살고위험군을 진단하는 인공지능 알고리즘 개발 (연구 III) ..... 49

제 1 절 연구대상 및 방법 ..... 51

제 2 절 결과 ..... 54

제 3 절 고찰 ..... 59

제 5 장 종합 고찰 ..... 65

참고문헌 ..... 128

Abstract ..... 152

# 표 목차

[표 1-1] **Comparison of demographics according to depressive episodes** ..... 78

[표 1-2] **Clinical characteristics by depressive episode** ..... 80

[표 1-3] **Difference of voice features by depressive episode** ..... 81

[표 1-4 **Jonckheere–Terpstra test result of voice features by depressive episode** ..... 83

[표 1-5] **Machine learning model performance through voice features** ..... 85

[표 1-6] **Comparison of demographics between healthy control and major depressive disorder and bipolar disorder** ..... 86

[표 1-7] **Comparison of clinical characteristics between healthy control and major depressive disorder and bipolar disorder** ..... 87

[표 1-8] **Voice and speech features between healthy control and major depressive disorder and bipolar disorder** ..... 88

[표 2-1] **Descriptive statistics of voice and speech features** ..... 90

[표 2-2] **The results of applying the diagnostic algorithm from Study I to the newly recruited test set from Study II** ..... 91

[표 2-3] **Demographics comparison between healthy control group and current depression group** ..... 92

[표 2-4] **Clinical characteristics between HC and CD** ..... 93

[표 2-5] **Voice and speech features comparison between healthy control and current depression** ..... 94

[표 2-6] **Classification results by voice and speech features between HC and CD** ..... 96

[표 2-7] **Classification results by text between HC and CD** ..... 97

[표 2-8] **Correlation analysis between voice and speech features and clinical characteristics in healthy control group** ..... 98

[표 2-9] **Correlation analysis between voice and speech features and clinical characteristics in current depression group** ..... 100

[표 2-10] **Mediation analysis between voice and speech features and clinical characteristics in healthy control group** ..... 102

[표 2-11] **Mediation analysis between voice and speech features and clinical characteristics in current depression group** ..... 103

[표 3-1] **Demographics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : BSS)** ..... 104

[표 3-2] **Clinical characteristics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : BSS)** .... 105

[표 3-3] **Voice and Speech features between DLSR and DHSR (suicidal risk : BSS)** ..... 106

[표 3-4] **Classification results by voice and speech features between DLSR and DHSR (suicidal risk; BSS)** ..... 108

[표 3-5] **Classification results by voice and speech features between DLSR and DHSR (suicidal risk : BSS)** ..... 109

[표 3-6] **Demographics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : MINI)** ..... 110

[표 3-7] **Clinical characteristics between low suicidal risk and depression with high suicidal risk (suicidal risk : MINI)** ..... 111

[표 3-8] **Voice and Speech features between DLSR and DHSR (suicidal risk : MINI)** ..... 112

[표 3-9] **Classification results by voice and speech features between DLSR and DHSR (suicidal risk; MINI)** ..... 114

[표 3-10] **Classification results by voice and speech features between DLSR and DHSR (suicidal risk : MINI)** ..... 115

# 그림 목차

[그림 1-1] **Difference of voice features by depressive episode**..... 116

[그림 1-2] **AUC curve predicting minor and major episodes using MLP**..... 117

[그림 2-1] **ROC curve of classification HC and CD by voice and speech features** ..... 118

[그림 2-2] **Feature importance for classifying HC and CD** ..... 119

[그림 2-3] **Text mining of important features in HC and CD** .... 120

[그림 2-4] **Classification results between HC and CD by text**.... 121

[그림 2-5] **Mediation analysis between voice and speech feature and clinical characteristics in healthy control group** ..... 122

[그림 2-6] **Mediation analysis between voice and speech feature and clinical characteristics in current depression group**..... 123

[그림 3-1] **ROC curve of classification DLSR and DHSR by voice and speech features (suicidal risk : BSS)**.....124

**[그림 3-2] Classification results of DLSR and DHSR by text and clinical characteristics (suicidal risk : BSS)** .....125

**[그림 3-3] ROC curve of classification DLSR and DHSR by voice and speech features (suicidal risk : MINI)**.....126

**[그림 3-4] Classification results of DLSR and DHSR by text and clinical characteristics (suicidal risk : MINI)**..... 127

# 약어 목록

AN: anorexia nervosa

ANOVA: analysis of variance

ANCOVA: analysis of covariate

AP: antipsychotics

AUC: area under curve

AUD: alcohol use disorder

BAI: Beck Anxiety Inventory

BHS: Beck hopelessness scale

BIS: Barratt Impulsiveness Scale-11

BMI: body mass index

BN: bulimia nervosa

BP: bipolar disorder

BSS: beck’s suicide ideation scale

BW: bandwidth

B-H test: Benjamini-Hochberg test

HDRS: Hamilton Depression Rating Scale

C: closed phased

CD: current depression group

CP: closing phase

DHSR: depression with high suicidal risk

DLSR: depression with low suicidal risk

DSM-V: Diagnostic and Statistical Manual of Mental Disorders,

Fifth Edition

F: female

GAD: generalized anxiety disorder

GNB: Gaussian Naïve Bayes

HC: healthy control

ICD-10: International Classification of Diseases and Related Health Problems, 10th Edition

IQ: intelligence quotient

J-T: Jonckheere-Terpstra

LR: logistic regression

M: male

MANCOVA: Multivariate analysis of covariance

MDD: major depressive disorder

MDE: major depressive episode

MINI: Mini International Neuropsychiatric Interview

MLP: multi-layer perceptron

m: magnitude

mDE: minor depressive episode

N: number

ND: not depressed

OCD: obsessive-compulsive disorder

OECD: Organisation for Economic Co-operation and Development

OP: opening phase

OZP: olanzapine

PHQ: Patient Health Questionnaire-9

PTSD: post-traumatic stress disorder

p: pitch

RFL: reason for leaving scale

ROC: receiver operating characteristic

SD: Standard deviation

SE: Standard error

SES: social economic status

Sq: squared

Stdev: standard deviated

SVM: support vector machine

ZCR: zero crossing rate

# 제 1 장 서 론

세계보건기구(WHO)에 따르면 2017년 기준 전 세계적으로 3억 2,200만 명이 우울증을 앓고 있으며, 우울증은 비치명적 건강상실의 주요 원인이다(1). Leptine에 따르면 일반 인구에서 우울증의 평생 유병률은 10%에서 15% 사이인 것으로 보고되었고, 그 유병률 역시 매년 빠르게 증가하고 있다(2). 또한 2030년에는 단극성 우울증이 전 세계 사망의 두 번째 주요 원인이 될 것으로 예측되기도 한다(3). 매우 높은 유병률과 빠른 증가추세를 보이는 우울증으로 인한 사회경제적 부담 역시 전세계적으로 빠르게 증가하고 있다. 특히 2000년부터 우울증은 비치명적인 이유들 중에서 가장 높은 사회적 부담을 야기한다고 논의되고 있다(4). 또한 세계 각국에서도 우울증으로 인한 사회경제적 부담이 빠르게 증가하고 있다(5-8). 국내에서 우울증 환자의 사회경제적 부담 역시 매우 높다. 우울증으로 인한 직접의료비용 및 간접의료비용을 총 합한 비용은 약 4조 8041억에 다다르며(9), 한 해 우울증과 직접적 연관이 있는 직접의료비용만 800만원 가량으로 추정된다(10). 또한 우울증 환자의 자살 위험은 일반 인구의 20배 이상이며 우울증 환자의 약 15~20%가 자살로 생을 마감하는 것으로 알려져 있다(11).

자살은 역시 전세계적으로 매년 80만건이 발생하며, 많은 유럽 및 북미 및 아시아 태평양 국가에서 자살이 주요 사망 원인 10위 안에 있다(12). 또한 26년동안 전세계적으로 꾸준히 사망률이 증가하고 있는 상황이다(13). 특히 국내의 경우 2003년도부터 2016년도에 이르는 14년 동안 인구 10만

명 당 자살로 인한 사망자수인 연간 자살률이 경제개발협력기구(Organisation for Economic Co-operation and Development, OECD) 가입국 가운데 가장 높은 수치를 기록하였다. 2017년도의 경우는 2위로 낮아졌지만, 2018년도에는 OECD 평균인 12.1명의 2배를 상회하는 26.6명으로 나타나 다시 1위가 되었다(14). 또한 자살시도는 입원기간의 많은 의료비용을 초래하며, 장기 치료가 필요한 영구 장애와 수입의 손실 등을 겪게 되어 자살 한 건으로 인한 사회적 손실은 미국 기준으로 무려 4억원 가량으로 추정된다(15).

따라서 자살과 우울증을 조기에, 정확하게 진단하는 것은 효율적 개입으로 연결될 수 있는 아주 중요한 문제이다. 하지만 현재로서는 자살과 우울증의 진단은 오직 환자의 주관적 증상의 보고에 의존하고 있다. 전세계적으로 Diagnosis of depression is made by Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) 이나 International Classification of Diseases and Related Health Problems, 10th Edition (ICD-10)이 우울증을 진단함에 있어서 가장 최적의 기준으로 알려져 있다. 정신 장애 진단 및 통계 편람 제5판(DSM-V)에 따르면 주요 우울 장애는 (1) 우울한 기분 또는 (2) 흥미나 즐거움의 상실 중 하나 이상을 포함하여 5가지 이상의 다른 우울 증상이 있고, 그 증상들이 2주 이상 지속될 때 진단될 수 있다(16). 국제질병분류 (ICD-10) 역시 DSM과 유사하며, 환자의 주관적인 증상의 개수와 지속기간을 바탕으로 진단을 내리도록 권고한다(17). 이러한 기준에 따른 진단 방법은 증상을 축소 보고하는 경향을 가진 사회에서 우울증의 유병률을 낮추는 원인 중 하나이다. 실제로 미국, 유럽 등의 국가보다 우리나라와 일본 등 동아시아의 국가에서

주요우울장애의 유병률은 절반 가까이 낮게 측정되기도 한다(18-20). 또한 증상에 대한 경험이 많지 않거나, 심층적인 인터뷰를 진행하기 어려운 1차 의료 환경에서는 종종 오진이 내려지거나 우울증이 있음에도 진단을 받지 못하는 과소진단(under-diagnosis)이 발생하기도 한다(21, 22). 특히 메타분석 결과 일반의가 우울증을 진단함에 있어서 민감도는 50.1%에 불과하였다(23). 더욱이 자살의 경우, DSM-5에서 추후 연구를 위한 정의로서 '자살행동장애'를 정의하고 있으나 이것이 실제 죽음의 의도를 어떻게 평가하는지에 대해서는 구체적으로 명시되어 있지 않다(16). 또한 자살 위험을 평가하기 위한 다양한 척도가 개발되었지만 아직까지 가장 정확한 척도가 무엇인지조차 확립되지 않았다(24,25). 특히 한국은 자살에 대해 사회적으로 터부시 여기고 있어 자살로 사망한 사람들에 대해 사회적 낙인이 붙게 되고, 실제로 그러한 영향은 자살통계에까지 광범위한 영향을 미친다(26). 이러한 사회적 배경을 고려해볼 때, 자살 위험을 평가에 있어 환자는 자살 위험의 보고를 축소할 가능성이 있기 때문에 임상의는 환자의 대답 뿐 아니라, 표정, 행동, 말투, 눈맞춤 등 다양한 요소들까지 복합적으로 받아들이고 평가에 임해야 한다(27, 28). 하지만 대부분의 자살 위험에 대한 평가들은 응급실이나 콜센터 등에서 이뤄지게 되지만 이러한 환경은 정신건강의학과 인터뷰를 충분히 시행할 수 없는 시간적 한계와 더불어 응급실에서 근무하는 당직의사나 콜센터의 상담가 등은 정신과적 임상 경험이 풍부하지 않을 가능성이 있다. 이러한 한계들을 극복하고자 다양한 상황에서 우울증과 자살을 정확하고 객관적으로 진단할 수 있는 다양한 지표들에 대한 도전이 전세계적으로 지속되고 있다.

특히 전통적으로 정신건강의학과 평가에 있어서 가장 중요한 방법은

면담, 즉 인터뷰이다(29). 인터뷰를 통해서는 환자와의 대화 내용 뿐 아니라 비언어적인 정보들인 전반적인 모습부터 행동, 표정, 눈빛, 목소리 등의 다양한 정보들을 획득하고, 이를 임상적 경험을 토대로 진단 및 평가에 활용하게 된다. 다양한 비언어적 정보들 중에서도 목소리는 1921년 에밀 크레펠린에 의해 "우울한 환자들은 낮은 목소리로, 천천히, 주저하면서, 단조롭게, 때때로 말을 더듬고, 속삭이기도 하며, 말을 꺼내기 전에 여러 번 망설이고, 문장 중간에 말이 끊기기도 한다" 라고 우울한 목소리에 대해 언급되었다(30). 1970년부터 공학적 기술들이 발달하면서 목소리의 다양한 지표들을 수치로 객관화할 수 있게 되면서, 우울증이 있는 경우 발화 중간의 멈춤이 길어지고, 발화량이 줄어든다는 것이 객관적으로도 입증되었다(31, 32). 또한 공학의 기술이 발전함에 따라 목소리의 더욱 다양한 지표들에 대한 연구가 진행되었고, 음성지터 및 성문 스펙트럼 등의 다양한 지표들을 통해 우울증이 있는 경우의 목소리 변화들에 대해서 확인되었다(33). 더욱이 임상적으로 이러한 목소리의 미세한 변화를 임상의가 인지하고 판단에 적용하기 위해서는 많은 시간과 노력을 들인 임상 경험을 필요로 하지만, 인공지능이 발달하면서 인간이 감지하지 못했던 미묘한 차이를 대신 감지할 수 있게 되었다. 인공지능은 대규모의 데이터분석을 통해 많은 임상적 경험을 대체할 수 있게 되면서 목소리가 우울증이나 자살 위험에 대한 객관적 지표일 가능성이 제기되었다(34). 실제로 단순히 우울한 상태나 자살 위험에 따라 목소리의 변화만을 확인하는 것에 그치지 않고, 목소리의 변화를 인공지능에 학습시킴으로써 우울한 상태나 자살 위험을 진단할 수 있는지에 대한 연구들이 진행되었다. 목소리 사이의 공백 패턴 등 다양한 지표들을 이용하여 우울증과 자살 위험을 예측하는 연구들에 있어서 대체로 그

정확도는 75~95% 가량을 보였다(35-38). 자살 위험을 예측한 연구는 그보다 더 적었지만, 그 정확도 역시 75% 가량을 확인할 수 있었다(39,40). 하지만 이러한 연구들은 대체로 공학적인 기술에 대한 도전으로서, 임상적 지표들에 대한 고려가 부족하였다. 특히 우울하거나 자살 위험도가 높은 환자들은 많은 경우 항정신병약제를 복용하고 있으며, 항정신병약제는 추체외로 작용을 통해 목소리의 톤을 균일하게 만드는 등의 영향을 미친다(41,42). 하지만 이러한 약제에 대한 고려가 있었던 연구는 부재하였으며, 연구의 피험자 수 역시 16명에서 30명 가량으로 매우 적다는 한계를 가지고 있다. 또한 우울증상에 대한 면밀한 평가가 이뤄지지 않고 벡의 우울척도 등과 같은 척도의 점수를 기준으로 우울증을 진단하는 경우가 많았으며, 동반되는 다른 질환에 대한 고려가 이뤄진 연구는 아직까지는 존재하지 않았다. 더욱이 국내에서 음성과 우울증 혹은 자살 위험과 관련된 연구는 2012년에 발표된 연구 1편 뿐으로, 이 연구에서는 11명의 우울증 여성 환자와 12명의 정상군의 음성을 분석하여 우울증 환자들의 평균 주파수의 음성 범위가 좁아져 있으며, 평균 발화 주파수와 음량의 범위는 벡의 우울척도와 음의 상관관계를 보인다는 것을 확인하였다(43). 하지만 이 연구 역시 피험자의 수가 매우 적고, 여성 환자만을 대상으로 하였다는 한계가 있다.

음성과 더불어 텍스트 역시 자살의 위험성을 감지할 수 있는 객관적 지표로서 관심을 받고 있다. 자살 사망자가 남긴 유서를 통해 실제 사망자와 시도자를 구별하는 연구에서도 실제 사망자가 사용하는 단어들은 조금 더 긍정적이며 미래지향적이었다(44). 또한 소셜 미디어에 게시된 글을 기반으로 하여 우울증과 더불어 자살 위험을 감지하는 연구들 역시 진행되고

있다(45-47). 이런 분석에는 자연어 처리 (natural language processing)이라는 분석 기법을 주로 이용되었고, 그 중에서도 의미 분석을 통해 게시된 글에 사용된 단어들의 의미를 바탕으로 자살의 위험성을 예측하는 방법이 많이 사용되었다. 하지만 텍스트 분석의 대부분은 이미 적혀져 있는 글, 즉 게시가 완료된 글이나, 유서 등 문자화가 완성된 데이터를 기반으로 분석을 한 연구들이 대부분이다. 즉 정신과적 면담이라는 실시간의 상황에서 환자가 사용하는 단어에 대한 연구는 부족한 현실이다. 공학적인 측면에서 면담 시 사용한 텍스트를 음성 등 다양한 지표들과 함께 활용하였던 연구가 있지만(49), 이 것은 실제 피험자와 임상가의 면담이 아닌, 면담에 사용되는 여러 질문 항목들을 모아 가상의 인터뷰를 하는 환경에서 획득한 정보들이다. 따라서 실제 임상의 임상의와의 상호작용 등이 반영되기 힘들고, 환자의 감정이 더욱 풍부하게 표현되기 어렵다는 한계가 있다. 특히 국내에서는 트윗에 사용된 텍스트와 영상 등으로부터 감성을 인식하는 방법을 통해 사용자의 감정 흐름을 예측한 연구(50)가 있었으나 피험자의 총 수가 40명이라는 한계가 있다.

더욱이 최근 인공지능이 눈부신 발전을 이루면서, 의학분야에서도 인공지능의 다양한 활용을 기대하고 있다. 인공지능이 의학적 판단을 대신 해줄 것이라는 처음 대중의 기대와 다르게, 현재까지 인공지능의 역할은 임상의사결정의 보조적인 역할을 하고 있다(51). 인공지능이 인간의 도움없이 의학적 결정을 단독으로 내릴 수는 없지만, 인공지능 기반의 임상의사결정지원시스템의 유용성과 효과는 다양한 의학 분야에서 인정받고 있다(52). 하지만 다른 과와는 다르게 정신건강의학과에서는 이러한 인공지능의 활용이 다소 더디게 진행되고 있다(53). 객관적인 기준이 있는

혈압 등의 질환과는 다르게 정신건강의학과의 질병들은 피험자의 주관적 증상과 사회적 고통이 심각하다는 기준을 두고 진단을 내리기 때문에 그 기준이 굉장히 모호하다는 특징이 있다(16, 17). 이러한 기준의 모호성으로 인하여 하나의 진단 속에도 다양한 병태생리와 증상들이 공존하게 되므로, 이를 인공지능에 학습을 시킨다 하더라도 그 결과값이 정확하기 어렵다는 한계가 생긴다. 이러한 한계로 인하여 정신건강의학과에서의 인공지능 연구는 굉장히 도전적이고 까다로운 분야이기도 하다. 그럼에도 불구하고 정신건강의학과에도 인공지능을 접목시키고자 하는 노력들은 시작되었고, 지속되고 있다. 특히 기존의 증상으로만 나누던 진단 체계에서 발전하여 다양한 생체지표들과 함께 고려한 체계로서 연구 도메인 진단 평가(research domain criteria, RDoC)가 제안되면서 다양한 생체 지표들을 진단에 활용하는 시도들이 시작되고 있다(54). 하지만 이러한 생체 지표들은 혈액 내 유전자 및 단백질과 뇌척수액의 단백질 등 지표의 채취해서 정보화하기에 시간이 많이 걸리기 때문에 즉각적인 도움을 줄 수 없다는 한계도 존재한다. 따라서 본 연구에서는 전문가의 인터뷰만큼 즉각적인 판단을 내리는데 도움을 줄 수 있는 생체지표로써 인터뷰시의 음성과 텍스트 데이터를 통해 우울증과 자살 위험을 평가할 수 있는 임상의사결정 시스템으로서 첫 도약을 내딛고자 한다.

본 연구는 국내에서는 최초로 면담시의 목소리와 사용된 텍스트 데이터를 활용하여 우울증과 자살 위험도를 진단하는 인공지능 알고리즘을 구축해보고자 하였다. 선행 연구들을 근거할 때, 우울증이 있는 환자에서는 발화시의 평균 주파수가 낮고, 발화가 멈춰 있는 비율이 높을 것으로 예상이 되며, 자살 위험도가 높은 환자에서는 긴장되어 있는 특징을 보이고 주파수의

변화폭 등이 적어져 있는 특징을 보일 것으로 가정하였다. 또한 인터뷰에서 사용되는 단어들 중에는 더 부정적 의미를 가지고 있는 단어가 많이 사용될 것으로 가설을 세웠다. 그리고 이러한 차이를 기반으로 인공지능을 통해 일부 데이터를 학습시켜 우울증과 자살 위험도를 진단할 수 있는 알고리즘을 구축하고, 남은 데이터를 통해 알고리즘의 정확도를 확인해보고자 하였다. 기존의 연구에서 많이 활용되었던 단순 과제를 통한 목소리 획득 방법보다는 임상가와 면담을 할 때의 피험자의 목소리에는 임상가와의 상호작용을 통해 피험자의 감정이 더 잘 반영되고, 긴장도가 낮아지며, 일상 생활의 언어가 더 많이 반영될 것으로 예상하였기 때문에, 면담 시 녹음을 통해 음성과 텍스트 정보를 추출하고자 하였다. 녹음을 진행하는 면담을 자유면담과 구조화된 면담 중에서 본 연구에서는 구조화된 면담을 진행하고자 하였고, 그 이유는 자유면담을 통해서는 우울증 환자군에서는 자발적 발화가 줄어들기 때문에 오히려 우울증군이 특징적으로 사용하는 단어들에 대한 반영이 줄어들 가능성이 있기 때문이다.

본 연구는 이러한 선행 연구들과 가설을 기반으로 진행된 “기계 학습을 기반으로 음성으로 자살 위험 예측(Predicting patients with suicidality: machine learning based voice analysis, Prism-V)” 이라는 연구의 일환으로서, 초기평가 자료를 통해 우울증을 감지하고, 자살 위험도를 진단할 수 있는 인공지능 알고리즘을 개발하여 그 정확도와 임상적 활용 가능성을 확인해보고자 하였다. 연구 I에서는 2차 년도까지의 초기 데이터를 이용하여 정상군과 경도 우울증, 중등도 우울증을 구분할 수 있는 음성 지표를 확인하고, 확인된 지표를 바탕으로 기계학습을 통해 알고리즘을 구축하고자 하였다. 경도 우울증이란 2주 이내의 단기간 우울증이거나, 주요 우울 장애로

진단하는 증상의 개수 등이 충족되지 않을 경우 진단되었다(55). 1992년 경부터 경도 우울증은 주요 우울 장애보다는 증상이 덜 심각하지만, 기능의 저하 및 동반되는 질환, 예후까지 모두 주요 우울증과 유사한 영향을 미친다는 것이 밝혀졌다(56). 더욱이 경미한 우울증은 기능저하 등을 통해 경제적, 사회적 부담에도 주요 우울장애과 비슷한 정도의 영향을 미친다(57,58). 또한 경도 우울증은 주요 우울장애의 위험인자이기도 하며, 사회적 기능의 악화를 예방하기 위해서 조기 진단할 수 있는 시스템을 필요로 한다(59-61). 따라서 연구 I은 주요 우울장애의 목소리 변화 뿐 아니라, 경도 우울증에서도 인터뷰 목소리에 차이가 발생하는지를 확인해보고, 경도 우울증에서도 정상군에서는 보이지 않는 목소리의 변화가 발생한다는 가설을 기반으로 경도 우울증을 진단할 수 있는 기계학습 알고리즘을 구축해보고자 하였다.

연구 II에서는 최종적으로 구축된 초기 평가 데이터를 기반으로 정상과 우울증 환자를 구별할 수 있는 음성의 지표를 확인하고, 더 나아가 인터뷰의 텍스트 분석을 시도하였다. 음성과 텍스트 데이터를 각각 이용하여 정상과 우울증을 진단할 수 있는 인공지능 알고리즘을 구축하여 두 알고리즘의 정확도를 확인해보고자 하였다. 이 것은 목소리 뿐 아니라 텍스트 데이터 자체가 우울증을 감지하기에 더 정확하다는 기존의 연구(62)를 바탕으로 텍스트 데이터와 목소리를 모두 활용하여 정확도를 확인해보고자 하였으며, 추후 텍스트 데이터와 목소리 모두를 통합할 수 있는 다중통합모델을 구축할 수 있는 가능성을 확인해보고자 하였기 때문이다.

마지막으로 연구 III에서는 연구 II의 초기 평가 데이터 중에서도 우울증 환자군의 데이터를 이용하여, 자살 고위험군과 저위험군을 구별하는 분석을

진행하였다. 자살로 사망한 경우 약 60%에서 90% 가량이 주요 우울장애 등의 기분 장애를 경험하였던 것으로 보고되었고(59-60), 본 연구에서 정상군의 경우 간이 국제 신경정신 인터뷰 (Mini International Neuropsychiatric Interview, MINI) 라는 구조화된 면담을 통해 정신과적 질환들과 자살사고 등에 대한 평가를 진행하였기 때문에 정상군에서 자살고위험군이 발생할 가능성이 매우 낮다. 또한 임상 현장에서 자살 위험성을 평가하는 환경을 고려해보면, 우울한 증상이 있으나 자살 위험이 높은지, 낮은지를 판단하는 것이 가장 어려운 임상적 난제이기 때문에 이러한 조건을 반영하고자 우울증 환자군 내에서 자살 고위험군을 구분할 수 있는 알고리즘을 구축해보고자 하였다. 연구 III 역시 목소리와 텍스트 데이터를 각각 활용한 알고리즘을 구축해보고자 하였으며, 텍스트 데이터의 경우에는 임상지표까지 통합한 모델을 구축하여 그 정확도를 확인해보고자 하였다.

이를 바탕으로 우울증과 자살 위험도를 임상 현장의 인터뷰를 통해 얻어진 목소리와 텍스트를 기반으로 평가할 수 있는 인공지능 알고리즘을 구축하고, 궁극적으로는 이 알고리즘을 기반으로 어플리케이션 등을 개발하여 임상 현장에서 위험성 평가에 도움이 되는 신의료기술의 기반을 설립하고자 하였다.

# 제 2 장 목소리를 통해 경도 우울증과 주요 우울증을 구별하는 인공지능 알고리즘의 개발 (연구 I)

연구 I에서는 2차 년도까지의 초기 데이터를 이용하여 정상군과 경도 우울증군, 주요 우울증군에서 목소리 차이에 대해 분석을 하고, 차이가 나는 지표들을 기계학습을 통해 학습시켜 3군을 구별할 수 있는 알고리즘을 구축하는 것이 그 목표이다. 경도 우울증은 주요 우울증의 위험요인 중 하나로서, 주요 우울증의 진단 기준에 적합한 증상의 정도나 개수, 기간 등을 충족시키지 않은 경우 진단할 수 있다(55). 경도 우울증은 그 증상은 비록 주요 우울증보다는 덜하지만, 사회적 손실 및 기능의 저하 등에 미치는 영향은 주요 우울증과 견줄만하다고 알려져 있다(56-60). 하지만 경도 우울증을 진단함에 있어서는 여전히 환자의 증상 보고를 기준으로 하고 있기에 경도 우울증을 진단함에 도움이 되는 객관적 지표들은 아직 부재하다. 따라서 본 연구에서는 목소리가 그 객관적 지표가 될 수 있을지 가능성을 알아보고자 한다. 연구 I의 결과는 추후 경도 우울증을 정밀하게 진단할 수 있는 지표 개발에 대한 기초 자료로서 사용될 수 있을 것이다.

## 제 1 절 연구대상 및 방법

### 1. 연구 참여자와 연구 설계

연구 참여자는 우울한 증상을 주소로 서울대학교병원 외래에 방문한 환자들 중, 19세에서 65세 사이의 후두부 질환이 없고 독립적으로 질문지를 읽고 이해할 수 있는 환자들을 대상으로 모집하였다. 정상대조군(HC)의 경우 학내 공고문과 온라인 광고를 통해 모집하였다. 간질 등의 기질적 원인에 의해 발생한 우울증상일 경우, 수술 등을 포함한 신경외과적 치료를 받은 경우, 알코올을 제외한 물질 남용의 기왕력이 있는 경우, 지능지수가 70 이하이거나 치매의 진단을 받은 경우 등은 연구에서 제외되었다. 모든 참가자들은 첫 방문 시 헬싱키 선언에 따라 서면 동의서를 작성하였고, 서울대학교병원의 윤리심의위원회의 승인을 받았다.

피험자들의 목소리는 구조화된 인터뷰인 Mini-International Neuropsychiatric Interview(MINI)가 진행되는 동안의 인터뷰를 녹음하여 획득하였다. 이 구조화된 면담의 녹음의 길이는 피험자에 따라 30분에서 50분 가량의 길이였다. 녹음된 파일에서 피험자가 발화하기 시작하는 시점과 발화가 종료된 시점을 기록하여 피험자가 발화하는 구간만을 추출하였다. 이러한 기법을 기반으로, 피험자 별 평균 18.05분의 음성 파일을 얻을 수 있었다.

모집된 피험자들 중에서 주관적인 우울증이나 MINI를 통해 확인된 우울 삽화가 없는 피험자들은 우울하지 않은 그룹(not depressed, ND)에 포함되었다. 주관적으로 우울함이 있다고 호소하지만, 현재의 우울증상이

MINI에 근거하여 주요 우울 삽화의 기준에 충분하지 않은 피험자들은 경도 우울 삽화군(minor depressive episode group, mDE)에 속하는 것으로 분류되었다. 마지막으로 주관적인 우울함을 호소하며, MINI를 통해 현재 주요 우울 장애의 진단 기준에 충족하는 우울 증상이 있음이 확인된 피험자들은 주요 우울 삽화군 (major depressive episode group, MDE)에 속하는 것으로 분류되었다.

2. 사회인구학적 정보와 항정신병약물

사회인구학적 정보로는 성별, 연령, 사회경제적 지위(SES), 정신과 치료력, 정신과 약물 사용, 참가자의 키와 체중을 포함한 정보들이 수집되었다. 정신건강의학과 문제가 아닌 다른 질환이 있고, 그로 인하여 복용하는 약물이 있는 경우 '기타 약물(drugs taken for non-psychiatric disorders)'로 분류되었다. 체질량지수(body mass index, BMI)는 수집된 키와 체중을 통해 계산되었다. 항정신병약물은 추체외로 증상을 통해 목소리에 영향을 미칠 수 있기 때문에(42), 본 연구에서는 복용 중인 항정신병약물의 용량을 확인하였다. 항정신병약물의 효과를 보정하기 위하여, 복용하는 항정신병약물의 복용량을 1일 약물 복용량 기준으로 olanzapine(OZP)을 기준으로 등가용량을 환산한 뒤 합산하였다(65,66).

3. 평가 항목

모든 피험자들은 간이 국제 신경정신 인터뷰(Mini International

Neuropsychiatric Interview, MINI) 7.0.2. 버전을 이용하여 우울증 및 정신과적 질환에 대해 평가되었다. MINI는 DSM-V 진단 기준에 따라 정신과적 질환을 정확하게 진단할 수 있도록 구성된 구조화된 인터뷰이다(67,68). ND 그룹에도 MINI를 통해 주요우울장애 및 양극성장애를 진단받은 이들이 포함되었으나, 이들은 현재 우울하지 않은 상태이므로 ND에 분류되었다. 이 것은 목소리의 다양한 지표들 중에는 성별 등의 특성을 반영하는 특성 지표(trait marker)도 있지만, 현재의 상태에 민감하게 반응하는 상태 지표(state marker)도 포함되어 있기 때문에 현재의 우울 상태에 민감한 상태 지표를 알아보고자 현재의 기분 상태를 중심으로 군을 분류하였기 때문이다.

해밀턴 우울증 등급 척도(Hamilton Depression Rating Scale, HDRS)는 피험자들의 객관적인 우울증 중증도를 평가하는 데 사용되었다. 이 설문지는 우울증상의 심각도와 관련된 17개의 항목으로 구성되어 있으며, 각 항목은 0(존재하지 않음)에서 4(심각함)범위의 5점 리커트 척도를 사용하여 평가된다. 또한 총점이 17점 이상이면 중등도의 우울증을, 24점 이상이면 중증 우울증을 시사한다(69-71).

환자 건강 설문지 (Patient Health Questionnaire-9, PHQ)는 참가자의 주관적 우울증을 평가하기 위해 사용되었다. PHQ는 우울증 선별 척도로 개발되었으며 0(전혀 그렇지 않음)에서 3(거의 매일)까지의 4점 리커트 척도를 사용하여 평가되는 9개 항목으로 구성되며, 10점 이상은 중등도에서 중증의 우울증을 나타낸다(72-74).

불안감에 대해서는 벡 불안 척도(Beck Anxiety Inventory, BAI)에 의해 평가하였다. BAI는 0(전혀 그렇지 않음)에서 3(심각함)까지의 4점 리커트

척도를 사용하여 평가되는 21개 항목으로 구성되며, 10점 이상은 가벼운 불안, 19점 이상은 중간 정도의 불안을 의미한다(75). 2016년에 실시한 메타분석을 통해 16~20점 이상일 경우 병리적 불안을 의미한다(76).

더불어 과거 연구를 통해 충동성이 남성의 우울증 및 불안과 연관되어 있다고 알려졌다(77). 따라서 충동성은 Barratt 충동성 척도(Barratt impulsiveness scale, BIS)를 사용하여 평가되었다. BIS는 1959년 충동적인 성격 특성을 평가하기 위해 개발되었으며 현재 가장 널리 사용되는 척도의 11번째 버전이다. 또한 BIS-11은 1(전혀 그렇지 않다)에서 4(항상)까지 5점 리커트 척도를 사용하여 평가되는 30개의 질문으로 구성되어 있다(78-80).

4. 목소리의 추출과 음성 지표 추출 방법

MINI 면담을 진행하는 중의 녹음 파일에서 피험자의 발화 시작 시점과 피험자의 발화 종료 시점을 기록하여 피험자가 발화하는 구간만을 추출하였다. 그 뒤 발화 길이가 3초 미만의 발화에는 단순 응답일 가능성이 높기 때문에 음성 데이터가 풍부하지 않아 제외하였다. 남은 발화구간은 분석에 활용하기 위하여 10초 단위로 절단하였다. 절단된 구간별로 아래의 방법을 이용하여 목소리의 지표들을 추출하였고, 피험자별 다양한 지표 값들에 대해서는 평균값을 계산하여 대표 값으로 지정하였다.

음성 특징은 성문, 템포 스펙트럼, 포먼트(formant) 및 기타 물리적 특징의 네 가지 측면에서 추출되었다. 모든 특징은 주로 각 발화 내에서 얻은 다음 전체 시간에 대해 평균을 추출하였다. 성문 특징은 소리가 성대에서 어떻게

분절되는지에 대한 정보로 구성되며 파형을 그린 후 각 숫자를 매개변수화하여 얻는다(81,82). 성문 폐쇄 인스턴스(glottal closure instance, GCI)가 먼저 계산된 후 역 필터링을 통해 다양한 미분으로 계산되었다. 그 다음으로 GCI와 미분 형태를 통합하여 성문 파형을 추정하였다. 큰 파형을 평탄화한 뒤, 다음 세 가지 매개변수, 즉 시작 단계(OP), 닫는 단계(CP) 및 닫힌 단계(C)가 추출되었다.

템포 스펙트럼 특징은 음악 정보 검색(MIR)에 주로 사용되는 음향 특징으로 "Librosa"라는 오디오 처리 툴킷을 통해 추출되었다(83). 이는 참가자가 발화를 계속하는 시간 또는 간격의 길이를 나타내는 시간적 특성과 시작의 주기성을 고려한 템포로 구성된다. 또한 평균 스펙트럼 중심, 스펙트럼 대역폭, 롤오프 주파수 및 제곱 평균 제곱근 에너지 등이 스펙트럼 특성으로 사용되었다.

포먼트 특징은 선형 예측 계수를 통해 얻은 음성학에서 일반적으로 사용되는 포먼트에 대한 정보를 나타낸다(84). 포먼트는 성대의 공명을 나타내며 스펙트럼의 극대값으로 이해할 수 있다. 다양한 포먼트 지표들 중에서, 첫번째부터 세번째 포먼트까지 이용되었고, 해당 대역폭을 추출하였다.

다른 물리적 속성의 경우 음높이와 크기의 평균과 분산, ZCR(Zero Crossing Rate), 음성 부분이 활용되었다(85). ZCR은 음성이 얼마나 강렬하게 발화되었는지를 나타내고 음성 부분은 얼마나 자주 나타나는지를 나타낸다. 특정 발화에 대한 ZCR의 평균을 계산한 후 ZCR이 평균보다 낮은 프레임을 무음으로 정의하였다. 해당 지표들을 추출한 코드는 아래와 같다.

```
y, sr = librosa.load(excerpt)
```

```
pitches, magnitudes = librosa.piptrack(y=y, sr=sr)
D = np.abs(librosa.stft(y))**2
ss, phase = librosa.magphase(librosa.stft(y))
S1 = librosa.feature.melspectrogram(S=D)
S2 = librosa.feature.melspectrogram(y=y,sr=sr,n_mels=128,fmax=8000)
```

https://librosa.github.io/librosa/feature.html#spectral-features

```
time  = librosa.get_duration(y=y,sr=sr)
cent  = librosa.feature.spectral_centroid(y=y,sr=sr)
cent  = cent.mean()
band  = librosa.feature.spectral_bandwidth(y=y,sr=sr)
band  = band.mean()
roll  = librosa.feature.spectral_rolloff(y=y,sr=sr)
roll  = roll.mean()
rmse  = librosa.feature.rms(y=y)
rmse  = -np.log(rmse).mean()
onset = librosa.onset.onset_strength(y, sr=sr)
tempo = librosa.beat.tempo(onset, sr=sr)
```

https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html

```
lpc = librosa.core.lpc(y,order=16)
```

```
roots_ = np.roots(lpc)
reals_ = np.real(roots_)
imags_ = np.imag(roots_)
angle = np.arctan(np.asarray(imags)/np.asarray(reals))
freqs = [(sr/(2*np.pi))*z for z in angle]
bands = [-0.5*(sr/(2*np.pi))*np.log(np.abs(z)) for z in roots]

# formant features
import matplotlib.pyplot as plt
import librosa.display
plt.figure(figsize=(10, 4))
librosa.display.specshow(librosa.power_to_db(S1,ref=np.max),
y_axis='mel', fmax=8000, x_axis='time')
plt.colorbar(format='%+2.0f dB')
plt.title('Mel spectrogram')
plt.tight_layout()

# Glottal features
sys.path.append('DisVoice/phonation')
from glottal_feature import GF_extract
def get_glottal(excerpt):
cycle, glottal_flow, glottal_flow_d=GF_extract
res = [z.mean() for z in np.transpose(cycle)]
```

```
# Pitch, Magnitude, Voice portion features
def get_pitmag(y,sr,zero_cross):
pitches, magnitudes = librosa.piptrack(y=y, sr=sr)
pitches = np.transpose(pitches)
magnitudes = np.transpose(magnitudes)
pit_seq = pip_to_pitch(pitches)
mag_seq = pip_to_pitch(magnitudes)
delay_p, mag_index, real_z  = find_delay(zero_cross)
mean_z = np.average(real_z)
mean_p, error_p, change_p = effective_analysis(pit_seq, mag_index)
mean_m, error_m, change_m = effective_analysis(mag_seq, mag_index)
return [mean_p, error_p, change_p, mean_m, error_m, change_m, mean_z, delay_p]
```

5. 통계학적 방법

인구통계학적 특징과 임상적 특징 중 범주형 변수를 피어슨 카이 제곱 검정(Pearson's chi-squared test)을 이용하여 비교 분석하였으며, 사후 검정은 피셔 정확 검정(Fisher’s exact test)을 이용하여 수행하였다. 연속형 변수의 경우 데이터가 정규분포를 만족하지 않았기 때문에 비모수적 방법인 크루스칼 왈리스 검정 (Kruskal-Wallis test)을 사용하여 세 그룹을

비교했으며, 사후 분석으로 맨- 휘트니 U 검정(Mann-Whitney U test)을 사용하였다. 또한 본 연구에서는 목소리의 다양한 특징들을 추출하여 다중 비교를 하였기 때문에, 제 1종 오류를 방지하기 위하여 벤자미니-호치버그 검정(Benjamini-Hochberg test)을 이용하여 사후 검정을 한 번 더 수행하였다.

피험자를 3군으로 분류한 결과, 한 군의 피험자가 30명이 넘지 않는 군이 있었기 때문에 정규성을 가정하기 충분하지 않았다. 하지만 음성의 특징을 비교할 때 사회인구학적 정보들과 임상의 특성의 다양한 변수들을 공변량으로 포함하여 분석할 필요가 있다. 따라서 본 연구에서는 목소리의 지표들 중에서 왜도의 절대값이 2 이하이거나, 첨도의 절대값이 7 이하일 경우 정규성을 보인다고 가정하였다. 만약 왜도가 2를 초과하는 경우 로그 변환을 하고, 왜도가 -2보다 작은 경우에는 제곱처리를 하여 정규성을 보정하였다(82). 한편 임상 변수들은 정규성에 만족하지 않더라도 변환을 하지 않았는데, 임상 척도의 점수들을 변환하게 되면 중등도 등을 평가하는 기준점과 공변량으로서 통계적 영향 모두 변화하기 때문이다.

이후 분산분석(analysis of variance, ANOVA)를 통해 세 군을 비교하였고, 세 군 간에 차이가 있는 연령, BMI 및 기타 약물 사용에 대해 공분산분석(analysis of covariance, ANCOVA)를 이용하여 p값을 보정하였다. Windows용 IBM SPSS Statistics 버전 25.0(SPSS Inc., Chicago, IL, USA)을 사용하여 분석을 수행하였다.

기계학습의 방법에는 로지스틱회귀분석(logistic regression, LR), 가우시안 나이브 베이즈 분석(Gaussian Naïve Bayes, GNB), 서포트 벡터 머신(support vector machine, SVM), 다층 퍼셉트론(multilayer perceptron,

MLP) 등의 다양한 방법이 존재한다(87,88). 음성의 지표에는 어떠한 분석 기법이 가장 우수한 성능을 보이는지에 대한 연구도 아직 부족하기 때문에, 본 연구에서는 위 4가지 기법을 모두 적용한 뒤 정확도를 확인하였다. 전체 데이터의 70%와 80%는 훈련 데이터로 사용되었고, 나머지 30%와 20%는 두 시나리오에 대한 예측 데이터로 사용되었다. 또한 본 연구에서는 데이터의 품질과 규모를 보강하기 위하여 소량의 노이즈를 추가하는 모델도 구축하였고, 이 모델의 경우 "augmented model"로 명명하였다. 또한 모든 음성지표를 사용하는 경우와 달리, 세 군에도 통계적 유의성을 보이는 지표만을 추출하여 구축한 모델의 경우 "selected model"로 명명하였다. 기계학습 모델의 정확도 평가를 위해서 정확도(accuracy), 신뢰 구간 95%의 곡선 아래 면적(area under the curve, AUC), 정밀도(precision), 재현율(recall)과 F1 점수를 추출하였다. LR, GNB 및 SVM은 Scikit-learn이라는 Python 패키지를 통해 구현되었으며 MLP는 Keras를 통해 구현되었다(89,90).

## 제 2 절 결과

### 1. 우울 삽화에 따른 사회인구학적 특성과 임상적 특성의 비교

2019년 1월 10일부터 2020년 4월 30일까지 총 93명의 피험자를 모집하였다. MINI 결과에 따라 우울증을 나타내는 60명의 피험자는 주요 우울 삽화(MDE)에 해당하는 34명의 피험자와 경미한 우울 삽화(mDE)에 해당하는 26명의 피험자 그룹으로 추가 분류되었다.

여성은 참가자의 70-79%를 구성했으며 연령과 SES는 통계적 차이를 보이지 않았다. mDE 그룹은 나이가 가장 많았으며 정신건강의학과 질환이 아닌 다른 질환으로 복용하는 약물이 더 많았다. MDE 그룹은 BMI가 가장 높은 군이었다. 또한 MDE 그룹은 항 정신병 약물을 복용하는 피험자가 가장 많았지만 복용량은 통계적으로 유의하지 않았다. 과거에 존재하였던 우울삽화 및 불안 증상, 조증 증상 등으로 인하여 MINI에서 진단을 받았더라도, 현재 정신과적 증상이 관해된 경우에는 ND 그룹으로 분류되었기 때문에, ND 그룹에서도 MINI를 통해 주요 우울장애를 진단받은 5명, 양극성 장애를 진단받은 2명이 포함되었다(표 1-1).

### 2. 우울 삽화에 따른 임상적 특징의 비교

객관적이고 주관적인 우울과 불안의 정도는 우울 삽화의 정도에 따라 증가하는 경향을 보였다. 그러나 충동성은 세 그룹 간에 통계적으로 유의한 차이가 없었다. 불안에 대한 MDE 그룹과 mDE 그룹 간의 차이는

통계적으로 유의하지 않았다(표 1-2).

3. 우울 삽화에 따른 목소리 지표들의 비교

추출된 21개의 음성 지표를 삽화군 별로 분석한 결과, 그룹별로 차이를 보이는 8개의 지표가 나타났다. 정상군과 mDE군 사이에 차이를 보인 음성 특징은 spectral centroid, spectral roll-off, sq mean pitch, standard deviation pitch, mean magnitude, ZCR과 voice portion 이었다. mDE 그룹과 MDE 그룹 간에는 통계적으로 유의미한 다른 목소리의 지표는 standard deviation pitch 하나가 있었다. 연령, 체질량지수(BMI), 약품 사용량 등을 보정한 결과 총 7개의 음성 특징이 통계적으로 유의한 것으로 나타났다: spectral centroid ($p$ = 0.008), spectral roll-off ($p$ = 0.012), formant BW2 ($p$ = 0.040), sq mean pitch ($p$ = 0.027), standard deviation pitch ($p$ = 0.020), ZCR ($p$ < 0.001), voice portion ($p$ = 0.020) (표 1-3).

Jonckheere-Terpstra 테스트는 각 그룹의 순위에 대한 통계적 검증을 시행하였다. 음성 특징은 에피소드의 심각도가 증가함에 따라 변화하는 경향을 보이지 않았다. 오히려 7개의 변수들은 ND에서 MDE, 그리고 mDE 순서로 증가 혹은 감소하는 추세를 보였다(표 1-4와 그림 1-1).

4. 기계학습을 통해 우울 삽화 예측 모델

기본적으로 21개의 음성 특성을 모두 사용하여 모델을 구성하였다. 하지만 “selected” 모델에는 그룹별로 차이를 보이는 7가지 목소리의 지표만을

이용하여 분석하였다(표 1-3). 결과는 시험에 대한 최상의 테스트 세트 정확도의 평균(최대), 즉 7:3 및 8:2 분할에 대해 각각 3회 및 5회 군을 다르게 분류하여 평가한 뒤, 측정치의 평균값을 보고하였다. MLP 분석 사양으로 128, 64 크기의 은닉층(hidden layer)을 사용하였다. 일반적으로 MLP 모델이 가장 좋은 성능을 보였다. 또한 7:3으로 학습을 시킨 모델보다, 8:3로 학습을 시킨 모델에서 더 높은 정확도를 보였다. 또한 각 군별로 통계적으로 유의성을 보이지 않은 목소리의 지표들을 모두 적용한 모델이 'selected' 모델보다 더 정확한 성능을 보였으며, 잡음(노이즈)를 함께 학습시킨 'augmented' 모델이 더 정확한 성능을 보였으며, precision의 평균값은 0.66, recall 의 평균값은 0.66를 보였다. MLP를 통해 AUC를 계산한 후 7:3 훈련 세트에 대한 결과는 AUC가 mDE를 구별함에 있어서 0.79, MDE를 구별함에 있어서 0.58로 측정되었다. 8:2 훈련 세트에서의 AUC값은 mDE를 구별함에 있어 0.69, MDE를 구별함에 있어서 0.67로 측정되었다(그림 1-2).

또한 LR과 GNB는 0.59-0.65의 AUC를 보였으며, 이 것은 MLP의 성능보다 우수하지는 않았다. 또한 LR 및 GNB 모델에서는 훈련 세트가 증가함에 따라 성능이 향상되는 경향이 관찰되지 않았다(표 1-5).

## 5. 진단명 별 임상지표와 목소리의 차이

추가적으로 본 연구에서는 진단명 별로 사회인구학적 정보와 임상정보, 그리고 목소리의 지표들도 분석을 진행하였다. MINI를 기준으로 과거 및 현재에도 정신과적 진단이 없는 정상대조군(HC) 26명과, 과거 혹은 현재

주요 우울 장애(MDD)로 진단이 내려진 주요 우울 장애군 15명, 양극성 장애(BP)로 진단이 내려진 52명으로 구분되어졌다.

이들의 사회인구학적 정보는 성별, 나이, SES, 정신건강의학과 외 문제로 복용중인 약물 유무와 BMI에서는 통계적 차이를 보이지 않았다. MDD 군에서 항정신병약물을 복용하는 피험자는 8명, BP 군에서 항정신병약물을 복용하는 피험자는 44명이었으며, 그 두 군간 맨-휘트니 U 검정을 이용하여 분석하였을 때 복용 용량은 통계적 유의성을 보이지 않았다. 하지만 과거 정신과적 치료력에서는 통계적으로 유의하게 BP군에서 더 많은 치료를 받았던 것으로 밝혀졌다(표 1-6).

각 군 별 임상지표들을 분석한 결과, HC, MDD, BP 순으로 객관적 우울 지표인 HDRS 점수, 주관적 우울 지표인 PHQ 점수, 불안 점수인 BAI 모두 순차적으로 통계적으로 유의한 차이를 보이며 커지는 것을 확인할 수 있었다. 하지만 BIS는 우울삽화 별 비교 분석과 유사하게 HC, MDD, BP 간의 통계적 차이를 보이지 않았다(표1-7).

또한 추출된 목소리 지표 21개를 비모수적 방법인 크루스칼-왈리스 검정을 통해 비교분석한 결과, spectral centroid, spectral roll-off, mean pitch, Stdev pitch, mean magnitude, ZCR, voice portion 7개의 지표에서 통계적 차이를 확인할 수 있었다(표1-8).

## 제 3 절 고찰

이 연구에서는 구조화된 면담을 통해서 피험자의 우울 삽화를 평가하였고, 목소리를 취득하였다. 또한 면담의 녹음을 통해 획득한 21개의 목소리의 지표들이 우울증의 심각한 정도에 따라서 변화하는지를 확인하였다. 21개의 지표들 중에는 성대의 특징을 반영하는 평균 음 높이에서부터, 발화지연 정도가 반영되는 발화시간 중 실제 발화시간의 비율 등 다양한 지표들이 포함되었다. 스펙트럼 중심(Spectral centroid)는 음성 스펙트럼의 중심을 나타내며, 이 것은 화자의 목소리의 높낮이를 반영한다(91). 스펙트럼 롤오프(spectral roll-off)는 전체 스펙트럼 에너지의 지정된 백분율 아래의 주파수로서, 임상적으로는 고음의 발화가 많아질수록 spectral roll-off가 커지게 된다(92). 포먼트는 발화된 스펙트럼에서 넓은 피크 혹은 일부 영역 중의 최대값으로 정의되며, 포먼트 BW2 (formant BW2) 값은 두 번째 피크값으로 발화되는 억양의 특징 중 하나이다(93). 본 연구에서는 제곱함수를 처리하였으나 상관관계를 고려할 때 평균음높이는 음성의 평균 주파수를 말하며 평균음높이가 높을수록 높은 음성을 의미한다. 음높이의 표준 편차(standard deviation pitch)는 평균 발화된 음높이를 기반으로 계산된 지표로서, 발화 음의 높이가 높고 낮음의 변화가 클수록 그 변화량이 커지게 된다. ZCR은 파형이 수평선을 가로지르는 비율로, 특정 프레임에 음성이 존재하는지 여부를 나타내는 지표로 종종 이용된다(94). 음성 존재 비율 (voice portion)은 ZCR을 기준으로 전체 프레임에 비하여 음성이 존재하는 프레임의 비율을 의미한다.

본 연구의 결과에 따르면 ND, MDE, mDE 그룹의 순서로 음성이

낮아지고, 말하는 동안 음높이 변화가 더 많이 발생하는 것으로 나타났다. Jonckheere-Terpstra test를 통해 group 3의 순서를 확인하였을 때에도 formant BW2와 sq_mean pitch를 제외한 음성의 변화는 ND, MDE, mDE 순임을 통계적으로 확인하였다(표 1-4).

이전 연구에 따르면 우울증 환자의 경우 목소리 톤이 단순해지고 생기가 없으며 볼륨이 낮아진다(95,96). 또한 47명의 우울한 환자와 57명의 우울하지 않은 참가자를 비교한 연구에서는 성대의 움직임이 느리고 참가자가 낮은 목소리로 말하는 것으로 나타났다고 보고하였다(97). 36명의 우울한 환자를 우울하지 않은 그룹과 비교한 추가 연구에서도 우울한 환자의 목소리가 낮다는 것이 확인되었다(98). 이러한 결과는 본 연구의 결과와 일치하며, 우울증군이 정상군에 비해 낮은 목소리를 내고 있음을 보여주었다.

우울증 환자의 중증도와 음높이 변동성에 대한 이전 연구는 다양한 결과를 보였다(99,100). 7명의 환자를 대상으로 한 2004년 연구에서 음높이 변동성은 우울증 중증도와 관련하여 감소한다는 결과를 보였으나, 35명의 환자의 목소리를 분석한 2007년 연구에서는 우울증이 증가할수록 음높이 변동성이 증가하는 것으로 나타났다(101,102). 본 연구에서는 음높이의 변화는 우울증군에서 정상군보다 더 큰 것으로 확인되었다. 또한 우울증상과 더불어 불안증상도 함께 증가하는 것도 확인할 수 있었다. 불안이 목소리에 미치는 영향에 대한 더 많은 연구가 필요하지만, 목소리의 떨림으로 표현될 수 있으며 음높이 변동성을 증가시킬 수 있다.

우울증의 중증도와 목소리 특성 사이의 관계에 대한 이전 연구는 우울증 진단을 받은 대상을 대상으로 수행되었다. 하지만 이 연구에서는 MDE보다 mDE군에서 더욱 음성의 변화가 크고 뚜렷하게 나타났다. 일반적으로

주요우울증보다 유병률이 높은 경도 우울증은 주요 우울 장애의 전조 증상으로 간주되며 주요우울증으로 진행될 가능성이 높다고 알려져 있다(103). 하지만 경도 우울증의 일부는 주요 우울 장애에서 증상이 일부만 개선이 된 군, 즉 부분 관해 된 상태일 가능성도 있다(104). 이 연구에서 mDE군은 MDE보다 나이가 많았고, 88.5%의 피험자가 평가 당시 항정신병 약물을 복용하고 있었다. 따라서 본 연구의 mDE는 부분 관해된 주요우울삽화일 가능성이 높다. 이러한 점에서 목소리 지표 변화를 해석해볼 때, 우울증의 증상이 호전되더라도 음성의 변화가 개선되지 않을 가능성이 있음을 시사한다.

기계 학습을 사용하여 우울 삽화의 심각도를 예측한 결과는 93개의 사례에서 8:2 훈련-테스트 분할로 60.0%의 정확도를 달성하였다. 이 60%의 정확도는 3군 비교에서 합당한 수준이며, 피험자의 수를 더 늘리면 정확도를 높일 수 있을 것으로 기대된다. 본 연구에서 60%의 정확도가 허용되는 이유는 본 연구와 유사하게 세 군을 구분하는 기존 연구에서의 F1 점수가 0.303에서 0.633 사이였고, 자연어 분석 등을 통해 세 군을 비교하는 기존 연구에서 F1 점수가 0.51에서 0.71 사이였기 때문이다(105,106). 본 분석에서는 ND, mDE, MDE, 세 군을 예측하기 위해 케이스별 추론을 수행하였기 때문에 정확도가 떨어질 수 밖에 없었다.

추가적으로 진단명에 따른 음성의 지표도 분석을 진행하였고, MDD 군이 BP군보다 우울증상, 불안증상이 덜하다는 것을 확인할 수 있었다. 음성 지표들의 분석 결과, 우울 삽화에 따라 차이를 보였던 7개의 목소리 지표들이 각 진단별로도 통계적 유의한 차이를 보이는 것 역시 확인할 수 있었다(표1-8). MDD 군이 우울증상이 더 경미하다는 것을 고려해볼 때,

목소리 지표들에 대한 분석결과는 우울증상을 보정한 뒤에도 진단별로 차이를 보이는 목소리 지표를 확인하는 것이 필요하다. 하지만 MDD군이 15명에 불과하여 비모수적 방법을 활용할 수 밖에 없었기 때문에 우울 증상 및 불안 증상을 보정한 목소리 지표의 차이를 확인할 수 없었다. 따라서 추후 진단에 따른 목소리 지표는 추가적인 데이터가 확보된 뒤 정밀한 분석이 필요할 것이다.

이 연구는 여러 강점을 가지고 있다. 기존 연구와 달리 구조화된 면담을 통해 충분한 시간(평균 18분) 동안 피험자의 목소리를 녹음하였다. 따라서 오디오 파일은 단순한 문장의 반복을 포함하지 않고 대신 다양한 구어체 및 준언어적 표현을 반영되었다. 또한, 기존 연구들과 달리 인구학적 특성과 임상적 변수들을 교정한 후에도 목소리의 지표와 우울 정도의 통계적 차이가 확인되었다. 연령, 다른 약물의 사용, 체질량지수(BMI)와 같은 다양한 요인 외에 항정신병 약물이 성대와 성대에 미치는 영향까지 고려한 분석을 수행하였다. 중요한 것은 경도 우울증군에서의 목소리 변화가 확인되었는데, 이는 주요 우울증군에서 변화보다 더 큰 것으로 나타났다. 따라서 이 연구를 통해 목소리를 바이오 마커로 사용하여 기계 학습을 통해 경도 우울증을 감지하고 진단할 수 있는 가능성을 확인하였다. 또한 주요 우울증과 경도 우울증을 진단할 때 음성을 객관적 지표로 사용할 가능성을 제시한다. 또한, 본 연구에서는 다양한 음성의 21가지 특징을 추출하였고, 목소리의 다양한 지표들 중에서 성별과 같이 주체의 특성을 반영하는 지표(state marker)가 있을 것이고, 주체의 상태를 반영하는 지표(trait marker)가 있을 것이라고 가정하였고, 각 군 별로 통계적 차이를 보였던 7개의 지표는 trait marker로서의 가능성을 시사한다.

물론 이 연구에는 다양한 한계가 있다. 첫 번째이자 가장 중요한 한계로 본 연구의 표본 크기는 충분히 크지 않다는 점이다. 본 연구에서는 인터뷰 전체를 녹음한 뒤, 피험자의 목소리 부분을 따로 추출하는 과정을 걸쳤다. 이 과정은 전체 인터뷰를 청취하면서 피험자의 발화의 시작과 끝을 기록하는 방법으로 이루어졌으며, 이 과정은 인터뷰 시간의 약 3배가 소요되며, 인력에 의하여 진행되는 과정이었다. 따라서 인터뷰 녹음을 이용한다는 과정 자체가 큰 표본을 대상으로 하기에는 장애가 되는 과정이기도 하였다. 따라서 앞으로는 확보된 인터뷰 녹음 파일을 기반으로 진단에 필요한 내용을 자동으로 판별해주는 프로세스와 더불어 발화자 간을 자동으로 선별해주는 오디오 신호 처리 방법의 개발이 필요하다. 또한, 본 연구는 각 피험자의 평균 발화 시간이 충분히 길다는 장점이 있으나, 표본의 크기가 충분하지 않아 보다 큰 규모의 연구를 통해 본 연구의 결과를 재현할 수 있는지 여부를 확인할 필요가 있다. 둘째로 ANCOVA를 이용하여 임상인구통계를 보정하고 비교하였지만, 음성지표는 각 군에서 불안의 정도에 따라 보정되지 않았다. 또한 불안의 정도가 우울함에 따른 목소리의 변화를 매개했을 가능성도 있다. 이와 관련하여 우울 환자의 불안 정도가 음성 지표를 어떻게 변화시키는지에 대한 매개 분석을 포함한 추가 연구가 필요하다. 셋째로 본 연구에서는 우울 삽화의 심각도와 목소리 특징 사이 관계를 단면적으로 확인하였기 때문에, 경도 우울증에서의 목소리 특징이 복용 중인 약물에 의한 영향일 가능성도 배제할 수 없으며, 선후관계를 확인할 수가 없다. 따라서 향후에는 더욱 다수의 피험자를 대상으로 한, 정기적 추적관찰을 통해 개인의 목소리 변화에 대한 연구가 추가적으로 필요하겠다.

# 제 3 장 목소리와 텍스트 데이터를 이용하여 우울증 진단 인공지능 알고리즘의 개발 (연구 II)

연구 II에서는 연구 I의 데이터에서 확장된, 3차 년도까지 최종적으로 모집된 초기평가 데이터를 기반으로 하였다. 연구 II의 목적은 정상군과 우울증군 간의 차이가 있는 목소리의 지표들을 분석하고, 이를 이용하여 정상군과 우울증 군을 구별할 수 있는 인공지능 알고리즘을 재정립하는 것이다. 더 나아가 MINI에서 환자가 사용한 텍스트 데이터를 추출하여 나이브 베이즈(Naïve Bayes) 기법을 적용하여 기계학습을 통해 텍스트를 기반으로 한 우울증 진단 알고리즘 역시 구축하여 그 정확도를 평가해보고자 한다.

연구 I을 통해 주요 우울 삽화 환자일수록 우울 증상 뿐 아니라 불안 증상에 대한 점수도 함께 높아지는 것을 확인할 수 있었다. 따라서 연구 II에서는 이러한 우울증상과 불안 증상과 함께 자살 사고의 심각도를 함께 측정하여 우울과 불안, 자살사고 그리고 목소리 지표와의 복합적인 관계에 매개 분석을 이용하여 알아보고자 한다.

또한 연구 I에서 사용되었던 알고리즘의 성능을 연구 II에서 추가적으로 모집된 피험자를 대상으로 확인하고 발전시켜보고자 하였다. 특히 연구 I의 알고리즘에 포함된 21개의 음성 지표들 중 glottal feature의 경우 추출까지의 시간이 24시간 이상이 소요되었다. 하지만 본 연구에서는 실시간 인터뷰를 통한 우울증 및 자살 위험도를 진단하는 알고리즘을 구축하는 것이 목적이기 때문에, 추출까지의 시간이 과도하게 걸리는 지표를 알고리즘에

포함할 경우 임상적 활용 가치가 떨어지게 된다. 따라서 연구 II에서는 연구 I에서 구축된 알고리즘과 별개로, glottal feature들을 제외한 음성의 지표를 기반으로 진단 알고리즘을 새로이 구축하고자 하였다.

더욱이 연구 I에서는 자살 고위험군의 사례가 적어, 자살 위험성에 대한 진단 알고리즘 구축에 어려움이 있었다. 이러한 한계를 극복하고, 자살 고위험군의 발병률이 작다는 한계를 극복하기 위하여, 연구 II에서는 정상군과 우울증군을 구별함에 초점을 맞추고, 연구 III에서는 우울증 군 내에서의 자살고위험군을 구분하는 인공지능 알고리즘을 구축하여 자살고위험군 진단의 정확도를 높이고자 계획하였다. 연구 II의 진단 알고리즘은 정상군과 우울증 군을 구별할 수 있는 인공지능 알고리즘 구축을 통해 우울증 진단의 정확성을 높이고, 객관적 지표로서 목소리 뿐 아니라 텍스트가 가지고 있는 가능성을 밝힘으로써 다양한 분야의 연구로 발전할 수 있는 초석을 닦는 기회가 될 것이다.

## 제 1 절 연구대상 및 방법

### 1. 연구 대상자

환자군은 우울증상을 호소하며 서울대학교병원 정신건강의학과 기분장애 클리닉을 방문한 환자들 중에서, 19세에서 65세의 성인 환자를 대상으로 모집하였으며, 정상군의 모집은 온라인 홍보와 학내 게시판을 통하여 모집하였다. 피험자들은 2019년 1월 10일부터 2021년 8월 31일까지 모집되었으며, MINI를 통해 정신과 진단을 확인했고, 정상군에서는 MINI를 통해 과거 또는 현재의 정신질환이 발견된 경우 연구에서 제외하였다. 또한 환자군 중 자기보고식 설문지를 작성하지 않은 피험자도 본 연구에서 제외하였다.

### 2. 평가 항목

자살 사고에 대해서 Beck's suicidal ideation scale (BSS)에 의하여 평가되었다(107). BSS는 자살시도에 대한 19개 문항으로 구성되어 있으며 아직 정해진 절단 기준점은 없으나 BSS의 점수가 높을수록 자살의 위험이 높다(108,109).

### 3. 목소리 지표의 추출

목소리 지표의 방법은 연구 I과 유사하다. 하지만 연구 I의 분석에 사용되었던 glottal 지표들은 추출에 시간과 컴퓨터 사양에 로딩이 많이 걸리기 때문에, 본 연구를 기반하여 인터뷰 음성을 실시간으로 지표화하는 앱을 개발하기에는 활용성이 떨어지게 된다. 따라서 본 연구에서는 연구 I에서 활용되었던 glottal 지표 3가지는 연구 II에서는 제외하였다.

연구 II에서는 피험자들의 한 문장의 길이를 'time' 이라는 변수로 추출하여 새로운 변수로 선정하였고, 스펙트럼 특징들과 포먼트 지표들은 연구 I과 동일한 방법으로 추출되었다. 음높이(pitch)와 강약(magnitude)에 대한 변수들로 평균값(mean)과 평균을 기준으로 각 프레임에서 가장 먼 정도의 절대값(error), 그리고 각 발화구간 별 평균값의 변화량(change)을 추출하였다.

ZCR(Zero Crossing Rate)은 연구 I과 동일하게 추출되었으며, ZCR을 이용하여 음성이 있는 구간을 연구 I에서 추출하였다면, 연구 II에서는 공백이 있는 구간을 'pause portion' 으로 추출하였다. 또한 'total count' 는 발화된 문장의 총 개수를 추출하여 새로운 변수로 선정하였다.

4. 텍스트 데이터의 추출

연구 I과 동일한 방법으로 분할된 음성 파일은 파이썬 라이브러리인 음성 인식 툴킷을 사용하여 텍스트로 변환되었다. 툴킷의 음성 인식기는 구글 API를 사용하였으며, 한국어 음성을 인식할 수 있는 "ko-KR"로 설정하였다. 모든 발화의 텍스트를 독립적으로 기록하였고, 이 과정에서 내용이 인식되지 못한 발화는 분석에서 제외하였다.

5. 통계학적 방법

범주형 변수는 피어슨 카이 제곱 검정(Pearson's chi-squared test)를 통해 비교하였고 연속형 변수는 스튜던트 T 검정(student T test)를 통해 분석하였다. 많은 수의 목소리의 지표들을 동시에 분석하였기 때문에 제 1종 오류를 감소시키고, 사회인구학적 정보와 임상 변수들에 대한 영향을 보정하기 위하여 다변량 공분산분석(Multivariate analysis of covariance, MANCOVA)방법을 적용하였다.

여러 목소리의 지표들이 추출된 뒤, 연구 I에서와 같이 왜도의 절대값이 2 이하일 경우와 첨도의 절대값 7 이하일 경우 모두를 만족할 때 정규성을 띈다고 가정하였다(110). 따라서 목소리의 지표들 중 정규성의 조건을 만족하지 않을 경우, 왜도가 양의 값이면 로그 지환을 하였다. 로그 치환 후 왜도와 첨도의 기준이 정규성의 기준이 부합한 것을 확인하였다(표 2-1). 모든 통계는 Windows용 IBM SPSS Statistics 버전 25.0(SPSS Inc., Chicago, IL, USA)을 사용하여 분석을 수행하였다.

또한 매개 회귀 분석(mediation analysis)을 시행하기 위하여 각 목소리의 지표들과 임상 변수간의 피어슨 상관관계(pearson’s correlation analysis)를 통해 분석하였다. 우울증상과 불안, 그리고 자살사고는 밀접한 매개 관계에 있음이 여러 선행 연구들에 의해 밝혀져 있다(111, 112). 따라서 본 연구에서는 우울증상이 목소리 지표에 미치는 영향에도 안과 자살사고가 다중으로 매개한다는 가정하에 연속 다중 매개 매개 분석 (serial multiple mediator model)을 시행하였다. 매개 분석은 SPSS PROCESS

macro. 3.5 버전의 모델 6번을 이용하여 정상군과 우울증 군에서 각각 분석을 시행하였다.

6. 기계학습의 방법

우선 연구 I에서 구축된 음성지표의 알고리즘을 연구 II에서 추가적으로 모집된 피험자를 대상으로 적용하였다. 그리고 연구 II의 알고리즘을 새로이 구축하기 위하여 음성 지표들을 기반으로 우울증 환자를 구별하는 알고리즘 구축을 위해 연구 I에서 진행하였던 4가지 기계학습 기법 외, quadratic discriminant analysis(QDA)기법과 Adaptive boosting(Adaboost)기법을 추가로 적용하였다. QDA 방법은 LR와 유사하게 측정된 지표들이 정규성을 띈다는 가정하에 각 군을 구별할 수 있는 기준선을 찾아가는 방법이지만, QDA는 LR에 비하여 비교적 작은 수의 학습에서 더 안정적인 모델이다(113). 연구 II와 연구 III에서는 정상군과 우울증군을 구분하고, 순차적으로 우울증군 내에서 자살 저위험군과 자살 고위험군을 구분하는 단계별 접근을 계획하였고, 이러한 단계별 접근의 방법으로서 QDA는 LR보다 더 적합한 모델일 수 있다(114). Adaboost 기법은 단순 분류를 시행한 뒤 분류의 결과가 좋은 샘플에게 가중치를 주고, 분류 성능이 좋지 못한 문제에 더 집중하여 새로운 분류의 규칙을 세워 나가면서 모델의 성능을 개선시키는 방법이다(115).

또한 연구 II에서는 전체 데이터의 80%를 학습에 활용(training)하였고, 남은 20%의 데이터를 정확도 확인(validation)에 이용하였으며, 모든 기계학습의 성능에 대한 지표들은 5중 교차검증을 통한 값들의 평균 수치로

산정되었다.

텍스트 데이터를 기반으로 우울증 환자를 구별하는 방법에서는 텍스트 하나하나가 각 군에 속할 확률을 계산하는 조건부 확률 기반 기계 학습 알고리즘인 나이브 베이즈 (Naïve Bayes) 방법을 활용하였다. 텍스트 영역에서 나이브 베이즈 방법은 전체 문장에 나타나는 단어의 빈도를 계산한 다음 이를 기반으로 통계 모델을 훈련한다. 따라서 연구 II에서는 정상군와 우울증 환자군에 따른 단어 빈도의 분포를 학습한 뒤, 피험자가 말한 문장의 단어를 기반으로 그 피험자가 어느 그룹에 속하는지 확률을 예측하였다.

특히 앙상블 모델의 경우, 각 군의 피험자 수가 30명을 넘어갈 경우 가우스 정규 분포를 만족한다는 가정(116-119)하에 인구통계학적 분포로부터 가우스 밀도 분포를 생성한 뒤, 주어진 인구통계학적 지표가 각 그룹에 속할 확률을 계산하였다. 이후 텍스트 기반의 Naive Bayes 분류기에 이 확률을 구현하여 앙상블 모델을 구축하였다. Naive Bayes 분류기는 Natural Language Toolkit 패키지(120)에서 사용되었으며 scikit-learn 패키지는 밀도 추정 및 분석에 사용되었다. 모든 실험은 5중 교차 검증을 통해 수행되었다. 정상군와 우울증군을 비교하는 앙상블 모델에서는 두 군에서 통계적 차이를 보였던 사회인구학적 지표들과 함께 성별까지 포함하여 모델을 구성하였다. 마지막으로 각 기계학습 모델의 ROC 값은 DeLong test에 의하여 비교하였다(121).

## 제 2 절 결과

1. 연구 II에서 추가적으로 모집된 피험자에 대한 연구 I 알고리즘의 결과

연구 II에서 2019년 1월 10일부터 2021년 8월 31일까지 총 85명의 우울증 환자군(current depression group, CD)이 모집되었고, 그 중 자가보고 설문지를 완료하지 않은 2명이 연구에서 배제되어 분석에는 총 83명의 데이터가 이용되었다. 105명의 정상군(healthy control group. HC)이 모집되었으며, 그 중 MINI를 통하여 과거 혹은 현재 정신과적 병력이 밝혀진 22명은 분석에서 제외되어, 정상군에는 총 83명이 분류되었다.

이들 중 연구 I의 알고리즘 구축 분석에 포함되지 않았던 83명을 대상으로 연구 I의 알고리즘을 적용한 결과를 확인하였다. 83명에는 정상군이 58명, 경도 우울증군이 6명, 주요 우울증군이 19명이었다. 그 결과 정확도는 SVM 모델에서 0.706으로 가장 높았으며, 민감도까지 포함된 F1 점수는 GNB 모델에서 0.406으로 가장 높았다(표 2-2).

2. 정상군과 우울증군 간의 비교

사회인구학적 정보를 살펴보면, 정상군에서 평균 연령은 37.072세로 우울증군의 평균연령 30.916세보다 높았다. 성별에서는 통계적 차이를 보이지 않았으며, SES는 정상군이 우울증군보다 더 낮은 경향을 보였다. BMI의 경우 우울증군의 평균이 24.809로, 정상군의 평균 22.364보다 높았으며, 정신건강의학과 외 다른 질환으로 복용하고 있는 약물이 있는 피험자는 우울증군에서 더 많음을 확인할 수 있었다(표 2-3).

우울증상에 대한 척도, 불안 증상에 대한 척도와 자살사고에 대한 척도

점수는 모두 우울증군에서 통계적으로 유의하게 높았으나, BIS 점수, 즉 충동성에 대한 점수는 양 군에서의 차이를 확인할 수 없었다(표 2-4).

3. 목소리 지표를 통한 기계학습 알고리즘 결과

스튜던트 T 검정을 통해 정상군과 우울증군을 비교하였을 때에는 spectral_band, ZCR, pause portion, total count_log 총 4가지 목소리 지표에서 통계적으로 유의한 차이를 확인할 수 있었다. 하지만 사회인구학적 정보 중에서 양 군의 차이가 있었던 나이, SES, BMI와 다른 약제 복용 유무를 MANCOVA를 통해 보정한 뒤에는 time_log, spectral_band, mean_pitch, change_pitch_log, ZCR, pause portion, total count_log 총 7개의 지표에서 통계적으로 유의한 차이를 확인할 수 있었다. 하지만 불안 척도 점수인 BAI의 결과를 위 MANCOVA 모델에 공변으로서 함께 보정을 한 뒤에는 spectral_band, ZCR, pause portion, total count_log 만이 통계적 유의성이 유지되었다(표 2-5).

목소리의 지표들을 기계학습의 다양한 모델에 학습을 시켰고, 기존의 연구 I에서 활용하였던 LR, GNB, SVM, MLP 등의 결과는 F1 수치가 0.514에서 0.777 가량의 성능을 보였으며, 그 중에서는 GNB를 통한 기계학습 모델이 가장 우수한 성능을 보였다. 이번에 새롭게 적용한 QDA 모델은 GNB 모델보다 F1 수치와, 정확도, 민감도의 특이도 모두 더 높은 성능을 보임을 확인할 수 있었다. 하지만 QDA 모델의 경우 두 군에서 통계적으로 유의함을 보였던 4개의 지표만을 추출하여 학습을 시켰을 경우 그 정확도가 감소하는 것을 확인할 수 있었다(표 2-6).

가장 우수한 성능을 보였던 QDA 모델의 ROC 곡선을 확인해보았을 때, 5중 교차 검증 후 평균 AUC는 0.806 이었고, 가장 높은 수치는 0.88을 기록하였다(그림2-1).

4. 텍스트 데이터를 이용한 기계학습 알고리즘 결과

피험자가 말한 전체 단어는 총 21,376단어였으며, 그 중 모든 피험자 중에서 한 번만 사용되었던 희귀 단어가 4.49%였고, 이는 분석에서 제외되었다. 피험자가 말하는 문장들 중 최대 길이는 1504글자, 평균 문장 길이는 33.91글자였다.

사용된 단어들 중에서 우울증환자에서 더 많이 사용되었기 때문에 중요도가 높은 단어들은 “입원”, “시도”, “실수”, “구토”, “쉬는”, “바닥” 등이 선정되었고, 이를 바탕으로 텍스트 마이닝(text mining)까지 진행하였다(그림 2-2, 그림 2-3).

현재 우울증 진단 모델은 인구통계학적 정보를 이용한 모델의 AUC가 0.761로, 텍스트를 이용한 모델의 AUC가 0.902로 더 높은 것을 확인하였다. 반면 인구통계학적 정보와 텍스트 정보로 학습한 앙상블 모델과 텍스트 정보만으로 학습한 모델 간에는 유의한 차이가 없었다(표 2-7). DeLong test를 통해 인구통계와 텍스트만을 이용한 각 모델의 ROC 곡선을 비교하였을 때 p-value 0.001의 통계적 유의하게 텍스트를 이용한 모델에서 더 성능이 우수함을 확인할 수 있었다. 하지만 텍스트만 이용한 모델과 앙상블 모델을 비교함에 있어서는 DeLong test 결과 통계적 유의성을 보이지 않았다(그림 2-4).

## 5. 목소리 지표들과 임상 변수와의 상관 분석

정상군과 우울증군에서 임상 변수의 차이가 크기 때문에, 임상 변수와의 상관 관계 분석은 두 군 내에서 이루어졌다. 정상군에서의 상관 분석에서 HDRS와 BAI, BAI와 BSS는 각각 통계적으로 유의한 상관관계를 보였다. 하지만 HDRS와 BSS의 관계는 통계적으로 유의한 상관관계를 보이지 않았다. 음성 지표들을 분석한 결과, HDRS와 상관을 보이는 지표는 pitch의 평균, error, pause portion과 total count_log 총 4개의 지표였고, BAI와 상관 관계를 보이는 지표는 time_log, spectral_band, formant0, change_p_log, change_m_log 총 5개의 지표였다. 또한 BSS와 상관 관계를 보이는 목소리 지표는 formant0, formant2 2개의 지표뿐이었다(표 2-8).

한편 우울증군에서는 HDRS, BAI, BSS 모두가 통계적으로 유의한 상관 관계에 있었다. 목소리 지표들의 경우 HDRS와 상관 관계를 보이는 지표는 spectral_rmse, formant_BW1 2개의 지표였고, BAI와 상관 관계를 보이는 지표는 time_log, change_p_log, change_m_log 3개의 지표였다. BSS와 상관 관계를 보인 지표는 spectral_centroid, spectral_band, spectral_rolloff, formant_BW1, ZCR, pause portion 총 6개의 지표가 있었다(표2-9).

## 6. 우울, 불안, 자살사고와 목소리와의 매개 분석

우울증이 목소리에 미치는 영향에 불안과 자살사고가 다중 매개를 함으로서 영향을 준다는 가정을 하에 목소리 지표와의 매개 분석을 하기

위하여 앞서 상관 분석을 진행하였다. 그 중 정상군은 우울, 불안 및 자살사고 위험도의 점수 자체가 낮게 분포되어 있기 때문에, 그 매개 관계를 잘 설명하기 위해서는 우울증군에서 자살사고와 상관관계를 보이는 목소리 지표들을 위주로 매개 분석을 시행하였다.

Spectral_band 지표에 대한 매개분석 결과, 정상군에서는 우울증상이 spectral_band에 미치는 직접 및 간접적 영향은 모두 통계적 유의성을 보이지 않았다(표 2-10, 그림 2-5). 하지만 우울증군에서는 우울증상 자체가 spectral_band에 미치는 직접적인 영향은 통계적으로 유의성을 보이지 않았으나, 우울증상이 불안과 자살사고에 의하여 연속적으로 매개되어 spectral_band에 미치는 간접적인 영향력과 우울증상이 자살사고에 의해 매개되어 spectral_band에 미치는 간접적인 영향력을 확인할 수 있었다(표 2-11, 그림 2-6).

## 제 3 절 고찰

기존의 연구 I의 타당성검증(validation)에서 가장 우수한 성능을 보였던 모델은 MLP 를 적용한 모델이었다. 하지만 연구 II의 새로운 데이터를 기반으로 평가한 검증(test)의 결과에서는 정확도가 가장 높은 모델은 SVM 모델이었으며, 민감도를 모두 포함한 성능이 가장 우수한 모델은 GNB 모델이로, 타당성검증시와는 다른 모델에서 각각 우수한 성능을 확인하였다. 이는 하나의 데이터 세트에 다양한 모델을 적용할 때의 성능은 데이터의 불균형 등에 영향을 받기 때문으로 판단된다. 따라서 향후에는 다양한 모델의 성능을 각각 비교하는 연구들이 추가로 진행되어, 음성 및 텍스트 데이터의 분석에 가장 유리한 모델이 어떠한 것인지에 대한 기준이 정해져야 할 것이다.

이어 연구 II에서는 정상군과 우울증군에서 목소리의 지표들과 인터뷰에서 사용하는 단어가 다를 것이라는 가정하에, 양 군 간의 차이를 유발하는 지표들에 대해서 알아보았다. 특히 연구 I에서 불안척도를 보정하지 않았다는 한계가 있으나 연구 II에서는 임상지표들을 보정한 뒤에도 통계적 유의성을 보이는 음성의 지표 7개를 발굴하였고, 이 중 3개의 지표는 불안 점수를 보정한 뒤에도 통계적 유의성이 유지되었다. 사회인구학적 변수들을 보정 뒤 정상군과 우울증군에서 차이를 보인 7개의 지표는 각각 time_log, spectral_band, mean_p, change_p_log, ZCR, pause portion, total count_log였다. Time_log는 피험자가 발화한 총 시간의 log 치환값, spectral_band는 성대의 특징을 반영하는 지표이다. Mean_p는 발화의 평균 음높이, change_p_log는 각 문장마다 평균 음높이의 변화정도의 log 치환값,

ZCR이 음성이 존재하는 비율, pause portion은 발화 중 공백이 존재하는 비율, total count_log는 발화된 전체 문장의 개수를 의미한다. 이를 해석해보자면 우울한 군에서 상대적으로 더 길게 발화하며, 발화되는 문장의 수가 많았다. 하지만 음성의 존재비율은 적고 공백이 많았으며 음높이의 평균 높이가 낮아져 있고, 각 발화구간 별 음높이의 차이가 커져 있었다. 우울한 군에서 음성을 머뭇거리는 특징과 평균 음높이가 낮은 특징을 보인다는 것은 기존의 연구와 일치하는 결과를 보였다(32,34,123). 하지만 우울한 군에서 발화시간과 양이 더 많다는 것은 기존의 연구와는 다른 결과였다. 이 것은 MINI라는 구조화된 인터뷰가 가진 특징을 살펴보면 이해할 수 있다. MINI는 우울한 기분부터 시작하여 성격장애에 이르기까지 다양한 정신 병리에 대해서 질문을 하도록 구성되어 있다. 특히 정신 병리가 있을 경우에는 그에 대한 부가적인 탐색을 하여 진단기준에 부합한지 확인하고, 정신 병리가 없을 경우에는 그 다음 파트로 넘어가도록 지시한다(124). 이러한 점을 고려하면 정신 병리가 없는 정상군에서는 질문의 항목 자체가 더욱 적었을 것이며, 그로 인한 대답 역시 줄어드는 결과를 보였을 것이다. 불안지표인 BAI를 보정한 뒤에도 ZCR, pause portion, total_count_log는 통계적 유의정을 유지하고 있었고, 이는 우울감이 발화중의 여백을 증가시킨다는 것을 더욱 강하게 지지한다.

정상군과 우울증군이 대답에서 활용된 단어들 중에서 빈도의 차이가 큰 단어들 상위 10개를 살펴보았을 때, “입원”, “시도”, “실수”, “구토”, “휴식", “바닥”, “충(벌레)”, “죽는”, “부정”, “퇴원” 등이 존재하였다. 본 연구에서는 한국어 감정 분류 사전의 한계로 인하여 감정어를 분류하지는 않았으나, “입원”, “실수”, “구토”, “바닥”, “충”, “죽는”, “퇴원” 등의 부정적 의미를 가진

단어들이 다수 포함되어 있는 것을 확인할 수 있었다. 이 것은 기계학습을 기반으로 한 텍스트 분석이 임상적으로도 중요한 단어를 추출하는 능력이 충분함을 보여주는 근거이다.

목소리의 지표를 이용한 모델과 텍스트 데이터를 이용한 각각의 인공지능 알고리즘을 구축하고, 그 구축된 알고리즘의 정확도를 확인해보았다. 목소리 지표를 이용한 모델의 경우에는 QDA 방법을 이용한 모델이 가장 높은 성능을 보였으며 모든 목소리 지표를 활용하는 모델에서 F1 값을 포함한 정확도, 민감도와 특이도 모두가 높아지는 것을 확인하였다. 5중 교차 검증의 결과, AUC값은 0.806, F1 score 0.820 상당히 우수한 수치를 보였다. 또한 텍스트 데이터를 이용한 모델의 경우에는 AUC 0.905의 높은 값을 보였으며, 임상지표와 함께 구축한 앙상블 모델과도 성능의 차이가 없었다. 이러한 수치는 기존의 연구들에서 우울증 존재를 분류하는 기계학습 모델의 정확도는 74~80%였고(25, 125), 여러 모델을 비교하였을 때에는 SVM이 Gaussian mixture models (GMM) 이나 MLP보다 더 나은 성능을 보인 연구 결과도 있었다(126). 하지만 본 연구에서는 기존의 연구에서는 다루지 않았던 Adaboost 기법과 QDA 기법을 통해 더 좋은 성능을 확인할 수 있었다. 이러한 기법들이 본 연구에서 더 좋은 성능을 보일 수 있는 것은 본 연구의 샘플 수가 적은 편이고, 목소리의 지표는 20가지가 들어가기 때문에 이들 간의 상호작용과 다양한 지표 등에 대한 학습이 이뤄지는 것이 그 원인으로 생각할 수 있다. 텍스트를 기반으로 한 분류 결과는 놀라운 정도였고, 민감도는 76.2% 였으나 특이도는 95.1%정도로 높고, AUC는 0.907 이라는 놀라운 지표를 확인하였다. 또한 이러한 지표는 통계적으로 유의한 차이를 보였던 사회인구학적 정보들인 나이, SES, BMI과 비정신과적

약물 복용력, 그리고 성별까지 종합하여 예측한 모델보다 더 정확한 수치를 보였다. 한편 사회인구학적 정보와 텍스트 데이터를 모두 사용한 앙상블 모델과 텍스트데이터만을 사용한 모델과는 성능의 차이가 없었다. 이러한 정확도를 보일 수 있었던 것은 MINI 인터뷰를 통해서 텍스트 데이터가 얻어졌기 때문에 비교적 열린 질문에 대한 대답이 많아 단어의 수가 상당히 많고 다양하고, 우울증 환자에서 부정적인 단어 사용의 빈도가 높기 때문이다. 또한 텍스트 모델과 앙상블 모델에서 성능의 차이를 보이지 않았던 것은 텍스트 자체가 두 군을 분류함에 있어서 우수한 지표임을 시사한다.

더욱이 연구 II에서는 HDRS, BAI, BSS와 음성 지표들 간의 상관관계까지 확인하였고, 환자군에서 우울정도와, 불안정도, 또한 자살 위험성과 상관관계를 보이는 음성의 지표가 각각 다르다는 것을 확인하였다. 이를 바탕으로 매개분석까지 진행하였고, 그 결과 환자군에서 불안감 자체가 우울증상과 목소리의 변화 사이를 매개하는 영향은 확인할 수 없었으며, 자살사고가 동반될 경우에는 목소리의 변화를 매개하고, 그 관계에서는 불안감도 중요한 역할을 한다는 점까지 확인할 수 있었다.

본 연구는 인터뷰를 통해 획득한 환자의 목소리와 환자가 사용한 텍스트 데이터를 기반으로 세계 최초로 우울증을 분류하는 알고리즘을 구축하였고, 다양한 기계학습 모델을 접목시켜본 최초의 연구로서, 다양한 강점을 가지고 있다. 첫째로 사용된 텍스트가 실시간 인터뷰에서 사용된 텍스트라는 점이다. 본 연구 II에서는 음성 녹음 파일을 텍스트화시키는 toolkit를 활용하였기 때문에, 인터뷰 녹음을 즉각적으로 텍스트화 할 수 있다는 장점이 있다. 또한 툴킷으로 인하여 텍스트화된 데이터 중에서 오류에 의해 인식되지 않은 텍스트만을 바탕으로 분류하는 것에도 상당히 좋은 성능을 보였기 때문에

향후 실제 임상 현장에서의 인터뷰를 진행하면서 자동으로 텍스트화하여 활용할 수 있다는 가능성이 있다. 둘째, 텍스트와 인구통계학적 데이터를 활용할 수 있는 앙상블 알고리즘을 구축하였다는 점이다. 앙상블 모델을 구축할 때에는 텍스트의 정확성을 확인하기 위하여 포함되는 변수를 통계적 유의성을 보이는 지표들 외 추가되는 변수를 최소화하였으나 그 정확도는 텍스트를 기반으로 한 알고리즘과 유사한 정도였다.

하지만 본 연구에서는 여러 가지 한계 역시 가지고 있다. 첫째, 본 연구에 사용된 텍스트 내용은 MINI 질문에 대한 대답만을 기반으로 하였기 때문에 중요한 데이터가 포함되지 않았을 가능성이 있다. 하지만 그림 2-2를 미루어 볼 때, 같은 질문에 대해서도 정상군과 우울증군이 다른 단어를 사용하며 대답을 하고, 우울증군에서 중요하게 사용되는 단어들은 부정적 의미를 가진 단어들이 많았기 때문에 본 한계를 다소간 극복할 수 있다. 하지만 향후에는 더 개방된 질문에 대한 인터뷰 등을 기반으로 정상군과 우울증군이 사용하는 단어에 대한 대규모의 연구가 진행될 필요가 있다. 둘째로 본 연구 II의 참가자 중 80%는 여성이었으며, 86.75%는 양극성장애를 진단받았다. 즉 본 연구의 결과가 모든 인류를 대상으로 일반화할 수 없다는 한계가 있다. 본 연구에 참여한 참가자들의 선택 편향(selection bias)가 있을 가능성도 있다. 또한 본 연구는 3차 의료기관에서 이루어진 연구이기 때문에 “입원” 등과 같은 단어가 더 중요하게 사용되었을 가능성도 있다. 만약 입원 병상이 없는 병원이었다면 “전원” 등의 단어가 더 중요하게 사용되었을 가능성도 있다. 따라서 다양한 연령대와 다양한 임상현장의 인터뷰 내용을 바탕으로 본 연구의 결과가 재확인되는지 살펴 보아야 한다. 또한 최근에는 목소리의 지표를 분석하는 방법으로서 멜 주파수 스펙트로그램 (mel frequency

spectrogram) 등 지표를 계산하여 값으로 산정하지 않고, 스펙트로그램 자체를 학습시키는 연구들 역시 진행되고 있다(127, 128). 또한 텍스트분석에는 자연어 처리 기법(natural language processing) 중 감정어 분류를 통한 연구가 진행되고 있다(105). 하지만 이러한 기법을 본 연구에 적용하기에는 목소리 변화에 대한 임상적 의미를 찾을 수가 없었고, 감정어 분류의 경우 현존하는 국내 감정 분류 사전 3가지를 살펴보았지만 본 연구에서 추출된 단어들에 대한 분류가 미흡하여 활용할 수 없는 상황이었다. 따라서 향후 본 연구의 데이터에도 더 다양한 분석 기법을 적용하여 알고리즘을 정밀화하는 과정이 필요하겠다.

# 제 4 장 목소리와 텍스트 데이터를 통해 자살고위험군을 진단하는 인공지능 알고리즘 개발 (연구 III)

연구 III에서는 연구 II의 피험자 중 우울증군 83명에 대한 하위집단 분석으로서, 우울증군 내에서도 자살 저위험군(depression with low suicide risk, DLSR)과 자살 고위험군(depression with high suicide risk, DHSR)으로 하위 집단을 구분하여 분석을 시행하였다.

임상의가 자살의 위험성을 판단해야 하는 환경은 다양하지만, 정신과적 문제가 없는 사람을 대상으로 자살의 위험성을 판단하는 경우는 많지 않다. 임상 현장, 즉 외래나 입원 병실, 그리고 응급실에서 정신과적 문제와 함께 자살에 관련된 생각을 호소하는 이들 중 정말로 자살 위험성이 어느 정도 일지를 판단하는 것이 어려운 점이다(129). 따라서 본 연구 III에서는 우울한 환자들 중에서 자살 고위험군과 저위험군을 구분하고, 구분된 양 군의 목소리 지표의 차이를 알아보고자 한다. 또한 목소리 지표와 연구 II에서 활용되었던 텍스트 데이터를 기반으로 두 군을 구별할 수 있는 기계학습 알고리즘을 구축하고 그 정확도를 알아보고자 한다.

자살 고위험군을 정의함에 있어서도 다양한 방법이 있을 수 있다. 본 연구 III에서는 첫번째로는 BSS의 점수를 이용해 자살 고위험군을 정의하였다. BSS는 정해진 절단 기준점이 없는 점수이지만 점수가 높을수록 자살의 위험도가 높다는 것을 시사한다(108,109). 한 연구에서는 BSS의 절단 기준점을 24점으로 할 경우 자살 사망자를 하지만 본 연구에서 BSS를

24점을 기준으로 분류하면 민감도 100%, 특이도 90%라는 연구 결과도 있었지만, 하나의 연구에서 사용된 절단 기준점이기 때문에 공인된 기준으로 판단하기 어렵다(130). 따라서 연구 III에서는 전체 피험자 166명의 BSS 점수 분포를 살펴보고, 상위 25%의 절단 기준점인 17을 기준으로 양 군을 구분하는 방법을 우선적으로 선택했다.

둘째로는 본 연구에서는 MINI를 사용하였고, MINI에 포함된 자살성 모듈은, 최근 한 달의 상황을 묻는 5개 항목과 평생 동안의 상황을 묻는 1개 항목으로 구성되어 자살 위험성을 평가하는 평가자 평정 척도이다. 적어도 한 문항에 그렇다고 응답한 경우 현재 자살 위험성의 존재를 시사한다. 각 문항의 점수를 합산하여, 현재 자살 위험성을, 저위험(1-5점), 중위험(6-9점), 고위험(10-33점)으로 평가한다(67). 따라서 MINI에서 최근 1달 이내 자살 위험성이 있고, 그 위험성이 고위험으로 평가된 피험자를 자살 고위험군으로 분류하고, 그 외의 우울증군은 자살 저위험군으로 분류하여 분석을 진행하였다. MINI의 자살성 모듈은 다양한 연구에서도 자살 위험을 평가하는 기준점으로 활용되고 있다(131, 132).

자살 예방 대책의 효과적인 접근을 위해 자살 고위험군을 정의하고 이에 대한 효과 등을 확인하는 과정은 필요하다(133). 하지만 자살 위험군에 대한 정의는 분명하지 않으며, 선정하는 범위에 따른 진단의 정확도가 달라질 가능성이 있다. 자살 고위험군의 범위에 따라서 본 연구 III에서 개발하는 진단 알고리즘의 민감도와 특이도가 변화할 수 있다(134). 따라서 연구 III에서는 비교적 자살 위험도 평가에서 전세계적으로 많이 활용되는 BSS와 MINI의 자살성 모듈 두가지를 각각 적용하였을 때의 자살 고위험군을 정의하고, 이에 따른 진단의 알고리즘을 구축하여 평가해보고자 한다.

## 제 1 절 연구대상 및 방법

### 1. 연구 대상자

연구 II에서 우울증군으로 분류된 83명을 연구 III의 대상으로 선정하였다.

### 2. 평가 항목

연구 II를 통해 모집한 사회인구학적 정보와 임상 정보 자료, MINI, HDRS, PHQ, BAI, BIS, BSS 척도 자료를 사용하였다.

연구 III에서는 연구 II와 같이 정신질환 외 다른 질병으로 복용하고 있는 약물에 대해 확인하였다. 피험자들이 복용하는 약물에는 현재 여성호르몬제, 피나스테리드, 레보티록신, 테노포비르, 아미트립틸린, 메트포르민 등 다양한 약물을 복용하고 있지만 그 중 음성변화와 관련되었다고 보고된 약물은 여성호르몬 대체요법 뿐이었다(135). 다만, 현재 1명만이 여성호르몬 대체요법 중이었기 때문에 이로 인한 목소리의 변화는 본 연구 III에서는 고려하지 않았다. 연구 III에서는 추가적으로 질병으로 인한 주관적 통증에 대한 정보를, 1에서 5까지의 리커트 척도를 통해 수집하였다. 이는 자살 고위험군과 저위험군의 타과 약제를 복용하는 비율이 같더라도, 통증 등의 지병으로 인한 고통의 정도가 다를 수 있고, 스스로 느끼는 고통의 정도가 우울감에 미치는 영향이 있을 수 있기 때문이다(136).

그 외 연구 III에서는 질병으로 인한 주관적 괴로움을 추가적으로 분석하였다. 질병이 있는 유무만으로는 그 주관적 괴로움의 정도는 다를 수

있으며, 주관적 괴로움이 심할수록 우울감 등의 기분에 미치는 영향 역시 커지기 때문이다(137). 또한 연구 I에서와 같이 복용하는 항정신병약물의 복용량을 1일 약물 복용량 기준으로 olanzapine(OZP)을 기준으로 등가용량을 환산한 뒤 합산하였다(65,66).

3. 목소리 지표와 텍스트 데이터

연구 II를 통해 획득한 목소리 지표와 텍스트 데이터를 동일하게 사용하였다.

4. 통계학적 방법

범주형 변수는 피어슨 카이 제곱 검정(Pearson's chi-squared test)를 통해, 연속형 변수는 스튜던트 T 검정(student’s T test)를 통해 분석하였다. 목소리의 지표들을 다변량 공분산분석(Multivariate analysis of covariance, MANCOVA)방법을 적용하여 임상 변수들에 대한 보정을 시행하였다. 통계분석에는 Windows용 IBM SPSS Statistics 버전 25.0(SPSS Inc., Chicago, IL, USA)을 사용되었다.

5. 기계학습

음성 지표들을 기반으로 우울증 환자를 구별하는 이분법 과제에서는 연구 II에서 진행하였던 LR, GNB, Adaboost, QDA. SVM, MLP 6가지 방법을

모두 적용하여 성능을 확인하였다. 또한 연구 II에서와 같이 전체 데이터의 80%를 학습시키고, 남은 20%의 데이터를 정확도를 확인에 이용하였으며, 기계학습의 성능에 대한 지표들은 5중 교차검증을 통한 값들의 평균 수치로 산정되었다.

텍스트 데이터를 기반으로 우울증 환자를 구별하는 방법 역시 연구 II와 같이 나이브 베이즈 (Naïve Bayes) 방법을 활용하였다. 연구 III에서의 앙상블모델의 경우에도 사회인구학적 정보들 중 양 군에서 통계적 유의성을 보이며 차이가 있는 변수를 포함하여 구성하였다. Naive Bayes 분류기는 Natural Language Toolkit 패키지(120)와 scikit-learn 패키지가 사용되었다. 또한 각 기계학습 모델의 ROC 값은 DeLong test에 의하여 비교하였다(121).

## 제 2 절 결과

### 1. 기준 별 자살 위험이군 분류의 차이

연구 II에서 정상군과 우울증군 모두를 합쳐 BSS값의 분포를 확인하였을 때, 최소값은 0, 최대값은 35였으며 평균은 9.627이었으며, 상위 75% 이상은 18점이 기준이었으나, 18점을 기준으로 절단하기에 17점에 분포되어 있는 피험자가 6명으로 많았다. 따라서 BSS를 기준으로 분류함에 있어서는 자살 고위험군을 다소 넓은 범주로 정의하여 분석하기 위해, 17점 이상인 피험자 50명이 자살 고위험군으로 분류되었고, 그 외 33명의 피험자는 자살 저위험군으로 분류되었다.

한편 MINI의 자살성 모듈을 기준으로, 최근 1달 이내 자살 위험이 있고, 그 위험에 대한 평가가 높다고 평가된 피험자는 총 31명이었다. 이 31명이 자살 고위험군으로 분류되었고, 그 외 52명의 피험자가 자살 저위험군으로 분류되었다.

### 2. 자살 저위험군과 자살 고위험군의 비교 (BSS를 기준으로)

BSS를 기준으로 자살 저위험군 33명과 자살 고위험군 50명에 대한 사회인구학적 정보를 비교하였을 때, 양 군에서 BMI, SES, 성별, 정신과 외 다른 질병으로 복용 중인 약물 유무, 그로 인한 주관적 고통 모두 통계적으로 유의한 차이를 보이지 않았다. 오직 나이에서만 차이를 보였고, 자살 저위험군의 평균 연령은 33.788으로, 자살 고위험군(평균 연령 29.02)보다

나이가 많았다. 또한 양군에서 복용중인 항정신병약물을 olanzapine을 기준으로 등가용량으로 치환하여 총 합을 하였을 때, 용량에서의 차이를 보이지 않았다.

한편 양 군에서 MINI를 통해 진단된 정신건강의학과적 질병 유무를 비교해보았을 때, 자살 고위험군에서 통계적으로 유의하게 공황장애가 많음을 확인할 수 있었다. 그 외 질환들은 양 군에서 차이를 보이지 않았다(표 3-1). 임상적 변수에 대해서 살펴볼 때, BIS를 제외한 PHQ, HDRS, BAI, BSS 모두 통계적으로 유의하게 자살 고위험군에서 높음을 확인할 수 있었다(표 3-2).

목소리 지표들을 양군 간 비교하였을 때, 아무 보정이 없이 자살 저위험군과 고위험군을 비교한 결과 formant_BW0, 1, 2 총 3개의 지표에서 통계적으로 유의한 차이를 보였다. 하지만 사회인구학적 정보 중 양 군에서 차이가 있었던 나이를 보정한 뒤에도 통계적으로 유의한 차이를 유지하는 지표는 formant_BW1, 2 뿐이었다. 한편 자살 고위험군에서 공황장애를 동반한 환자가 많았고, 실제 임상지표에서도 BAI가 높았기 때문에, BAI를 보정한 뒤에 통계적으로 유의한 차이를 보이는 지표를 확인해보았을 때에는, formant_BW1, 2와 formant0까지 총 3개의 지표가 자살 위험도에 따른 차이를 보였다(표 3-3).

3. 자살 저위험군과 자살 고위험군의 분류 알고리즘 (BSS를 기준으로)

20개의 목소리 지표를 바탕으로 자살 고위험군과 저위험군을 구별하는 알고리즘에서는 LR 기법과 SVM 방법이 F1 점수 0.789, 정확도 0.718으로

가장 높은 점수를 보였다. 전반적으로 각 모델에서 민감도는 낮아지고 특이도가 높아진 경향을 확인할 수 있었다. MLP 모델이 민감도가 0.535으로 가장 높았으며, 특이도는 SVM이 0.840으로 가장 높음을 확인할 수 있었다(표 3-4, 그림 3-1).

사회인구학적 변수를 예측변수로 사용하는 모델에서 자살 저위험군과 자살 고위험군 간의 통계적으로 유의한 지표였던 연령을 포함하여, 항정신병 약물 용량, SES, BMI 및 두 그룹 간에 차이가 나는 BAI 점수 등이 포함되었다. 사회인구학적 모델을 통해 자살 고위험군과 저위험군을 구분할 때에는 AUC 0.511 이었으나, 텍스트를 기반으로 자살 위험군을 분류할 때에는 AUC가 0.626, 정확도 0.602, 민감도 0.489, 특이도 0.750 이었다. 또한 사회인구학적 정보와 텍스트 모두를 활용한 앙상블 모델에서는 AUC 0.874, 정확도 0.795, 민감도 0.703, 특이도 0.870 이었다. 또한 앙상블 모델의 성능은 텍스트를 이용한 모델보다도 DeLong 검정에 의하여 더 우수함이 확인되었으나, 텍스트를 이용한 모델의 경우 사회인구학적 모델과 비교하였을 때 그 우수성이 통계적으로 입증되지 않았다($p$ value = 0.308)(표 3-5, 그림 3-2).

4. 자살 저위험군과 자살 고위험군의 비교 (MINI를 기준으로)

MINI의 자살성 모듈을 기준으로 하였을 때, 자살 고위험군은 31명, 자살 저위험군은 52명이었다. 이들의 사회인구학적 정보를 비교하였을 때, 자살 고위험군의 평균 연령은 27.645로 저위험군의 평균연령 32.865보다

낮았으나, 그 외 BMI, 성별, 다른 질환으로 인한 주관적 괴로움 등에서는 통계적으로 유의한 차이를 보이지 않았다. 또한 자살 고위험군이 복용하는 항정신병약물은 olanzapine 기준 8.082mg 으로, 자살 저위험군의 복용 용량 4.705보다 높은 용량으로 복용하고 있었다. 그뿐 아니라 공황장애 뿐 아니라 사회불안장애도 자살 고위험군에서 더 많음을 통계적으로 유의하게 확인할 수 있었다. 하지만 그 외 BMI, 성별, SES, 타과 약제 복용 유무 및 타과 질환으로 인한 괴로움의 정도 모두 통계적으로 유의한 차이를 보이지 않았다(표 3-6).

임상 변수들은 BSS 기준으로 분류한 것과 마찬가지로 BIS를 제외한 모든 변수에서 자살 고위험군의 점수가 더 높다는 것을 통계적으로 유의하게 확인할 수 있었다(표 3-7).

한편 목소리 지표들을 스튜던트 T 검정에 의하여 비교하였을 때에는 pause portion 만이 양 군에서 차이를 보였다. 하지만 통계적 차이를 보였던 나이와 항정신병약물 용량을 보정한 뒤에는 formant_BW1 지표만이 통계적으로 유의하게 두 군간 차이를 보였다. 그리고 BAI까지 보정한 뒤에는 그 통계적 유의성 마저 보이지 않게 되었다(표 3-8).

5. 자살 저위험군과 자살 고위험군의 분류 알고리즘 (MINI를 기준으로)

목소리 지표를 바탕으로 기계학습 모델을 통해 자살 고위험군을 분류하였을 때, Adaboost 모델에서 F1 점수 0.499, 정확도 0.576, 민감도 0.490으로 가장 성능이 우수하였다. 특이도의 경우 QDA 모델에서

0.983으로 가장 우수하였다(표 3-9, 그림 3-3)

사회인구학적 정보를 기반으로 MINI 기준의 자살고위험군을 구분하였을 때 AUC는 0.499에 불과하였으나, 텍스트를 기반으로 예측하는 모델의 AUC는 0.632로 높아지는 것을 확인할 수 있었다. 그러나 인구통계학적 특성과 텍스트 모델을 포함한 앙상블 모델의 경우 AUC가 0.800으로 두 모델보다 높았다(표 3-10, 그림 3-4).

## 제 3 절 고찰

연구 III에서 자살 위험군에 대한 정의는 서론에서 언급하였듯 아직 전세계적으로 통용되는 기준은 부재하였다. 따라서 본 연구에서는 자살사고에 대한 자가보고 설문지인 BSS를 기준으로 한 분류와, 관찰자 평가 설문지인 MINI를 기준으로 한 분류를 모두 사용하였다. 그 결과, BSS 기준과 MINI의 자살성 모듈 기준의 자살 고위험군이 각각 다르게 분류되었다. BSS 척도는 절대적 기준이 없었기 때문에 본 연구에서의 기준값이 17점은 다소 넓은 기준점일 가능성이 있고, 실제 본 연구의 결과에서도 BSS의 기준으로는 자살 고위험군이 50명, MINI의 기준으로는 자살 고위험군이 31명으로 분류되는 것을 확인할 수 있었다.

자살 저위험군과 고위험군간 사회인구학적 정보를 비교해보았을 때, 자살 고위험군에서 통계적 유의성을 보이며 나이가 더 젊고 공황장애의 진단이 많이 동반되었다. 양극성 장애 환자 중에서 자살 고위험군의 특징을 살펴보았던 선행의 연구들에서도, 자살 고위험군이 더 나이가 젊었고, 공황장애 등의 동반 질환이 많음을 확인할 수 있었다(138-140). MINI를 기준으로 한 더 좁은 의미의 자살 고위험군에서는 사회불안장애 역시 통계적으로 유의하게 많이 동반하고 있었다. 사회적 불안은 사회적 소속감을 저하시키고, 사회적 상황에서의 부담감을 증가시키며 그로 인하여 우울증과 함께 자살의 위험을 높인다고 알려져 있다(141,142)는 기존의 연구의 결과도 일치하였다.

BSS를 기준으로 할 때와 MINI를 기준으로 할 때 모두 임상적 지표들에 대해서는 BIS를 제외하고 자살 고위험군에서 모두 높은 점수를 확인할 수

있었다. 충동성이 높은 피험자들이 자살의 행동을 하는 경향성이 높다는 선행의 연구들도 있으나(143,144), 충동성이 높은 피험자에서 꼭 자살 시도율이 높지는 않았다는 보고 역시 있었다(145). 본 연구에서는 충동성에 대한 평가지표로서 BIS를 활용하였고, BIS는 Barratt이 제안한 충동성과 불안은 성격적 특성 중 하나라는 점에서 착안된 척도이다(146). 초기의 BIS는 불안과 관련된 특성 중 하나로서 충동성을 평가하였다면, 점차 충동성에 대한 이해가 커지면서 행동적 충동성, 인지적 충동성 그리고 비계획적인 충동성 이렇게 3가지 측면에서 평가하는 것으로 발전하였다(147). 하지만 충동성이라는 것은 타고난 성향적인 영향도 있으나, 기분의 변화 및 주변 환경의 변화에 따라서도 달라질 수 있는 지표이기도 하다(148, 149). 따라서 본 연구에서 기분과 자살 위험에 따른 충동성의 차이를 관찰하기에 BIS는 다소 적합한 척도가 아니었을 가능성이 있다. 그 외 HDRS, PHQ, BAI, BSS는 모두 현재 환자의 상태를 반영한 지표이기 때문에 자살 고위험군에서 더 높은 점수를 보이는 것은 충분히 이해할 수 있다.

BSS 기준으로 자살 고위험군에서 목소리 지표의 차이를 비교해보자면, formant_BW1, 2 지표가 사회인구학적 인자들을 보정한 뒤에도 통계적 유의성을 보였다. 성대 특성의 변화는 기분 변화에 중요한 역할을 하는 GABA 신경전달물질에 의하여 근육 긴장이 증가함에 의해 영향을 받는다(150). 또한 화자의 정신 상태 변화와 관련된 타액 분비 및 점액 분비의 변화에서도 영향을 받는다(151). 이러한 변화는 성대의 음향 공명과 관련된 정보를 포함하는 포먼트 특징에 의해 자세히 포착된다(152). 특히 BSS를 기준의 자살 고위험군에서는 나이와 불안감까지 보정한 뒤에도

통계적 유의성을 보였던 지표들이 3가지 있었다. Formant0는 성대진동율(rate of vocal fold vibration)으로서, 성대의 물리적 상태의 지표이자 정서적 상태의 지표로서, 우울증과 관련되어 변화하는 지표임을 여러 논문을 통해 밝혀져 있다(153-155). 또한 formant_BW1,2의 경우 기존의 연구에서는 우울증 및 자살 위험군에서 정상군보다 더 낮아져있는 경향을 보였다(156). 기존의 연구에서는 formant_BW의 경우 우울증상이 있을 경우 근육의 긴장이 증가하여 성대의 공명이 감쇠하기 때문에 그 수치가 작아지는 효과가 있다고 알려져 있다(157). 하지만 본 연구에서는 그 수치가 자살 고위험군에서 더 증가되어 있었다. 이 것은 우울증상과 자살 위험군은 다른 성향을 가지기 때문으로 해석된다. 근긴장도를 포함한 정신운동기능을 평가한 연구 결과 근긴장이 이완되어 있는 군에서 5년 간 추적관찰 시 자살 시도를 더 많이 하는 것이 확인되었다(158). 통상적으로 근육을 이완시키는 방법은 불안 등을 낮추는 행동치료의 기법 중 하나로서, 우울, 불안이 높을수록 근육이 긴장되어 있다고 알려져 있다(159). 이 것은 자살 고위험군의 경우는 우울증과는 또 다른 심리적 상태이기 때문에 근긴장도에도 차이가 있을 가능성이 있다. 자살 행동을 하는 사람들이 보고하는 가장 일반적인 동기는 혐오스러운 상황에서 벗어나 참을 수 없는 마음 상태에서 휴식을 취하는 것이다(160). 즉 다방면에서 생긴 내적 문제로 인하여 발생한 많은 갈등으로부터 자살이라는 방법으로 심리적 고통을 해결한다는 선택을 내린 것이다. 따라서 이미 이들은 마음 속에서 해결 방법을 결정하였기 때문에 오히려 더 희망적이고, 미래 지향적이 될 수 있다. 이 것은 실제 자살 시도자의 유서에 남긴 단어를 통해서도 입증되어 있다(44).

하지만 MINI를 기준으로 한, 조금 더 좁은 의미로 자살 고위험군을 분류하여 목소리 지표를 비교하였을 때에는 통계적 유의성을 보이는 지표를 찾을 수 없었다. 불안을 보정하지 않은 상태에서는 BSS와 같이 formant_BW1이 통계적으로 유의하게 자살 고위험군에서 높음을 확인할 수 있었으나 불안증상을 보정한 뒤에는 그 통계적 유의성이 사라지게 되었다. 이 것은 자살 고위험군의 피험자 수가 충분히 크지 않았기 때문일 가능성이 있다.

각 기준의 자살 고위험군을 목소리 지표들을 이용하여 예측하는 기계학습 모델에서는 MLP에서 민감도는 0.535로 가장 높았으나, F1 점수가 0.789인 LR 모델과 SVM 모델이 가장 성능이 우수함을 확인하였다. 하지만 이 두 모델의 경우 민감도는 각각 0.495와 0.462로 0.5를 넘지 않는 정도로 진단적 지표로서 목소리를 활용하기에는 어려운 정도였다. MINI를 기준으로 한 자살 고위험군의 분류는 더욱 성능이 좋지 않았다. 이 것은 데이터 양 자체가 부족하고, 자살 저위험군과 고위험군에서 통계적 차이가 두드러지는 지표가 희소하기 때문일 것으로 추정한다. 하지만 목소리가 자살의 위험성과 연관되어 있다는 많은 선행 연구들이 있기 때문에, 더 대규모의 피험자를 대상으로 한 연구를 통해 본 연구의 결과는 다시 한 번 확인되어질 필요가 있다.

텍스트를 기반으로 한 분류 역시 연구 II와 정상군과 우울증군을 구별하는 모델보다 성능이 좋지 않았다. BSS를 기준으로 할 때의 모델 정확도는 0.602, 민감도 0.489, 특이도 0.750 에 불과하였고, MINI를 기준으로 하면 정확도 0.602, 민감도 0.744, 특이도 0.477 이었다. 하지만 텍스트를 기반으로 한 분류는 앙상블 모델을 통해서 성능의 부족함을 극복할 수

있었다. 앙상블 모델의 경우에 BSS 기준의 자살 고위험군을 분류하는 모델의 AUC는 0.874, MINI 기준의 모델 AUC는 0.800 으로 비교적 우수한 성능을 보였다. 본 연구 III과 같이 피험자의 수가 충분히 많이 않을 때에는 텍스트와 같은 지표들은 사회인구학적 정보와 함께 활용될 때에는 객관적 진단 도구로서 유용할 수 있음을 보여주었다.

본 연구에는 많은 강점이 있다. 첫째, 기존의 연구들과 다르게 자살 위험군을 우울증 군 내의 하위 집단간 비교를 통해 기계학습이 더 임상적 상황과 유사한 진단을 내리도록 한 것이다(40). 기존의 연구에서는 정상군와 우울증을 모두 포함하여 자살 위험군을 정의하고, 이들을 구분하도록 하였다. 본 연구에서도 자살 저위험군에 정상군을 포함하여 분석을 시행해 보았고, 그 결과는 아주 우수하였다. 하지만 이러한 결과는 연구의 목적에 맞지 않는 분석이기 때문에 보고하지 않았다. 본 연구의 목적은 목소리와 텍스트 데이터가 임상에서 활용될 수 있기 위한 초석으로서, 실제 임상에서 사용되는 인터뷰를 기반으로 데이터들이 확보되었다. 임상의 현장에서 자살 위험군을 평가하는 상황은 우울감과 함께 자살사고를 호소하는 피험자들에 대해서 우울함에 대한 평가를 시행하고 이후 자살 사고에 대해 더 추가적으로 평가를 하는 것이 기본적이다(161). 실제 연구 II에서 확인하였듯 정상군은 이미 우울증군과 다른 목소리의 지표들을 가지고 있으며, 사용하는 텍스트 역시 달랐다. 하지만 본 연구에서 확인해본 목소리 지표는 우울증 군 내에서 자살 저위험군과 자살 고위험군에서 크게 다르지 않았다. 따라서 이러한 아주 세밀한 차이를 기계학습을 통해 보완할 수 있는지에 대한 평가를 시행하였고, 본 연구의 결과 텍스트 지표를 임상 지표들과 함께 활용할 때 그 성능이 입증되었다. 둘째 연구 II와 같이 본 연구에 활용된 목소리와 텍스트

데이터가 실제 인터뷰를 통해 획득되었다는 점이다. 이 역시 실제 임상의 현장과 유사한 환경을 조성하기 때문에, 본 연구의 결과를 임상에 더 잘 활용할 수 있다. 셋째 텍스트와 인구 통계 데이터를 활용할 수 있는 앙상블 알고리즘을 구축하였다는 점이다. 본 연구에서 텍스트 데이터만으로는 자살 고위험군을 예측하는데 한계가 있었다. 하지만 거기에 그치지 않고 임상의 변수들과 통합한 앙상블 모델을 만들었고, 그 결과 텍스트데이터가 임상 지표와 함께 활용될 수 있는 가능성을 확인하였다. 이 앙상블 모델은 음성 지표들까지 모두 통합한 알고리즘까지 확대할 가능성을 열어두었기 때문에, 임상 현장에서 확보할 수 있는 더욱 표정 등의 더욱 다양한 객관적 지표들에 대한 연구의 초석이 될 것이다.

본 연구의 한계점 역시 존재한다. 가장 큰 한계점은 대상자의 수가 충분히 크지 않다는 것이다. 물론 기존의 연구들에 비하여 자살 고위험군이 31명 이상이라는 것은 적은 수는 아니지만, 통계적 차이를 입증할 만큼의 충분한 수는 아니었던 것으로 보인다. 따라서 본 연구의 결과가 대규모의 대상으로도 재현되는지에 대해서는 추가적인 연구가 필요하다. 둘째 피험자의 77%가 여성이고, 양극성장애의 환자가 전체의 86.7% 이라는 점에서 연구 II의 한계와 같이 피험자가 전체 자살 고위험군을 대표할 수 없다. 물론 자살 저위험군과 자살 고위험군의 주요 우울 장애와 양극성 장애 진단 비율은 통계적으로 다르지 않았고, 성별 역시 통계적 차이를 보이지 않았지만 본 연구 III의 결과로서 제시된 목소리 지표들의 차이들은 더 다양한 피험자에서도 확인할 필요가 있다.

# 제 5 장 종합 고찰

## 1. 연구의 목적

본 연구는 현재 정신건강의학과 진단 체계의 한계를 인식하고 이를 극복하기 위한 연구이다. 현재의 한계로서 첫째 우울과 자살 위험성에 대한 판단은 전적으로 대상자가 호소하는 증상을 위주로 평가된다는 한계가 있다. 물론 이러한 한계를 극복하기 위하여 임상가는 신뢰할 수 있는 정보를 제공하는 보호자와의 면담을 시행하여 증상의 객관적 증거들을 탐색하기도 한다. 하지만 이 역시 믿을 수 있는 보호자를 확보하기 어려운 환경, 즉 자살 예방 콜센터 혹은 보호자와 함께 내원하지 않은 응급실 등과 같은 경우에는 보호자와의 면담을 시행하기 어려울 때가 많다. 둘째 임상가의 주관적 경험을 바탕으로 내려진 평가를 위하여 많은 시간의 경험을 필요로 한다. 하지만 자살 위험을 평가하게 되는 순간은 주로 응급실 등과 같은 상황이 많고, 이러한 때 숙련된 정신과적 전문의로부터 평가를 받는 경우는 드물다. 따라서 전문의가 가지고 있는 많은 숙련된 경험을 보조할 수 있는 임상의사결정지원시스템을 구축한다면 임상적으로 많은 도움이 될 수 있으리라 판단하였다. 더불어 이러한 시스템이 인터뷰를 통해 즉각적으로 얻어진 정보를 바탕으로 구축된다면 그 임상적 활용 가치는 높아질 것으로 기대하였다.

따라서 본 연구에서는 인터뷰를 통해 획득한 목소리와 텍스트 데이터를 분석하고, 이를 바탕으로 우울증과 자살 고위험군을 진단하는 알고리즘을

개발하고자 하였다. 이러한 알고리즘을 개발하기 위하여, 3개의 세부 연구로 나누어 진행하였다. 연구 I에서는 2차 년도의 초기 데이터를 기반으로, 정상군과 경도 우울증군, 주요 우울증군으로 분류하여 목소리 지표들의 차이들을 비교 분석하였고, 다양한 기계학습 모델을 접목하여 세 군을 구분하는 초기 알고리즘을 구축하였다. 하지만 연구 I에서의 음성 지표들 중에서 성대의 특징을 반영하는 glottal feature들의 경우 값을 추출하는 시간이 24시간 이상이 필요하게 되었다. 이러한 음성지표가 반영된 알고리즘은 임상 현장에서의 활용가능성을 낮추기 때문에, 추후의 연구에서는 해당 지표를 제외하는 방향으로 알고리즘을 수정하게 되었다. 또한 연구 I에서는 자살 고위험군의 수가 작았기 때문에, 자살 위험을 진단하는 알고리즘을 구축하기에 어려움이 있었다. 하지만 임상적으로도 자살 고위험군은 정의에 따라 그 빈도가 다르지만, 유병률이 높지는 않은 상태이다. 따라서 자살 고위험군을 일차적으로 진단하게 될 경우, 임상적 활용도가 떨어지게 되므로, 연구 II와 III을 통하여 우울증을 정확하게 진단한 뒤 자살 위험성에 대한 평가가 진행되는 단계적 알고리즘을 구축하고자 하였다.

연구 II에서는 최종적으로 모집된 초기평가 데이터를 통해 정상군과 우울증군의 구분을 시도하였다. 또한 목소리 지표들 뿐 아니라 환자들이 면담에서 사용한 텍스트 데이터를 추가로 추출하여 진단 알고리즘을 구축하였다. 연구 III의 경우 연구 II 에서 모집된 우울증군을 자살 저위험군과 자살 고위험군으로 하위그룹을 구분하여, 자살 고위험군의 목소리와 텍스트를 기반으로 자살 고위험군 진단 알고리즘을 구축하였다. 더욱이 텍스트 데이터는 임상적 지표들과 통합적으로 활용할 수 있는 앙상블 모델을 구축하여 그 기능을 살펴보았다.

## 2. 연구의 결과 요약

연구 I의 결과 경도 우울증군은 통계적으로 유의하게 다른 두 군보다 나이가 많았고, 88.5%에서 항정신병약물을 복용하고 있었다. 따라서 경도 우울증군보다는 부분 관해된 우울증군으로 보는 것이 더욱 합당하였다. 이 점을 고려하여 음성의 지표를 분석하였을 때, 정상군과 우울증군, 그리고 경도 우울증의 순서대로 목소리 지표들이 변화하는 것을 확인할 수 있었다. 이들을 구분하는 목소리의 지표로 총 7개의 지표가 확인되었고, 이를 해석해볼 때 정상에서 우울, 그리고 우울 후 부분 관해가 되어도 평균적으로 발화하는 음의 높이는 낮아져 있었고, 음높이의 변화는 오히려 더 증가하는 경향을 보았다. 이러한 지표들을 기반으로 기계 학습을 통해 세 군 분류를 시도하였고, 그 정확도는 MLP 모델에서 F1 값 0.589로 가장 높은 수치를 보였다. 연구 I의 알고리즘을 추가로 모집된 연구 II의 피험자에게 적용해본 결과, GNB 모델에서 F1 값이 가장 높았고 이는 데이터 불균형 정도에 따라서 적합한 모델이 달라질 수 있다는 것을 시사하였다.

이어 연구 II에서는 정상군과 우울증군을 구별할 수 있는 지표로서 불안 정도까지 보정한 뒤 4개의 지표가 선정되었다. 텍스트의 분류에서도 우울증군에서 더 다빈도로 사용되는 단어들을 살펴보았고, 상위권에는 부정적 의미를 보이는 단어들이 다수 포함되어 있었다. 목소리 지표를 기반으로 우울증군을 진단하는 알고리즘을 구축하였을 때에는 QDA 모델이 AUC 0.806의 성능을 보였다. 또한 텍스트를 기반으로 우울증군을 진단하는 알고리즘은 AUC 0.905 였으며, 이는 임상 지표를 이용하여 진단하는

알고리즘보다 더 훌륭한 성능임을 확인하였다.

연구 III에서는 연구 II의 우울증 군을 각각 BSS 17점 이상과 MINI의 자살성 모듈의 고위험군을 기준으로 자살의 고위험군을 분류하였다. 그 뒤 각각의 분류에 대한 목소리의 차이를 살펴보았고, 목소리 지표와 텍스트를 기반으로 한 자살 고위험군 진단 알고리즘을 구축하여 그 정확도까지 확인하였다. 그 결과 자살 고위험군에서의 차이를 보인 목소리 지표는 우울증을 구분하는데 도움이 되었던 지표와는 다르게, BSS, MINI 기준 모두에서 불안을 보정하기 전에는 Formant0가 통계적 유의하게 다름을 확인하였다. 목소리를 기반으로 자살 고위험군을 분류하는 진단의 알고리즘은 임상적으로 활용할 수 있는 정도의 성능을 보이지는 않았다. 한 편 텍스트를 기반으로 자살 고위험군을 분류하는 알고리즘 역시 좋은 성능을 보이지 않았지만, 텍스트 데이터를 사회인구학적 지표들과 함께 활용하는 알고리즘은 AUC 0.800~0.847로 임상적 가치를 확인할 수 있었다. 따라서 목소리와 텍스트 모두는 우울증 군을 진단함에 있어서 그 성능이 모두 우수하였으나, 자살 고위험군을 진단할 때에는 단독적으로 사용하기에는 어려움이 있었다. 이 것은 자살 고위험군과 저위험군 간의 데이터 불균형으로 인한 성능의 저하로 인한 가능성도 있으나, 텍스트 데이터의 경우 임상 데이터와 함께 활용된 앙상블 모델은 진단의 성능이 비교적 우수하였기 때문에 이에 대한 추가적인 연구가 필요할 것이다.

### 3. 연구의 종합적 강점

기존의 연구들과 가장 큰 차이점이자, 본 연구의 강점으로는 환자가 임상

현장에서 실제로 인터뷰를 한 것을 바탕으로 목소리와 텍스트 데이터를 추출하였다는 것이다. 기존의 연구들과 같이 일정의 문장을 주고 읽는 과제를 수행한다면 피험자 목소리 데이터의 추출에는 피험자 발화만을 추출하지 않아도 되기 때문에 연구의 방법은 더 간단해질 수 있다. 하지만 과제 수행시의 근긴장도로 인하여 음성 자체에 영향을 미쳤을 것이다(162). 또한 이러한 일정한 문장을 발화하거나, 자기소개를 하는 등의 짧은 과제를 통해서는 피험자가 사용하는 텍스트 데이터에 대해서는 연구를 진행하기 어려웠을 것이다. 둘째로는 텍스트 데이터가 이미 작성된 소셜 미디어 및 전자 의무 기록 등이 아닌, 인터뷰의 음성 녹음을 통해서 획득되었고, 그 획득 과정 역시 자동화되어 있는 툴킷을 활용하였다는 점이다. 따라서 텍스트 데이터를 활용한 연구의 경우 인터뷰를 녹음하기만 하면 자동으로 텍스트화하여 그 데이터를 기반으로 우울증을 진단하고 자살의 위험성을 감지할 수 있기 때문에 그 활용성이 높아지게 된다. 마지막으로 기존의 음성 및 텍스트 데이터를 통한 연구들과 다르게 MINI를 구조화된 진단 인터뷰를 활용하여 명확하게 진단을 구분하였다는 것 역시 장점이다. 기존의 연구들은 서론에서 언급하였듯, 다양한 임상 척도를 기준으로 우울증군을 구분하였다. 하지만 이러한 기준은 정확한 진단 기준이 아니기 때문에 각 연구별로의 우울증 군의 차이가 발생할 수 있다. 어떠한 연구에는 경미한 우울증까지도 포함이 되었을 가능성이 있으며, 다른 연구에서는 심각한 우울증만을 포함되었을 가능성도 있는 것이다. 따라서 본 연구에서는 전세계적으로 정신건강의학과 진단의 기준점으로 여기는 DSM-5를 기준으로 구성된 MINI를 통해 우울한 상태에 대한 정의가 확실하게 이루어졌기 때문에, 이를 바탕으로 구축된 진단의 알고리즘의 성능이 인정받을 가능성이 높다.

## 4. 연구의 한계

물론 본 연구에는 여러 가지 한계점 역시 존재한다. 첫째, 가장 중요한 한계점으로서 본 연구에 사용된 텍스트 내용은 MINI의 구조화된 질문들에 대한 대답을 통해 추출되었다는 한계점이 있다. MINI라는 구조화된 인터뷰는 특정 증상이 있으면 그 것에 대해 추가적인 질문과 대답을 요구하는 구조로 구성되어 있기 때문에, 그로 인한 텍스트 데이터의 선택 편향을 무시할 수 없다. 그럼에도 불구하고 본 연구를 통해 우울증 환자에게 다빈도로 사용되는 단어들은 같은 질문에 대해서도 다른 것으로 확인할 수 있었다.

둘째로는 피험자의 목소리와 기분 상태가 단면적으로 평가되었다는 점이다. 본 연구에서는 초기 평가시의 기분과 목소리 상태를 바탕으로 우울증 및 자살 위험성을 평가하고자 하였다. 이를 통해 우울과 자살 위험에 따라 달라지는 목소리의 지표들을 각각 발견할 수 있었으나, 이 것이 기분 상태로 인한 변화인지, 혹은 기분 상태에 대한 치료로 인한 변화인지에 대해서는 평가할 수가 없었다. 더욱 기분이 악화되거나 호전됨에 따른 개인 내에서의 목소리 변화가 더욱 진단적 중요점일 가능성도 있다.

셋째로 연구 II에서 논의되었듯이 최근에는 목소리를 분석하는 딥러닝의 방법으로서 멜 주파수 스펙트로그램 (mel frequency spectrogram) 등 지표를 계산하여 값으로 산정하지 않고, 스펙트로그램 자체를 학습시키는 연구들 역시 진행되고 있으며(127, 128), 녹음된 파일 전체를 학습시켜 임상가와 피험자의 목소리를 구별하여 피험자의 발화 내용 중에서도 중요한 부분을 스스로 학습하도록 딥 러닝을 활용할 수도 있다. 하지만 이러한

기법을 본 연구에 적용하면, 우울증상 및 자살 위험에 따른 목소리 변화에 대한 임상적 의미를 확인하기 어렵기 때문에 이러한 기법을 추가로 적용하지는 못하였다. 본 연구를 통해 국내에서도 우울증 및 자살 위험도에 따른 목소리의 변화를 확인하였기 때문에, 추후에는 딥러닝의 다양한 기법들이 적용된 보다 정밀한 진단 알고리즘의 개발이 필요하겠다.

넷째로 본 연구 III에서 자살고위험군의 정의 기준은 아직 전세계적으로 합의된 기준이 없는 상황으로서, 자살 고위험군을 주관적 보고 혹은 임상가 관찰 평가만으로 정의하는 것에는 다양한 한계들이 존재한다. 본 연구에서도 자살고위험군을 분류하는 기준으로서는 임상평가를 활용하였고, 그를 바탕으로 구축된 인공지능 알고리즘 역시 그 한계를 내포하였을 것이다. 하지만 본 연구는 객관적 진단 기준을 설립해보고자 하는 시도로서, 본 연구에서 개발된 목소리와 음성의 바탕으로 자살고위험군을 진단하는 인공지능 알고리즘의 임상적 효용성에 대한 추가적인 연구가 필요하다. 이러한 추가적인 연구들이 진행됨에 따라서, 본 연구를 통해 개발된 자살 고위험군 인공지능 알고리즘이 자살고위험군을 진단하는 새로운 기준이 될 가능성도 있겠다.

마지막으로 본 연구에서는 실제 목소리에 대한 분석만을 진행하였다는 것이다. 본 연구에서는 우선 실제 임상 현장에서의 즉각적 임상의사결정을 보조해줄 수 있는 시스템을 구축하고자 하였고, 국내에서는 우울 및 자살 위험에 따른 목소리의 특징에 대한 연구가 부족하였기 때문에 실제의 목소리를 바로 녹음하여 분석하였다. 하지만 외래 진료실 등과 같이 실제 목소리를 바탕으로 임상의사결정을 내려야 하는 진료 환경도 있지만, 서론에 언급하였듯 자살 예방 콜센터 등과 같이 전화상의 목소리를 기반으로

임상의사결정을 내려야 하는 순간들도 존재한다. 전화 상의 목소리는 실제의 목소리의 변질이 생기기 때문에, 본 연구에서 추출된 지표들을 바탕으로 한 알고리즘을 전화상의 목소리에 적용하기 어렵다는 한계가 있다.

## 5. 추후 연구에 대한 제언

본 연구는 국내에서는 최초로 인터뷰를 통해 획득할 수 있는 음성과 텍스트를 기반으로 즉각적으로 우울증과 자살 위험을 진단하는 알고리즘을 구축하였으나, 살펴본 바와 같이 다양한 한계를 가지고 있다. 따라서 본 연구를 바탕으로 향후 필요한 연구에 대한 제언을 남기고자 한다. 첫째로 자유 인터뷰에서 사용된 텍스트 데이터를 기반으로 한 추가 연구가 필요하겠다. 임상의와 환자의 자유로운 인터뷰를 기반으로 데이터를 수집하면 환자가 사용하는 단어의 범위가 더 넓어지고 텍스트를 통한 진단의 정확도가 낮아질 가능성도 있기 때문에 보다 광범위한 데이터를 기반으로 한 연구가 필요하다.

둘째로 피험자 한 사람 안에서도 종적인 추적관찰을 통해 기분 상태 변화에 따른 음성과 텍스트의 변화에 대한 연구가 필요하다. 임상 상황에서 첫 면담 시 우울증의 유무와 자살의 위험성을 평가해야 하는 상황도 있지만, 외래 진료 등과 같은 환경에서는 지속적으로 변화하는 기분 상태와 자살 위험성에 대해서 판단을 내려야 하는 상황도 존재한다. 따라서 한 사람 안에서도 기분 상태에 따라서 사용하는 텍스트가 달라지는지, 목소리의 지표들이 변화하는지 등을 살펴보고, 우울한 상태로 변화할 때 혹은 자살 위험성이 증가할 때의 목소리와 텍스트의 변화를 함께 확인한다면 진단

알고리즘의 정밀화가 가능할 것이다. 또한 음성의 다양한 특징 속에는 남성인지 여성인지에 따라서 차이가 발생하는 pitch의 평균값 등은 피험자의 특성을 대표하는 인자(trait marker)도 있으며, 발화중의 공백 시간 등과 같이 심리 상태를 더욱 정밀하게 반영하는 상태를 대표하는 인자(state marker) 등이 모두 공존하고 있을 것이다. 기존의 다양한 연구들을 통해 우울한 상태에 따라서 변화하는 F0 등을 포함한 formant 인자들, 발화중의 공백 비율 등이 밝혀져 있으나, 다양한 음성의 지표들 중 어떠한 것이 우울증상 및 자살 위험 등과 더욱 밀접한 관련이 있는 상태표지자(state marker)로서의 기능을 할 지에 대한 종적인 관찰 연구를 통해 밝혀질 필요가 있다.

셋째로 향후에는 더욱 대규모의 인구를 대상으로, 다양한 인터뷰 상황에 대한 빅데이터의 수집이 필요하겠다. 또한 본 연구는 3차 의료기관에서 피험자 모집이 이루어졌으나, 일차 의료 현장과 자살 예방 콜센터 등과 같이 다양한 환경에서의 데이터 수집 역시 필요할 것이다. 또한 본 연구의 피험자는 대부분 수도권의 억양을 보였기 때문에, 각 지역별 방언에 대한 고려가 필요하지 않았다. 따라서 본 연구의 결과가 국내의 각 지역에서도 재현되는지, 다양한 지역에서도 연구도 필요하겠다.

넷째로 최근 원격의료가 각광을 받고 있고, 실제 임상에서 자살 예방 콜센터 등과 같은 곳에서는 전화상의 목소리를 바탕으로 판단은 해야 하는 상황들이 다수 있다. 목소리를 바로 녹음을 하는 경우와, 전화 등의 전자 기기를 거친 목소리는 다양한 지표들의 변화가 발생할 수 있다. 본 연구를 통해 실제의 목소리 변화를 확인하였기 때문에, 향후에는 전화 목소리 등의 다양한 변화에 대한 추가적인 연구들 역시 진행되어야 할 것이다.

마지막으로 더욱이 본 연구와 같이 인터뷰를 통한 음성의 획득 방법은 피험자의 발화 과정을 추출하는 후처리 과정의 어려움이 있기 때문에, 발화 구분을 정확하게 시행하는 기술이 개발되기 전까지는 고정된 과제를 통한 음성을 통한 진단 알고리즘 역시 구축을 하여, 각각의 효용성을 검토하고 활용 가능성을 비교하는 연구 또한 필요할 것이다. 화자 구분의 기술이 발전하고, 음성의 빅데이터 분석을 통해서 인공지능 진단 알고리즘이 더욱 고도화되고 정밀화된다면, 그 후 실제 임상현장에 적용을 하여 임상가의 진단에 도움이 되는 지에 대한 추가적인 평가가 필요할 것이다.

6. 본 연구의 임상적 활용 방안

마지막으로 본 연구를 통해 획득한 음성과 텍스트를 이용한 인공지능 기반의 진단 알고리즘의 임상적 활용계획을 제시하고자 한다.

첫째로 연구 결과로 개발된 음성을 기반으로 우울증을 진단하는 알고리즘을 활용하여 임상에서 사용 가능한 어플리케이션을 준비하고자 한다. 본 연구에서는 인터뷰시의 발화를 활용하였으나, 추후 연구에서 제언하였듯 현재는 화자구분 알고리즘의 정확도가 우수하지는 않기 때문에 전체 인터뷰를 녹음하여 결과를 표현하기에 한계가 있다. 또한 본 연구의 큰 한계 중 하나인, MINI를 통해 획득된 텍스트 데이터라는 한계를 극복하기 위하여는 향후 자유 인터뷰의 데이터 취득이 필요함에 대해 언급하였다. 또한 전화상의 목소리에 대한 연구들이 진행된다면, 실제 목소리와 전화 상의 목소리의 차이에 대한 변형을 거치면 현 알고리즘을 전화 상의 목소리에 적용할 가능성도 있기 때문에 향후의 다양한 활용가능성을 고려하여 본

알고리즘을 적용시킨 어플리케이션은 가능한 경량화하여 추후 다양한 알고리즘을 접목시킬 수 있도록 계획하였다. 더불어 꼭 장시간의 인터뷰가 아니더라도 간단한 과제 등을 통해 획득되는 짧은 목소리에 대한 데이터를 획득하고 이를 기반으로 정확한 진단을 내릴 수 있는 것 역시 임상적 활용 가능성을 높일 수 있다. 이러한 고려사항을 바탕으로 2021년 11월 해당 어플리케이션의 초기 버전이 개발하였다. 현재 어플리케이션은 피험자의 나이, 성별을 입력하고, 문장 읽기 과제를 통해 3초 가량의 짧은 목소리를 녹음하여, 녹음된 목소리를 기구축된 알고리즘에 적용한 결과를 확인할 수 있도록 제작하였다.

제작된 어플리케이션을 통한 실질적 임상에서의 진단보조시스템으로써 그 효용성을 확인함과 동시에, 어플리케이션의 녹음 기능을 통해 짧은 과제를 통한 대규모의 음성 데이터를 함께 수집하고자 한다. 이렇게 획득되는 대규모의 데이터는 우리가 제시한 필요한 후행 연구의 기본 데이터로서 활용될 것이다. 또한 어플리케이션의 도움을 받은 진단과 임상의의 진단을 비교 분석을 하는 무작위 대조 시험 등을 고려해볼 수도 있을 것이다. 이는 정신건강의학과 분야에서는 최초의 디지털 진단 기법으로서 가치가 있다.

2011년 전국 정신질환실태조사에 따르면, 기분 장애를 경험하는 국민은 전체 국민의 3.6%에 달하였으나, 그 중에서 정신건강의학과 전문의로부터 치료를 받는 경우는 28.5%에 불과하였다. 전문의를 방문하지 않은 이유로는 74.3%에서는 스스로 해결할 수 있을 것이라고 생각하였고, 67.8%에서는 나는 정신질환이 없다 생각했기 때문으로 보고하였다. 즉 기분 장애를 경험하고 전문의의 도움을 받아야 하는 67.8%는 스스로 질환이 있는지도 잘 모르고 있다는 의미이다. 이러한 문제를 해결하기 위하여는 질환의 상태에

대한 객관적이고 정확한 진단 방법이 구축되는 것이 필요하다. 특히 스스로의 기분 상태에 대한 인식이 부족하고 그에 대한 보고를 축소 보고하는 경향이 있기는 아시아문화권에서는 객관적인 진단 알고리즘은 이러한 한계를 극복할 수 있는 방법으로서, 조기 진단과 개입을 이룰 수 있는 방법이 될 것이다.

하지만 현재 정신건강의학과에서의 디지털 헬스 케어 서비스는 대부분 디지털 치료제에 집중되어 있다. 특히 가상 현실 기술과 어플리케이션 기반의 마음 챙김 등의 인지행동 치료제 등이 많이 개발되어 있다(163,164). 하지만 이러한 디지털 치료제가 전달되는 대상은 여전히 임상가와의 면담과 평가를 통해 진단이 되어져야만 이러한 디지털 치료제에 접근할 수 있다는 한계를 극복할 수 없다. 따라서 향후에는 디지털 치료제 뿐 아니라, 디지털 진단 기법에 대해서도 더욱 많은 관심을 가지고 다양한 연구들이 진행되는 것이 필요하겠다. 그런 의미에서 본 연구를 통해 인터뷰의 목소리와 텍스트 데이터를 통한 우울증의 진단 성능에 대해서는 그 정확성을 확인할 수 있었고, 자살 위험성을 진단함에 있어서는 임상 지표와 통합하였을 때의 활용 가능성을 확인하였기 때문에, 향후 더욱 다양한 디지털 진단 기법 연구의 기초가 될 수 있기를 기대한다.

표 1-1. Comparison of demographics according to depressive episodes

| | | ND ‡ | mDE | MDE | p Value | Post Hoc Test |
|---|---|---|---|---|---|---|
| N | | 33 | 26 | 34 | | |
| sex | M | 8 (24.2%) | 8 (30.8%) | 7 (20.6%) | 0.689 | |
| | F | 25 (75.8%) | 18 (69.2%) | 27 (79.4%) | | |
| Age * | Mean ± SD | 28.12 ± 4.827 | 34.58 ± 11.497 | 29.68 ± 9.914 | 0.022 | 1 < 2*, 2 > 3* |
| SES | Very low | 0 (0%) | 2 (7.7%) | 1 (2.9%) | 0.397 | |
| | Low | 10 (30.3%) | 5 (19.2%) | 7 (20.6%) | | |
| | Middle | 18 (54.5%) | 15 (57.7%) | 18 (52.9%) | | |
| | High | 5 (15.2%) | 3 (11.5%) | 4 (11.8%) | | |
| | Very high | 0 (0%) | 1 (3.8%) | 4 (11.8%) | | |
| BMI *** | Mean ± SD | 21.356 ± 1.861 | 23.470 ± 4.575 | 25.620 ± 5.396 | <0.001 | 1 < 3 ** |
| drugs taken for non-psychiatric disorders * | Yes | 2 (6.1%) | 8 (30.8%) | 4 (11.8%) | 0.025 | 1 ≠ 2 * |
| | No | 31 (93.9%) | 18 (69.2%) | 30 (88.2%) | | |
| Antipsychotics *** | Yes | 0 (0%) | 23 (88.5%) | 29 (85.3%) | <0.001 | 1 ≠ 2 ***, 1 ≠ 3 *** |
| | No | 33 (100%) | 3 (11.5%) | 5 (14.7%) | | |
| Antipsychotics dose | Mean ± SD | | 5.142 ± 4.589 | 8.254 ± 7.893 | 0.100 | |
| diagnosis by MINI *** | No psychiatric disorder | 26 (78.8%) | 0 (0%) | 0 (0%) | 0.000 | |
| | Major depressive disorder | 5 (15.2%) | 9 (34.6%) | 1 (2.9%) | | |
| | Bipolar disorder | 2 (6.1%) | 17 (65.4%) | 33 (97.1%) | | |

| | | | | |
|---|---|---|---|---|
| | Anxiety disorders † | 2 (6.1%) | 2 (7.7%) | 13 (38.2%) |
| | Obsessive compulsive disorder | 1 (3.0%) | 1 (3.8%) | 3 (8.8%) |
| | Alcohol use disorder | 2 (6.1%) | 1 (3.8%) | 3 (8.8%) |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001.

† Combined all types of anxiety disorders, including panic disorder, generalized anxiety disorder, and social anxiety disorder.

‡ Currently, there are no symptoms of depression, but past major episodes of depression are included.

표 1-2. Clinical characteristics by depressive episode

| | | ND | mDE | MDE | p Value | Post Hoc Test |
|---|---|---|---|---|---|---|
| N | | 33 | 26 | 34 | | |
| HRDS *** | mean | 3.879 | 13.346 | 18.706 | <0.001 | 1 < 2 < 3 *** |
| | SD | 2.902 | 4.127 | 4.414 | | |
| PHQ *** | mean | 1.576 | 11.615 | 16.294 | <0.001 | 1 < 2, 3 ***, 2 < 3* |
| | SD | 2.332 | 5.947 | 6.279 | | |
| BAI *** | mean | 1.394 | 20.385 | 25.206 | <0.001 | 1 < 2 ***,<br>1 < 3 *** |
| | SD | 2.609 | 16.346 | 17.562 | | |
| BIS | mean | 60.909 | 65.615 | 62.088 | 0.127 | |
| | SD | 5.598 | 8.750 | 8.155 | | |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001.

표 1-3. Difference of voice features by depressive episode

| | | ND | mDE | MDE | p Value | B–H Test | Adjusted p Value # |
|---|---|---|---|---|---|---|---|
| N | | 33 | 26 | 34 | | | |
| log_glottal_OP* | mean | 0.890 | 0.846 | 0.925 | 0.037 | | 0.051 |
| | SD | 0.105 | 0.096 | 0.137 | | | |
| log_glottal_CP | mean | 0.710 | 0.692 | 0.759 | 0.094 | | 0.097 |
| | SD | 0.124 | 0.110 | 0.132 | | | |
| log_glottal_C | mean | −0.675 | −0.694 | -0.627 | 0.094 | | 0.097 |
| | SD | 0.124 | 0.110 | 0.132 | | | |
| log_spectral_time | mean | 2.345 | 2.583 | 2.457 | 0.191 | | 0.461 |
| | SD | 0.463 | 0.529 | 0.497 | | | |
| spectral_centroid *** | mean | 7.471 | 7.375 | 7.398 | <0.001 | 1 > 2 ***, 1 > 3** | 0.008 ‡‡ |
| | SD | 0.058 | 0.075 | 0.105 | | | |
| spectral_bandwidth | mean | 7.444 | 7.422 | 7.430 | 0.343 | | 0.968 |
| | SD | 0.050 | 0.057 | 0.069 | | | |
| spectral_roll-off *** | mean | 8.118 | 7.994 | 8.026 | <0.001 | 1 > 2 ***, 1 > 3** | 0.012 ‡ |
| | SD | 0.082 | 0.110 | 0.132 | | | |
| spectral_rmse | mean | 4.358 | 4.058 | 4.329 | 0.180 | | 0.468 |
| | SD | 0.540 | 0.668 | 0.760 | | | |
| log_spectral_tempo | mean | 4.771 | 4.779 | 4.772 | 0.093 | | 0.327 |
| | SD | 0.012 | 0.019 | 0.012 | | | |
| formant1 | mean | 6.230 | 6.239 | 6.218 | 0.552 | | 0.553 |
| | SD | 0.062 | 0.069 | 0.082 | | | |
| formant2 | mean | 7.374 | 7.349 | 7.349 | 0.490 | | 0.304 |
| | SD | 0.085 | 0.095 | 0.101 | | | |
| formant3 | mean | 8.043 | 8.022 | 8.026 | 0.526 | | 0.919 |
| | SD | 0.080 | 0.068 | 0.079 | | | |
| formant_BW1 | mean | 42.083 | 39.681 | 44.119 | 0.331 | | 0.159 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | SD | 10.233 | 8.399 | 14.107 | | | |
| formant_BW2 * | mean | 180.213 | 201.080 | 198.655 | 0.200 | | 0.040 ‡ |
| | SD | 53.886 | 50.843 | 45.060 | | | |
| sq_formant_BW3 | mean | 50622.267 | 52003.924 | 44566.728 | 0.133 | | 0.094 |
| | SD | 13671.136 | 18982.013 | 14031.728 | | | |
| sq_mean_pitch ** | mean | 87561.420 | 73835.557 | 81997.509 | 0.002 | 1 > 2 **, | 0.027 ‡ |
| | SD | 12409.867 | 15014.241 | 15820.365 | | | |
| stdev_pitch ** | mean | 0.287 | 0.344 | 0.300 | 0.003 | 1 < 2**, 2 > 3* | 0.020 ‡ |
| | SD | 0.057 | 0.065 | 0.067 | | | |
| mean_magnitude ** | mean | 69.894 | 61.002 | 65.060 | 0.009 | 1 > 2* | 0.059 |
| | SD | 11.454 | 11.292 | 9.902 | | | |
| sq_stdev_magnitude | mean | 0.787 | 0.748 | 0.852 | 0.140 | | 0.237 |
| | SD | 0.146 | 0.252 | 0.215 | | | |
| ZCR *** | mean | 0.055 | 0.044 | 0.047 | <0.001 | 1 > 2***, 1 < 3*** | <0.001 ‡‡‡ |
| | SD | 0.007 | 0.006 | 0.010 | | | |
| voice portion ** | mean | 0.665 | 0.695 | 0.681 | 0.001 | 1 < 2** | 0.020 ‡ |
| | SD | 0.023 | 0.031 | 0.033 | | | |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001.

# Adjusted for BMI, age, non-psychiatric medication.

‡ adjusted p value < 0.05, ‡‡ adjusted p value < 0.01, ‡‡‡ adjusted p value < 0.001

표 1-4. Jonckheere-Terpstra test result of voice features by depressive episode

| | | ND | mDE | MDE | J–T Test |
|---|---|---|---|---|---|
| N | | 33 | 26 | 34 | |
| log_glottal_OP* | mean | 0.890 | 0.846 | 0.925 | 0.381 |
| | SD | 0.105 | 0.096 | 0.137 | |
| log_glottal_CP | mean | 0.710 | 0.692 | 0.759 | 0.069 |
| | SD | 0.124 | 0.110 | 0.132 | |
| log_glottal_C | mean | −0.675 | −0.694 | -0.627 | 0.068 |
| | SD | 0.124 | 0.110 | 0.132 | |
| log_spectral_time | mean | 2.345 | 2.583 | 2.457 | 0.343 |
| | SD | 0.463 | 0.529 | 0.497 | |
| spectral_centroid *** | mean | 7.471 | 7.375 | 7.398 | <0.001 |
| | SD | 0.058 | 0.075 | 0.105 | |
| spectral_bandwidth | mean | 7.444 | 7.422 | 7.430 | 0.384 |
| | SD | 0.050 | 0.057 | 0.069 | |
| spectral_roll-off ** | mean | 8.118 | 7.994 | 8.026 | 0.001 |
| | SD | 0.082 | 0.110 | 0.132 | |
| spectral_rmse | mean | 4.358 | 4.058 | 4.329 | 0.794 |
| | SD | 0.540 | 0.668 | 0.760 | |
| log_spectral_tempo | mean | 4.771 | 4.779 | 4.772 | 0.286 |
| | SD | 0.012 | 0.019 | 0.012 | |
| formant1 | mean | 6.230 | 6.239 | 6.218 | 0.562 |
| | SD | 0.062 | 0.069 | 0.082 | |
| formant2 | mean | 7.374 | 7.349 | 7.349 | 0.221 |
| | SD | 0.085 | 0.095 | 0.101 | |
| formant3 | mean | 8.043 | 8.022 | 8.026 | 0.621 |
| | SD | 0.080 | 0.068 | 0.079 | |
| formant_BW1 | mean | 42.083 | 39.681 | 44.119 | 0.729 |
| | SD | 10.233 | 8.399 | 14.107 | |

| | | | | | |
|---|---|---|---|---|---|
| formant_BW2 | mean | 180.213 | 201.080 | 198.655 | 0.094 |
| | SD | 53.886 | 50.843 | 45.060 | |
| sq_formant_BW3 | mean | 50622.267 | 52003.924 | 44566.728 | 0.292 |
| | SD | 13671.136 | 18982.013 | 14031.728 | |
| sq_mean_pitch | mean | 87561.420 | 73835.557 | 81997.509 | 0.149 |
| | SD | 12409.867 | 15014.241 | 15820.365 | |
| stdev_pitch | mean | 0.287 | 0.344 | 0.300 | 0.520 |
| | SD | 0.057 | 0.065 | 0.067 | |
| mean_magnitude | mean | 69.894 | 61.002 | 65.060 | 0.110 |
| | SD | 11.454 | 11.292 | 9.902 | |
| sq_stdev_magnitude* | mean | 0.787 | 0.748 | 0.852 | 0.045 |
| | SD | 0.146 | 0.252 | 0.215 | |
| ZCR *** | mean | 0.055 | 0.044 | 0.047 | < 0.001 |
| | SD | 0.007 | 0.006 | 0.010 | |
| voice portion * | mean | 0.665 | 0.695 | 0.681 | 0.021 |
| | SD | 0.023 | 0.031 | 0.033 | |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001.

표 1-5. Machine learning model performance through voice features

| | | 7:3 | | | | 8:2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model | LR | GNB | SVM | MLP | LR | GNB | SVM | MLP |
| Accuracy Mean (max) | augmented | 45.2 (53.6) | 48.8 (57.1) | 46.4 (50) | 51.2 (53.6) | 43.2 (57.9) | 43.2 (57.9) | 45.3 (52.6) | 60 (68.4) |
| | selected-augmented | 47.6 (57.1) | 48.8 (57.1) | 40.5 (53.6) | 51.2 (57.1) | 43.2 (63.2) | 43.2 (57.9) | 35.8 (42.1) | 51.6 (57.9) |
| AUC Mean (max) | augmented | 63.4 (68) | 64.7 (70.2) | 60.3 (61) | 59.7 (65.1) | 64.5 (72.1) | 64.5 (73.9) | 61.1 (70.6) | 65.9 (72.1) |
| | selected-augmented | 63.6 (75.7) | 63.6 (72) | 58.8 (66.5) | 62.9 (70.7) | 62 (74.6) | 60.3 (67.5) | 56.8 (64.9) | 62.6 (69.9) |
| Precision (sensitivity) Mean (max) | augmented | 45.1 (55.6) | 49.1 (58) | 47.3 (55) | 51.4 (71.5) | 46.6 (70.9) | 41.6 (60) | 48.7 (68.7) | 65.6 (76.7) |
| | selected-augmented | 57.2 (64.6) | 54 (66.7) | 36.3 (56.3) | 60 (62.6) | 44 (61.3) | 42.3 (58.9) | 34.6 (42.1) | 62.6 (72.2) |
| Recall (specificity) Mean (max) | augmented | 44.8 (55.1) | 49.6 (60.2) | 45.1 (48.6) | 48.5 (55.8) | 43.3 (61.7) | 42.6 (64.2) | 45.6 (52.5) | 66.2 (69.7) |
| | selected-augmented | 46.1 (55.8) | 48.1 (56.7) | 48.4 (64.6) | 54.8 (62.5) | 43.8 (64.6) | 43 (58.1) | 37.7 (44.2) | 93.3 (100) |
| F1 Mean (max) | augmented | 43.1 (50.8) | 46.4 (55.5) | 43.5 (49.4) | 44.3 (52.3) | 42.1 (58.2) | 39.6 (55.6) | 43.4 (49.2) | 58.9 (69.7) |
| | selected-augmented | 45.4 (56.8) | 47.8 (57.5) | 40.7 (64.6) | 49.3 (57.9) | 41 (62.1) | 42.1 (58.3) | 33.1 (35.7) | 58.7 (71.4) |

Bold: mean or max is greater than or equal to 60

표 1-6. Comparison of demographics between healthy control and major depressive disorder and bipolar disorder

| | | Healthy Control | Major depressive disorder | Bipolar disorder | P value | Post hoc |
|---|---|---|---|---|---|---|
| N | | 26 | 15 | 52 | | |
| sex | M | 8 (30.8%) | 3 (20.0%) | 12 (23.1%) | 0.682 | |
| | F | 18 (69.2%) | 12 (80.0%) | 40 (76.9%) | | |
| age | Mean ± SD | 28.54 ± 5.140 | 31.13 ± 9.164 | 31.29 ± 10.811 | 0.694 | |
| SES | 1 | 0.000 | 0.000 | 3 (5.8%) | 0.380 | |
| | 2 | 8 (30.8%) | 5 (33.3%) | 9 (40.9%) | | |
| | 3 | 14 (53.8%) | 8 (53.3%) | 29 (55.8%) | | |
| | 4 | 4 (15.4%) | 2 (13.3%) | 6 (11.5%) | | |
| | 5 | 0.000 | 0.000 | 5 (9.6%) | | |
| Medication for non-psychiatric disorders | Yes | 2 (7.7%) | 3 (20.0%) | 9 (17.3%) | 0.450 | |
| | no | 24 (92.3%) | 12 (80.0%) | 43 (82.7%) | | |
| BMI | Mean ± SD | 21.291 ± 1.735 | 22.230 ± 3.046 | 24.981 ± 5.294 | 0.490 | |
| antipsychotics dose† | N | 0 | 8 | 44 | 0.135 | |
| | Mean ± SD | 0 | 3.577 ± 1.605 | 7.478 ± 7.171 | | |
| Previous psychiatric illness** | Yes | 0 (0%) | 10 (66.7%) | 48 (92.3%) | 0.001 | 2≠3* |
| | no | 26 (100%) | 5 (33.3%) | 4 (7.7%) | | |

† Two groups of major depressive disorder and bipolar disorder were compared by the Mann-Whitney method

* p value < 0.05, ** p value < 0.01, *** p value < 0.001

표 1-7. Comparison of clinical characteristics between healthy control and major depressive disorder and bipolar disorder

| | | Healthy control | Major depressive disorder | Bipolar disorder | P value | post hoc |
|---|---|---|---|---|---|---|
| N | | 26 | 15 | 52 | | |
| HDRS*** | Mean | 3.308 | 10.400 | 16.712 | <0.001 | 1<2,3***, 2<3** |
| | SD | 2.205 | 6.468 | 4.904 | | |
| PHQ*** | Mean | 1.154 | 6.800 | 14.923 | <0.001 | 1<2<3*** |
| | SD | 2.073 | 5.722 | 6.419 | | |
| BAI*** | Mean | 0.500 | 12.000 | 23.846 | <0.001 | 1<2,3***, 2<3* |
| | SD | 0.949 | 14.962 | 16.950 | | |
| BIS | Mean | 60.500 | 62.267 | 63.846 | 0.191 | |
| | SD | 4.693 | 8.868 | 8.405 | | |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001

표 1-8. Voice and speech features between healthy control and major depressive disorder and bipolar disorder

| | | Healthy control | Major depressive disorder | Bipolar disorder | p value | Post hoc |
|---|---|---|---|---|---|---|
| N | | 26 | 15 | 52 | | |
| glottal_OP | Mean | 2.477 | 2.328 | 2.480 | 0.211 | |
| | SD | 0.297 | 0.165 | 0.344 | | |
| glottal_CP | Mean | 2.069 | 2.000 | 2.102 | 0.427 | |
| | SD | 0.282 | 0.214 | 0.279 | | |
| glottal_C | Mean | 0.518 | 0.500 | 0.526 | 0.427 | |
| | SD | 0.071 | 0.054 | 0.070 | | |
| spectral_time | Mean | 12.246 | 17.837 | 12.956 | 0.202 | |
| | SD | 11.857 | 14.269 | 7.290 | | |
| spectral_centroid*** | Mean | 7.480 | 7.409 | 7.389 | <0.001 | 1>2***, 1>3* |
| | SD | 0.056 | 0.083 | 0.093 | | |
| spectral_bandwidth | Mean | 7.450 | 7.430 | 7.425 | 0.075 | |
| | SD | 0.051 | 0.054 | 0.064 | | |
| spectral_rolloff*** | Mean | 8.131 | 8.034 | 8.013 | <0.001 | 1>2***, 1>3* |
| | SD | 0.075 | 0.116 | 0.124 | | |
| spectral_rmse | Mean | 4.349 | 4.000 | 4.297 | 0.228 | |
| | SD | 0.589 | 0.574 | 0.721 | | |
| spectral_tempo | Mean | 118.092 | 119.172 | 118.259 | 0.128 | |
| | SD | 1.560 | 2.771 | 1.342 | | |
| formant1 | Mean | 6.226 | 6.231 | 6.229 | 0.791 | |
| | SD | 0.056 | 0.074 | 0.078 | | |
| formant2 | Mean | 7.382 | 7.353 | 7.347 | 0.562 | |
| | SD | 0.084 | 0.110 | 0.093 | | |
| formant3 | Mean | 8.048 | 8.036 | 8.022 | 0.175 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | SD | 0.079 | 0.073 | 0.076 | | |
| formant_BW1 | Mean | 42.129 | 41.821 | 42.266 | 0.926 | |
| | SD | 10.774 | 10.038 | 12.253 | | |
| formant_BW2 | Mean | 179.890 | 194.088 | 198.864 | 0.290 | |
| | SD | 53.065 | 64.302 | 43.874 | | |
| formant_BW3 | Mean | 221.704 | 228.024 | 212.505 | 0.499 | |
| | SD | 33.144 | 34.371 | 41.485 | | |
| mean_pitch** | Mean | 299.709 | 272.529 | 280.279 | 0.002 | 1>2** |
| | SD | 19.406 | 34.976 | 26.624 | | |
| stdev_pitch** | Mean | 0.275 | 0.340 | 0.314 | 0.004 | 1<2* |
| | SD | 0.055 | 0.069 | 0.066 | | |
| mean_magnitude | Mean | 70.441 | 63.680 | 63.806 | 0.044 | 1>2** |
| | SD | 12.109 | 11.169 | 10.427 | | |
| stdev_magnitude | Mean | 0.882 | 0.870 | 0.891 | 0.925 | |
| | SD | 0.086 | 0.172 | 0.135 | | |
| ZCR*** | Mean | 0.056 | 0.049 | 0.046 | <0.001 | 1>2***, 1>3* |
| | SD | 0.007 | 0.010 | 0.009 | | |
| voice portion** | Mean | 0.665 | 0.682 | 0.686 | 0.002 | 1<2*** |
| | SD | 0.023 | 0.032 | 0.032 | | |

* p value < 0.05, ** p value < 0.01, *** p value < 0.001

표 2-1 Descriptive statistics of voice and speech features

| | Minimum | Maximum | Mean | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| Time | 5.155 | 110.115 | 14.885 | 12.640 | 4.088 | 22.625 |
| Time_log | 0.712 | 2.042 | 1.093 | 0.236 | 1.162 | 1.895 |
| Spectral_centroid | 7.119 | 7.603 | 7.381 | 0.110 | -0.212 | -0.650 |
| Spectral_band | 7.144 | 7.576 | 7.403 | 0.082 | -0.478 | -0.146 |
| Spectral_roll | 7.607 | 8.278 | 7.997 | 0.150 | -0.335 | -0.615 |
| Spectral_rmse | 2.805 | 5.678 | 4.027 | 0.686 | -0.001 | -1.104 |
| Formant0 | 5.809 | 6.417 | 6.230 | 0.085 | -1.094 | 3.651 |
| Formant1 | 6.967 | 7.633 | 7.327 | 0.118 | -0.337 | 0.601 |
| Formant2 | 7.789 | 8.329 | 8.018 | 0.082 | 0.196 | 0.673 |
| Formant_bw0 | 16.412 | 78.831 | 41.069 | 11.653 | 0.569 | 0.044 |
| Formant_bw1 | 105.766 | 323.104 | 202.755 | 47.935 | 0.095 | -0.705 |
| Formant_bw2 | 81.274 | 305.971 | 219.537 | 38.583 | -0.682 | 0.585 |
| mean_p | 177.710 | 340.239 | 278.612 | 29.312 | -0.487 | 0.199 |
| error_p | 0.145 | 0.488 | 0.327 | 0.070 | 0.155 | -0.510 |
| change_p | 7.877 | 216.424 | 38.863 | 30.063 | 3.229 | 12.821 |
| change_p_log | 0.896 | 2.335 | 1.513 | 0.239 | 0.804 | 1.309 |
| mean_m | 39.103 | 94.833 | 63.709 | 11.136 | 0.494 | -0.135 |
| error_m | 0.305 | 1.123 | 0.859 | 0.139 | -0.893 | 1.065 |
| change_m | 10.630 | 475.788 | 73.262 | 64.099 | 3.407 | 14.351 |
| change_m_log | 1.027 | 2.677 | 1.772 | 0.262 | 0.790 | 1.355 |
| ZCR | 0.030 | 0.073 | 0.048 | 0.009 | 0.306 | -0.502 |
| Pause portion | 0.611 | 0.743 | 0.681 | 0.028 | -0.097 | -0.470 |
| Total_count | 3.000 | 190.000 | 41.601 | 26.913 | 2.762 | 11.402 |
| Total_count_log | 0.477 | 2.279 | 1.548 | 0.253 | -0.441 | 2.325 |

표 2-2. The results of applying the diagnostic algorithm from Study I to the newly recruited test set from Study II

| | | |
|---|---|---|
| Logistic Regression | F1 score | 0.292 |
| | Accuracy | 0.682 |
| Gaussian Naive Bayes | F1 score | **0.406** |
| | Accuracy | 0.671 |
| Support Vector Machine | F1 score | 0.275 |
| | Accuracy | **0.706** |
| Multi-layer Perceptron | F1 score | 0.335 |
| | Accuracy | 0.671 |

Bold: Best value among four models

표 2-3. Demographics comparison between healthy control group and current depression group

| | | HC | CD | P value |
|---|---|---|---|---|
| N | | 83 | 83 | |
| Age** | mean | 37.072 | 30.916 | <0.001 |
| | SD | 11.421 | 10.820 | |
| Sex | F | 69 (83.1%) | 64 (77.1%) | 0.331 |
| | M | 14 (16.9%) | 19 (22.9%) | |
| SES * | Low | 30 (36.1%) | 19 (22.9%) | 0.015 |
| | Middle | 45 (54.2%) | 43 (51.8%) | |
| | High | 8 (9.6%) | 21 (25.3%) | |
| BMI*** | mean | 22.364 | 24.809 | <0.001 |
| | SD | 2.902 | 4.810 | |
| Non-psychiatric medication** | Yes | 2 (2.4%) | 15 (18.1%) | 0.001 |
| | no | 81 (97.6%) | 68 (81.9%) | |

* p<0.05, ** p<0.01, *** p<0.001

표 2-4. Clinical characteristics between HC and CD

| | | HC | CD | P value |
|---|---|---|---|---|
| N | | 83 | 83 | |
| PHQ*** | M | 0.892 | 15.205 | <0.001 |
| | SD | 1.440 | 6.636 | |
| HDRS*** | M | 4.000 | 16.638 | <0.001 |
| | SD | 3.008 | 4.805 | |
| BAI*** | M | 1.542 | 24.217 | <0.001 |
| | SD | 2.923 | 16.659 | |
| BIS | M | 61.470 | 63.530 | 0.058 |
| | SD | 5.840 | 7.922 | |
| BSS*** | M | 1.108 | 18.145 | <0.001 |
| | SD | 1.815 | 9.531 | |

* p<0.05, ** p<0.01, *** p<0.001

표 2-5. Voice and speech features comparison between healthy control and current depression

| | | HC | CD | P value† | Adjusted age, SES, BMI, suffering‡ | Adjusted age, SES, BMI, suffering and BAI‡ |
|---|---|---|---|---|---|---|
| N | | 83 | 83 | | | |
| Time_log | M | 1.070 | 1.113 | 0.243 | 0.045 | 0.943 |
| | SD | 0.249 | 0.224 | | | |
| Spectral_centroid* | M | 7.383 | 7.381 | 0.917 | 0.053 | 0.045 |
| | SD | 0.115 | 0.105 | | | |
| Spectral_band | M | 7.384 | 7.424 | 0.002 | 0.019 | 0.498 |
| | SD | 0.090 | 0.069 | | | |
| Spectral_roll | M | 7.997 | 8.000 | 0.881 | 0.087 | 0.055 |
| | SD | 0.155 | 0.144 | | | |
| Spectral_rmse | M | 3.952 | 4.101 | 0.165 | 0.611 | 0.411 |
| | SD | 0.586 | 0.776 | | | |
| Formant0 | M | 6.237 | 6.221 | 0.208 | 0.935 | 0.432 |
| | SD | 0.093 | 0.076 | | | |
| Formant1 | M | 7.312 | 7.343 | 0.093 | 0.383 | 0.507 |
| | SD | 0.129 | 0.107 | | | |
| Formant2 | M | 8.010 | 8.028 | 0.161 | 0.568 | 0.689 |
| | SD | 0.088 | 0.076 | | | |
| Formant _bw0 | M | 39.728 | 42.628 | 0.110 | 0.278 | 0.906 |
| | SD | 10.152 | 12.929 | | | |
| Formant _bw1 | M | 203.690 | 202.296 | 0.853 | 0.822 | 0.395 |
| | SD | 48.165 | 48.438 | | | |
| Formant _bw2 | M | 222.962 | 216.600 | 0.284 | 0.292 | 0.348 |
| | SD | 34.137 | 41.780 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean_p | M | 279.720 | 277.538 | 0.631 | 0.017 | 0.060 |
| | SD | 28.696 | 29.659 | | | |
| Error_p | M | 0.329 | 0.324 | 0.637 | 0.210 | 0.422 |
| | SD | 0.070 | 0.070 | | | |
| Change_p_log | M | 1.501 | 1.523 | 0.564 | 0.049 | 0.814 |
| | SD | 0.254 | 0.227 | | | |
| Mean_m | M | 64.293 | 63.055 | 0.477 | 0.126 | 0.213 |
| | SD | 11.340 | 11.043 | | | |
| Error_m | M | 0.850 | 0.872 | 0.300 | 0.306 | 0.433 |
| | SD | 0.122 | 0.152 | | | |
| Change_m_log | M | 1.759 | 1.784 | 0.548 | 0.412 | 0.350 |
| | SD | 0.260 | 0.265 | | | |
| ZCR*** | M | 0.050 | 0.046 | 0.000 | <0.001 | <0.001 |
| | SD | 0.008 | 0.009 | | | |
| Pause_portion** | M | 0.676 | 0.687 | 0.013 | 0.001 | 0.003 |
| | SD | 0.025 | 0.030 | | | |
| Total count_log** | M | 1.444 | 1.643 | 0.000 | <0.001 | 0.003 |
| | SD | 0.213 | 0.243 | | | |

† student T test

‡ multivariate analysis of covariance (MANCOVA)

* p<0.05, ** p<0.01, *** p<0.001

표 2-6. Classification results by voice and speech features between HC and CD

| | | Depression (binary) | Depression (binary, selected features) |
|---|---|---|---|
| Logistic Regression | F1 score | 0.582 | 0.624 |
| | Accuracy | 0.618 | 0.647 |
| | Sensitivity | 0.664 | 0.672 |
| | Specificity | 0.572 | 0.639 |
| Gaussian Naive Bayes | F1 score | 0.777 | 0.722 |
| | Accuracy | 0.776 | 0.741 |
| | Sensitivity | 0.743 | 0.759 |
| | Specificity | **0.823** | 0.736 |
| AdaBoost | F1 score | 0.798 | 0.789 |
| | Accuracy | 0.794 | 0.782 |
| | Sensitivity | 0.732 | 0.700 |
| | Specificity | **0.863** | **0.873** |
| Quadratic Discriminant Analysis | F1 score | **0.820** | 0.799 |
| | Accuracy | **0.824** | **0.806** |
| | Sensitivity | **0.806** | **0.801** |
| | Specificity | **0.846** | **0.819** |
| Support Vector Machine | F1 score | 0.514 | 0.374 |
| | Accuracy | 0.529 | 0.400 |
| | Sensitivity | 0.538 | 0.394 |
| | Specificity | 0.531 | 0.499 |
| Multi-layer Perceptron | F1 score | 0.527 | 0.499 |
| | Accuracy | 0.512 | 0.471 |
| | Sensitivity | 0.449 | 0.230 |
| | Specificity | 0.585 | 0.771 |

Bold: mean or max is greater than or equal to 0.800

표 2-7. Classification results by text between HC and CD

| | Demographic | Text | Ensemble |
|---|---|---|---|
| Accuracy | 0.681 | **0.831** | **0.831** |
| Sensitivity | 0.671 | **0.699** | 0.762 |
| Specificity | 0.692 | **0.964** | **0.951** |
| AUC | 0.761 | **0.905** | **0.907** |

Bold: mean or max is greater than or equal to 0.800

표 2-8. Correlation analysis between voice and speech features and clinical characteristics in healthy control group

| | | HDRS | BAI | BSS |
|---|---|---|---|---|
| HDRS | Pearson linear correlation coefficient | 1 | .338** | .112 |
| | P value | | .002 | .315 |
| BAI | Pearson linear correlation coefficient | .338** | 1 | .239* |
| | P value | .002 | | .029 |
| BSS | Pearson linear correlation coefficient | .112 | .239* | 1 |
| | P value | .315 | .029 | |
| Time_log | Pearson linear correlation coefficient | -.047 | .264* | .012 |
| | P value | .674 | .016 | .916 |
| Spectral_centroid | Pearson linear correlation coefficient | -.191 | -.201 | .009 |
| | P value | .084 | .069 | .938 |
| Spectral_band | Pearson linear correlation coefficient | -.215 | -.226* | .056 |
| | P value | .051 | .040 | .614 |
| Spectral_roll | Pearson linear correlation coefficient | -.177 | -.202 | .007 |
| | P value | .109 | .067 | .953 |
| Spectral_rmse | Pearson linear correlation coefficient | -.001 | -.205 | -.124 |
| | P value | .995 | .063 | .265 |
| Formant0 | Pearson linear correlation coefficient | .135 | -.265* | -.278* |
| | P value | .224 | .016 | .011 |
| Formant1 | Pearson linear correlation coefficient | -.210 | .104 | .171 |
| | P value | .057 | .351 | .123 |
| Formant2 | Pearson linear correlation coefficient | -.208 | .108 | .319** |
| | P value | .060 | .331 | .003 |
| Formant_bw0 | Pearson linear correlation coefficient | -.001 | -.039 | -.012 |
| | P value | .993 | .729 | .911 |
| Formant_bw1 | Pearson linear correlation coefficient | .022 | .094 | -.114 |
| | P value | .845 | .400 | .305 |

| | | | | |
|---|---|---|---|---|
| Formant_bw2 | Pearson linear correlation coefficient | -.075 | -.074 | .209 |
| | P value | .499 | .505 | .058 |
| Mmean_p | Pearson linear correlation coefficient | -.247* | -.119 | .000 |
| | P value | .025 | .283 | .997 |
| Error_p | Pearson linear correlation coefficient | .271* | .132 | .015 |
| | P value | .013 | .234 | .895 |
| Change_p_log | Pearson linear correlation coefficient | .023 | .292** | .017 |
| | P value | .834 | .007 | .878 |
| Mean_m | Pearson linear correlation coefficient | -.122 | -.157 | -.074 |
| | P value | .273 | .156 | .506 |
| Error_m | Pearson linear correlation coefficient | -.147 | -.060 | .043 |
| | P value | .185 | .590 | .703 |
| Change_m_log | Pearson linear correlation coefficient | -.089 | .224* | .034 |
| | P value | .421 | .042 | .761 |
| ZCR | Pearson linear correlation coefficient | -.202 | -.123 | .003 |
| | P value | .067 | .270 | .981 |
| Pause portion | Pearson linear correlation coefficient | .229* | .093 | -.022 |
| | P value | .037 | .404 | .847 |
| Total_count_log | Pearson linear correlation coefficient | .229* | .097 | .079 |
| | P value | .037 | .381 | .480 |

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 2-9. Correlation analysis between voice and speech features and clinical characteristics in current depression group

| | | HDRS | BAI | BSS |
|---|---|---|---|---|
| HDRS | Pearson linear correlation coefficient | 1 | .539*** | .458*** |
| | P value | | <.001 | <.001 |
| BAI | Pearson linear correlation coefficient | .539*** | 1 | .528*** |
| | P value | <.001 | | <.001 |
| BSS | Pearson linear correlation coefficient | .458*** | .528*** | 1 |
| | P value | <.001 | <.001 | |
| Time_log | Pearson linear correlation coefficient | .161 | .255* | .151 |
| | P value | .146 | .020 | .173 |
| Spectral_centroid | Pearson linear correlation coefficient | .036 | .149 | .239* |
| | P value | .748 | .180 | .030 |
| Spectral_band | Pearson linear correlation coefficient | .058 | .139 | .300** |
| | P value | .599 | .211 | .006 |
| Spectral_roll | Pearson linear correlation coefficient | .043 | .159 | .251* |
| | P value | .702 | .150 | .022 |
| Spectral_rmse | Pearson linear correlation coefficient | .341** | .205 | .130 |
| | P value | .002 | .064 | .243 |
| Formant0 | Pearson linear correlation coefficient | .120 | .183 | -.075 |
| | P value | .282 | .097 | .499 |
| Formant1 | Pearson linear correlation coefficient | .026 | -.026 | .037 |
| | P value | .815 | .816 | .739 |
| Formant2 | Pearson linear correlation coefficient | .117 | -.013 | .139 |
| | P value | .292 | .905 | .210 |
| Formant_bw0 | Pearson linear correlation coefficient | .132 | .169 | .199 |
| | P value | .235 | .127 | .071 |
| Formant_bw1 | Pearson linear correlation coefficient | .251* | .178 | .288** |
| | P value | .022 | .107 | .008 |

| | | | | |
|---|---|---|---|---|
| Formant_bw2 | Pearson linear correlation coefficient | -.129 | .048 | .114 |
| | P value | .244 | .663 | .303 |
| Mmean_p | Pearson linear correlation coefficient | .026 | .039 | .132 |
| | P value | .816 | .727 | .233 |
| Error_p | Pearson linear correlation coefficient | -.057 | .000 | -.154 |
| | P value | .608 | .999 | .164 |
| Change_p_log | Pearson linear correlation coefficient | .129 | .266* | .129 |
| | P value | .244 | .015 | .244 |
| Mean_m | Pearson linear correlation coefficient | .036 | .041 | .078 |
| | P value | .749 | .712 | .485 |
| Error_m | Pearson linear correlation coefficient | .060 | .016 | .049 |
| | P value | .591 | .887 | .661 |
| Change_m_log | Pearson linear correlation coefficient | .158 | .247* | .180 |
| | P value | .153 | .024 | .104 |
| ZCR | Pearson linear correlation coefficient | .053 | .149 | .219* |
| | P value | .634 | .180 | .046 |
| Pause portion | Pearson linear correlation coefficient | -.083 | -.112 | -.267* |
| | P value | .458 | .312 | .015 |
| Total_count_log | Pearson linear correlation coefficient | .067 | .048 | .035 |
| | P value | .546 | .665 | .757 |

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 2-10. Mediation analysis between voice and speech features and clinical characteristics in healthy control group

| Direct effect of X on Y | | | | | |
|---|---|---|---|---|---|
| Effect | se | t | p | LLCI | ULCI |
| -.0048 | .0034 | -1.4022 | .1648 | -.0116 | .0020 |

| Indirect effect(s) of X on Y: | | | | |
|---|---|---|---|---|
| | Effect | BootSE | BootLLCI | BootULCI |
| TOTAL | -.0016 | .0015 | -.0043 | .0015 |
| Ind1 | -.0020 | .0012 | -.0045 | .0004 |
| Ind2 | .0001 | .0009 | -.0011 | .0027 |
| Ind3 | .0003 | .0004 | -.0006 | .0010 |

| Indirect effect key: | | | | | | |
|---|---|---|---|---|---|---|
| Ind1 HDRS | -> | BAI | -> | band | | |
| Ind2 HDRS | -> | BSS | -> | band | | |
| Ind3 HDRS | -> | BAI | -> | BSS | -> | band |

* p<0.05, ** p<0.01, *** p<0.001

표 2-11. Mediation analysis between voice and speech features and clinical characteristics in current depression group

| Direct effect of X on Y | | | | | |
|---|---|---|---|---|---|
| Effect | se | t | p | LLCI | ULCI |
| -.0015 | .0019 | -.8125 | .4190 | -.0052 | .0022 |

| Indirect effect(s) of X on Y: | | | | |
|---|---|---|---|---|
| | Effect | BootSE | BootLLCI | BootULCI |
| TOTAL | .0024 | .0013 | -.0001 | .0052 |
| Ind1 | .0001 | .0011 | -.0021 | .0024 |
| Ind2* | .0012 | .0008 | .0000 | .0031 |
| Ind3* | .0010 | .0006 | .0002 | .0024 |

| Indirect effect key: | | | | | | |
|---|---|---|---|---|---|---|
| Ind1 HDRS | -> | BAI | -> | band | | |
| Ind2 HDRS -> | BSS | -> | band | | | |
| Ind3 HDRS -> | BAI | -> | BSS | -> | band | |

* p<0.05, ** p<0.01, *** p<0.001

표 3-1. Demographics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : BSS)

| | | DLSR | DHSR | P value |
|---|---|---|---|---|
| N | | 33 | 50 | |
| age* | mean | 33.788 | 29.020 | 0.049 |
| | SD | 11.376 | 10.109 | |
| BMI | mean | 24.447 | 25.048 | 0.581 |
| | SD | 4.602 | 4.973 | |
| sex | M | 8 (24.4%) | 11 (22.0%) | 0.812 |
| | F | 25 (75.8%) | 39 (78.0%) | |
| SES | low | 8 (24.2%) | 11 (22.0%) | 0.885 |
| | med | 16 (48.5%) | 27 (54.0%) | |
| | high | 9 (27.3%) | 12 (24.0%) | |
| Non-psychiatric medication | Yes | 7 (21.2%) | 8 (16.0%) | 0.546 |
| | no | 26 (78.8%) | 42 (84.0%) | |
| Subjective suffering due to non-psychiatric disease | mean | 0.210 | 0.340 | 0.372 |
| | SD | 0.545 | 0.668 | |
| AP_OZP | N | 33 | 50 | 0.071 |
| | mean | 4.361 | 7.026 | |
| | SD | 4.627 | 7.471 | |
| MINI | MDD | 6 (18.2%) | 5 (10.0%) | 0.332 |
| | BP | 27 (81.8%) | 45 (90.0%) | |
| | panic disorder* | 1 (3.0%) | 9 (18.0%) | 0.040 |
| | social anxiety disorder | 2 (6.1%) | 3 (6.0%) | >0.999 |
| | OCD | 2 (6.1%) | 5 (10.0%) | 0.527 |
| | PTSD | 1 (3.0%) | 2 (4.0%) | 0.817 |
| | AUD | 2 (6.1%) | 9 (18.0%) | 0.116 |
| | BN | 1 (3.0%) | 4 (8.0%) | 0.352 |
| | GAD | 4 (12.1%) | 8 (16.0%) | 0.623 |

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 3-2. Clinical characteristics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : BSS)

| | | DLSR | DHSR | P value |
|---|---|---|---|---|
| N | | 33 | 50 | |
| PHQ*** | mean | 11.455 | 17.680 | <0.001 |
| | SD | 5.900 | 5.937 | |
| HDRS** | mean | 14.424 | 18.100 | <0.001 |
| | SD | 4.430 | 4.509 | |
| BAI*** | mean | 15.939 | 29.680 | <0.001 |
| | SD | 13.122 | 16.595 | |
| BIS | mean | 63.000 | 63.880 | 0.623 |
| | SD | 7.850 | 8.030 | |
| BSS*** | mean | 8.576 | 24.460 | <0.001 |
| | SD | 5.728 | 5.312 | |

* p<0.05, ** p<0.01, *** p<0.001

표 3-3. Voice and Speech features between DLSR and DHSR (suicidal risk : BSS)

| | | DLSR | DHSR | P value† | Adjusted age‡ | Adjusted age, BAI‡ |
|---|---|---|---|---|---|---|
| N | | 33 | 50 | | | |
| Time_log | mean | 1.081 | 1.134 | 0.293 | 0.136 | 0.568 |
| | SD | 0.200 | 0.238 | | | |
| Spectral_centroid | mean | 7.367 | 7.390 | 0.330 | 0.618 | 0.995 |
| | SD | 0.091 | 0.113 | | | |
| Spectral_band | mean | 7.408 | 7.435 | 0.081 | 0.189 | 0.357 |
| | SD | 0.064 | 0.070 | | | |
| Spectral_roll | mean | 7.982 | 8.012 | 0.367 | 0.698 | 0.863 |
| | SD | 0.125 | 0.156 | | | |
| Spectral_rmse | mean | 4.078 | 4.117 | 0.823 | 0.808 | 0.287 |
| | SD | 0.746 | 0.803 | | | |
| Formant0* | mean | 6.233 | 6.213 | 0.256 | 0.109 | 0.012 |
| | SD | 0.066 | 0.081 | | | |
| Formant1 | mean | 7.344 | 7.342 | 0.946 | 0.852 | 0.929 |
| | SD | 0.092 | 0.116 | | | |
| Formant2 | mean | 8.020 | 8.033 | 0.467 | 0.659 | 0.527 |
| | SD | 0.078 | 0.075 | | | |
| Formant _bw0 | mean | 39.049 | 44.990 | 0.040 | 0.100 | 0.246 |
| | SD | 10.559 | 13.878 | | | |
| Formant _bw1* | mean | 184.473 | 214.059 | 0.006 | 0.006 | 0.021 |
| | SD | 42.094 | 49.133 | | | |
| Formant _bw2* | mean | 204.699 | 224.455 | 0.034 | 0.016 | 0.016 |
| | SD | 48.080 | 35.401 | | | |
| Mean_p | mean | 276.871 | 277.979 | 0.869 | 0.809 | 0.707 |
| | SD | 24.839 | 32.691 | | | |
| Error_p | mean | 0.327 | 0.322 | 0.770 | 0.894 | 0.922 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | SD | 0.064 | 0.075 | | | |
| Change_p_log | mean | 1.494 | 1.542 | 0.350 | 0.126 | 0.592 |
| | SD | 0.221 | 0.232 | | | |
| Mean_m | mean | 62.450 | 63.454 | 0.688 | 0.836 | 0.933 |
| | SD | 9.636 | 11.959 | | | |
| Error_m | mean | 0.883 | 0.865 | 0.591 | 0.503 | 0.437 |
| | SD | 0.127 | 0.167 | | | |
| Change_m_log | mean | 1.753 | 1.804 | 0.398 | 0.264 | 0.832 |
| | SD | 0.222 | 0.291 | | | |
| ZCR | mean | 0.044 | 0.047 | 0.219 | 0.333 | 0.620 |
| | SD | 0.007 | 0.010 | | | |
| Pause_portion | mean | 0.691 | 0.684 | 0.318 | 0.440 | 0.673 |
| | SD | 0.023 | 0.034 | | | |
| Total count_log | mean | 1.661 | 1.630 | 0.573 | 0.473 | 0.339 |
| | SD | 0.171 | 0.282 | | | |

† student T test

‡ multivariate analysis of covariance (MANCOVA)

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 3-4. Classification results by voice and speech features between DLSR and DHSR (suicidal risk; BSS)

| | | **SSI for depression** |
|---|---|---|
| Logistic Regression | F1 score | 0.789 |
| | Accuracy | 0.718 |
| | Sensitivity | 0.495 |
| | Specificity | **0.824** |
| Gaussian Naive Bayes | F1 score | 0.649 |
| | Accuracy | 0.576 |
| | Sensitivity | 0.503 |
| | Specificity | 0.620 |
| AdaBoost | F1 score | 0.580 |
| | Accuracy | 0.482 |
| | Sensitivity | 0.438 |
| | Specificity | 0.558 |
| Quadratic Discriminant Analysis | F1 score | 0.771 |
| | Accuracy | 0.647 |
| | Sensitivity | 0.167 |
| | Specificity | **0.912** |
| Support Vector Machine | F1 score | 0.789 |
| | Accuracy | 0.718 |
| | Sensitivity | 0.462 |
| | Specificity | **0.840** |
| Multi-layer Perceptron | F1 score | 0.673 |
| | Accuracy | 0.624 |
| | Sensitivity | 0.535 |
| | Specificity | 0.649 |

Bold: mean or max is greater than or equal to 0.800

표 3-5. Classification results by voice and speech features between DLSR and DHSR (suicidal risk : BSS)

| | Demographic-only | Text-only | Ensemble |
|---|---|---|---|
| AUC | 0.511 | 0.626 | **0.874** |
| Accuracy | 0.578 | 0.602 | 0.795 |
| Sensitivity | 0.444 | 0.489 | 0.703 |
| Specificity | 0.643 | 0.750 | **0.870** |
| p-value (Delong test) | | 0.308 | 0.012* |

Bold: mean or max is greater than or equal to 0.800

* p<0.05, ** p<0.01, *** p<0.001

표 3-6. Demographics between depression with low suicidal risk and depression with high suicidal risk (suicidal risk : MINI)

| | | DLSR | DHSR | P value |
|---|---|---|---|---|
| N | | 52 | 31 | |
| Age* | mean | 32.865 | 27.645 | 0.033 |
| | SD | 11.381 | 9.065 | |
| BMI | mean | 24.366 | 25.551 | 0.280 |
| | SD | 4.049 | 5.870 | |
| sex | M | 12 (23.1%) | 7 (22.6%) | 0.958 |
| | F | 40 (76.9%) | 24 (77.4%) | |
| SES | low | 14 (26.9%) | 5 (16.1%) | 0.209 |
| | med | 28 (53.8%) | 15 (48.4%) | |
| | high | 10 (19.2%) | 11 (35.5%) | |
| Non-psychiatric medication | Yes | 12 (23.1%) | 3 (9.7%) | 0.125 |
| | no | 40 (76.9%) | 28 (90.3%) | |
| Subjective suffering due to non-psychiatric disease | mean | 0.230 | 0.390 | 0.281 |
| | SD | 0.581 | 0.715 | |
| AP_OZP* | N | 52 | 31 | 0.023 |
| | mean | 4.705 | 8.082 | |
| | SD | 4.872 | 8.424 | |
| MINI | MDD | 7 (13.5%) | 4 (12.9%) | 0.942 |
| | BP | 45 (86.5%) | 27 (87.1%) | |
| | panic disorder** | 2 (3.8%) | 8 (25.8%) | 0.003 |
| | social anxiety disorder* | 1 (1.9%) | 4 (12.9%) | 0.040 |
| | OCD | 4 (7.7%) | 3 (9.7%) | 0.753 |
| | PTSD | 1 (1.9%) | 2 (6.5%) | 0.285 |
| | AUD | 7 (13.5%) | 4 (12.9%) | 0.942 |
| | BN | 3 (5.8%) | 2 (6.5%) | 0.899 |
| | GAD | 5 (9.6%) | 7 (22.6%) | 0.104 |

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 3-7. Clinical characteristics between low suicidal risk and depression with high suicidal risk (suicidal risk : MINI)

| | | DLSR | DHSR | P value |
|---|---|---|---|---|
| N | | 52 | 31 | |
| PHQ** | mean | 13.462 | 18.129 | 0.002 |
| | SD | 5.782 | 7.032 | |
| HDRS** | mean | 15.500 | 18.548 | 0.005 |
| | SD | 4.734 | 4.358 | |
| BAI** | mean | 20.577 | 30.323 | 0.009 |
| | SD | 15.301 | 17.294 | |
| BIS | mean | 63.981 | 62.774 | 0.505 |
| | SD | 7.229 | 9.043 | |
| BSS*** | mean | 13.692 | 25.613 | <0.001 |
| | SD | 7.935 | 7.017 | |

* p<0.05, ** p<0.01, *** p<0.001

표 3-8. Voice and Speech features between DLSR and DHSR (suicidal risk : MINI)

| | | DLSR | DHSR | P value† | Adjusted age, AP_OZP‡ | Adjusted age, AP_OZP and BAI‡ |
|---|---|---|---|---|---|---|
| N | | 52 | 31 | | | |
| Time_log | mean | 1.109 | 1.120 | 0.910 | 0.381 | 0.807 |
| | SD | 0.210 | 0.249 | | | |
| Spectral_centroid | mean | 7.368 | 7.403 | 0.094 | 0.356 | 0.515 |
| | SD | 0.091 | 0.122 | | | |
| Spectral_band | mean | 7.416 | 7.437 | 0.132 | 0.417 | 0.580 |
| | SD | 0.064 | 0.075 | | | |
| Spectral_roll | mean | 7.982 | 8.030 | 0.078 | 0.395 | 0.583 |
| | SD | 0.129 | 0.165 | | | |
| Spectral_rmse | mean | 4.084 | 4.131 | 0.693 | 0.509 | 0.288 |
| | SD | 0.794 | 0.757 | | | |
| Formant0 | mean | 6.228 | 6.209 | 0.217 | 0.229 | 0.073 |
| | SD | 0.070 | 0.085 | | | |
| Formant1 | mean | 7.340 | 7.349 | 0.895 | 0.656 | 0.627 |
| | SD | 0.099 | 0.119 | | | |
| Formant2 | mean | 8.028 | 8.028 | 1.000 | 0.853 | 0.863 |
| | SD | 0.075 | 0.079 | | | |
| Formant _bw0 | mean | 42.455 | 42.918 | 0.778 | 0.972 | 0.622 |
| | SD | 12.783 | 13.377 | | | |
| Formant _bw1* | mean | 194.014 | 216.188 | 0.058 | 0.021 | 0.052 |
| | SD | 43.338 | 53.868 | | | |
| Formant _bw2 | mean | 213.884 | 221.156 | 0.624 | 0.274 | 0.323 |
| | SD | 44.841 | 36.315 | | | |
| Mean_p | mean | 276.507 | 279.268 | 0.413 | 0.912 | 0.873 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | SD | 25.465 | 36.015 | | | |
| Error_p | mean | 0.332 | 0.310 | 0.221 | 0.392 | 0.353 |
| | SD | 0.068 | 0.073 | | | |
| Change_p_log | mean | 1.523 | 1.523 | 0.851 | 0.405 | 0.882 |
| | SD | 0.222 | 0.240 | | | |
| Mean_m | mean | 63.875 | 61.679 | 0.498 | 0.267 | 0.219 |
| | SD | 10.622 | 11.765 | | | |
| Error_m | mean | 0.875 | 0.867 | 0.829 | 0.552 | 0.554 |
| | SD | 0.137 | 0.176 | | | |
| Change_m_log | mean | 1.778 | 1.794 | 0.865 | 0.508 | 0.940 |
| | SD | 0.237 | 0.311 | | | |
| ZCR | mean | 0.044 | 0.047 | 0.247 | 0.456 | 0.615 |
| | SD | 0.008 | 0.011 | | | |
| Pause_portion | mean | 0.693 | 0.676 | 0.020 | 0.058 | 0.075 |
| | SD | 0.026 | 0.035 | | | |
| Total count_log | mean | 1.643 | 1.641 | 0.959 | 0.920 | 0.821 |
| | SD | 0.178 | 0.327 | | | |

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

표 3-9. Classification results by voice and speech features between DLSR and DHSR (suicidal risk; MINI)

| | | MINI for depression |
|---|---|---|
| Logistic Regression | F1 score | 0.310 |
| | Accuracy | 0.565 |
| | Sensitivity | 0.24 |
| | Specificity | **0.864** |
| Gaussian Naive Bayes | F1 score | 0.444 |
| | Accuracy | 0.529 |
| | Sensitivity | 0.441 |
| | Specificity | 0.646 |
| AdaBoost | F1 score | 0.499 |
| | Accuracy | 0.576 |
| | Sensitivity | 0.490 |
| | Specificity | 0.703 |
| Quadratic Discriminant Analysis | F1 score | 0.044 |
| | Accuracy | 0.541 |
| | Sensitivity | 0.025 |
| | Specificity | **0.983** |
| Support Vector Machine | F1 score | 0.180 |
| | Accuracy | 0.541 |
| | Sensitivity | 0.177 |
| | Specificity | **0.913** |
| Multi-layer Perceptron | F1 score | 0.294 |
| | Accuracy | 0.565 |
| | Sensitivity | 0.219 |
| | Specificity | **0.852** |

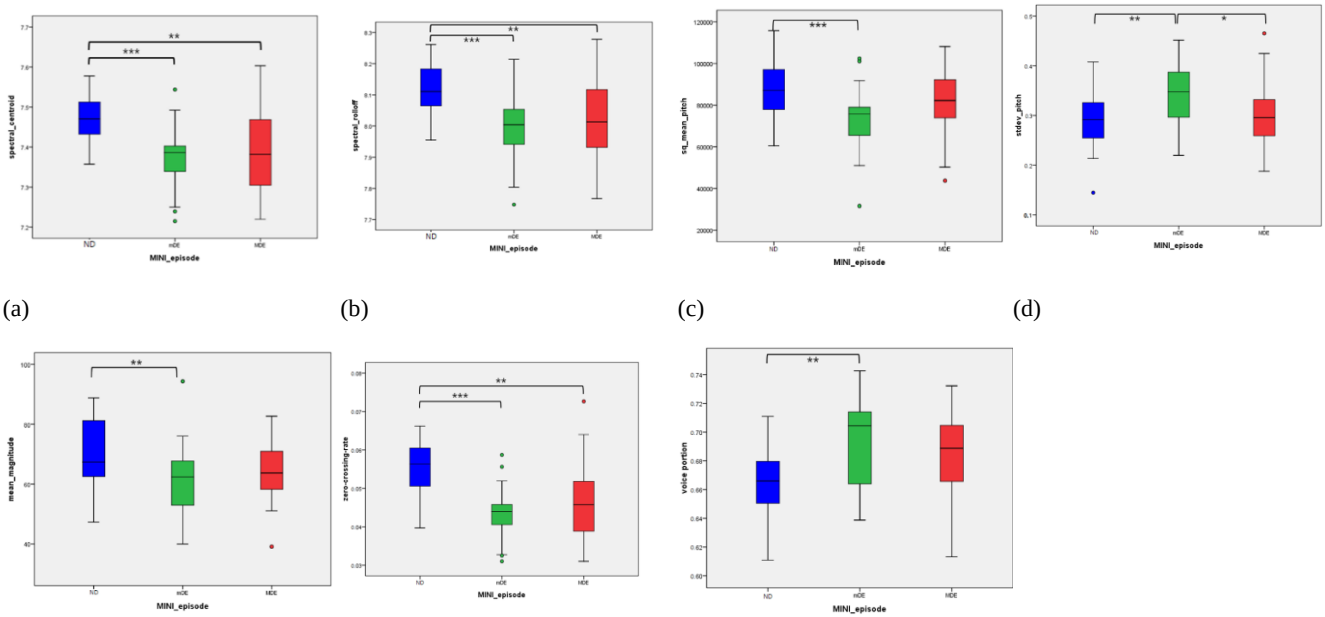
Bold: mean or max is greater than or equal to 0.800

표 3-10. Classification results by voice and speech features between DLSR and DHSR (suicidal risk : MINI)

| | Demographic | Text | Ensemble |
|---|---|---|---|
| Accuracy | 0.542 | 0.602 | 0.747 |
| Sensitivity | 0.625 | 0.744 | **0.816** |
| Specificity | 0.370 | 0.477 | 0.647 |
| AUC | 0.499 | 0.632 | **0.800** |

Bold: mean or max is greater than or equal to 0.800

그림 1-1. Difference of voice features by depressive episode



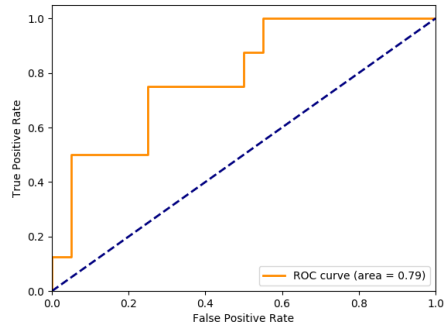

(a) (b) (c) (d)

(e) (f) (g)

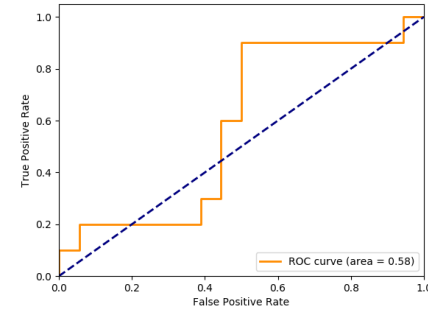
(a) Spectral_centroid between three groups; (b) spectral_rolloff between three groups; (c) sq_mean_pitch between three groups; (d) stdev_pitch between three groups; (e) mean_magnitude between three groups; (f) zero-crossing-rate between three groups; (g) voice portion between three groups.

* p value < 0.05, ** p value < 0.01, *** p value < 0.001

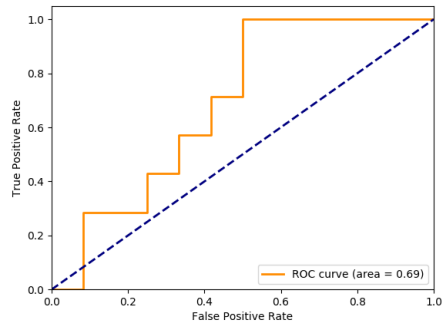
그림 1-2. AUC curve predicting minor and major episodes using MLP



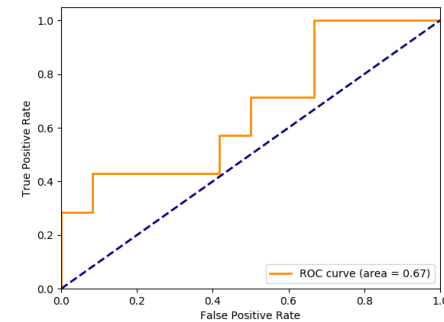

(a)




(b)




(c)




(d)

(a) AUC for minor episode, 7:3 training; (b) AUC for major episode, 7:3 training; (c) AUC for minor episode, 8:2 training; (d) AUC for major episode, 8:2 training;

그림 2-1. ROC curve of classification HC and CD by voice and speech features

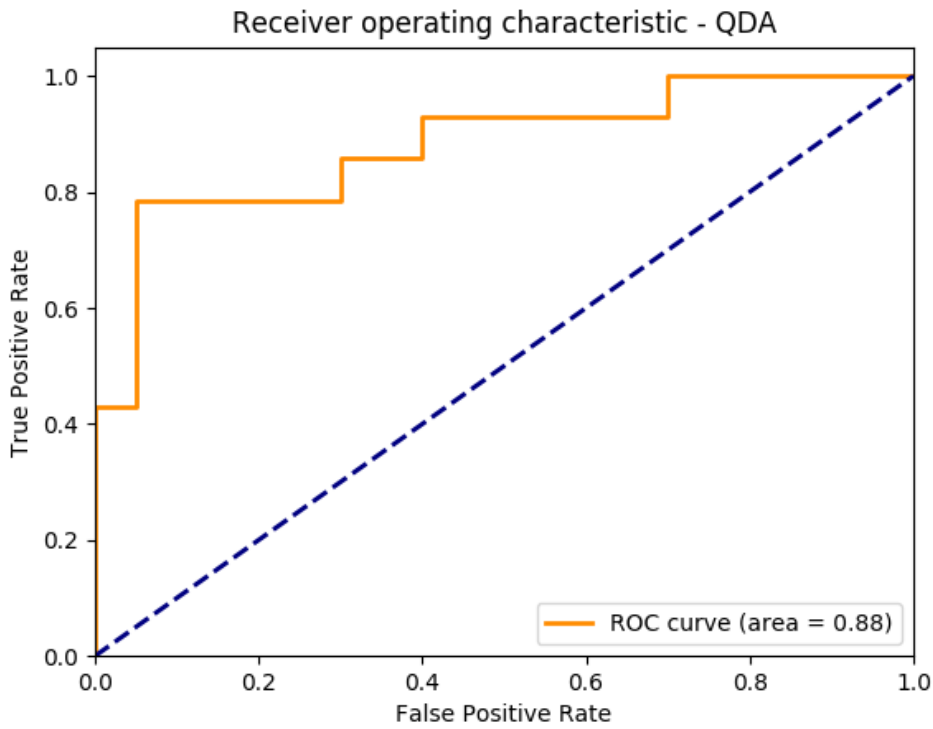


그림 2-2. Feature importance for classifying HC and CD

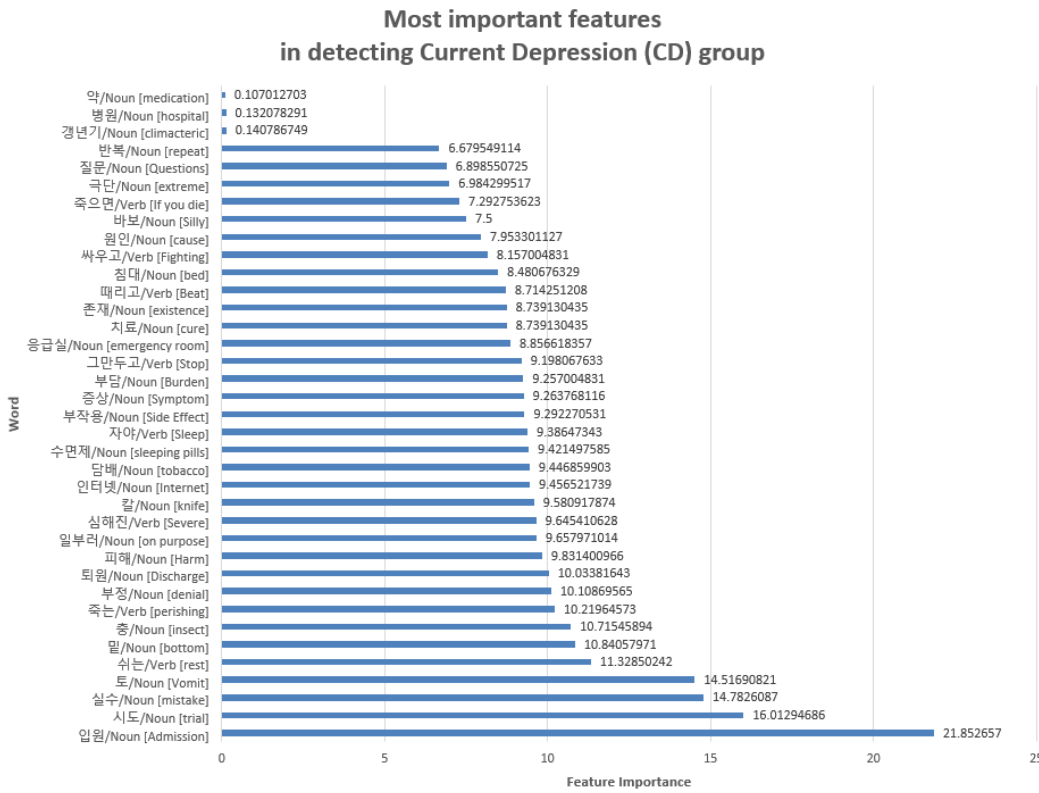


그림 2-3. Text mining of important features in HC and CD

그림 2-4. Classification results between HC and CD by text

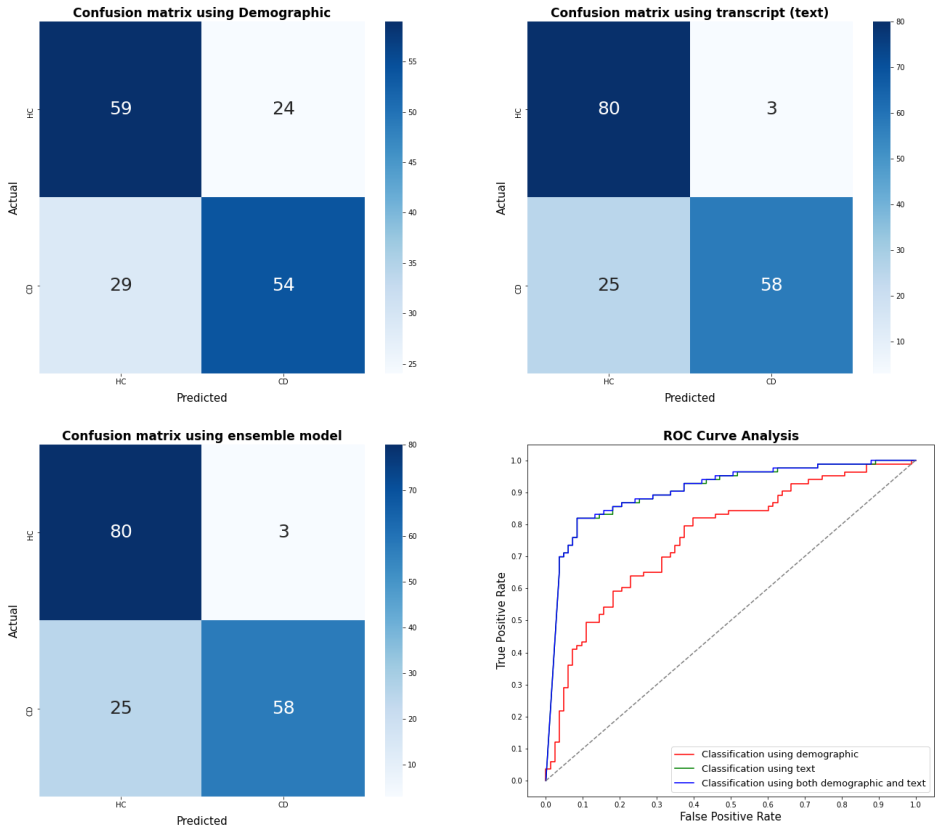
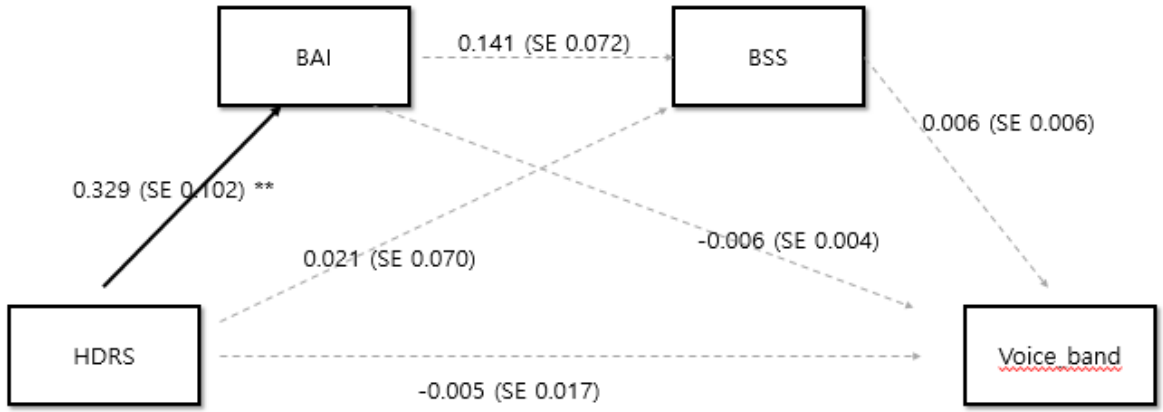
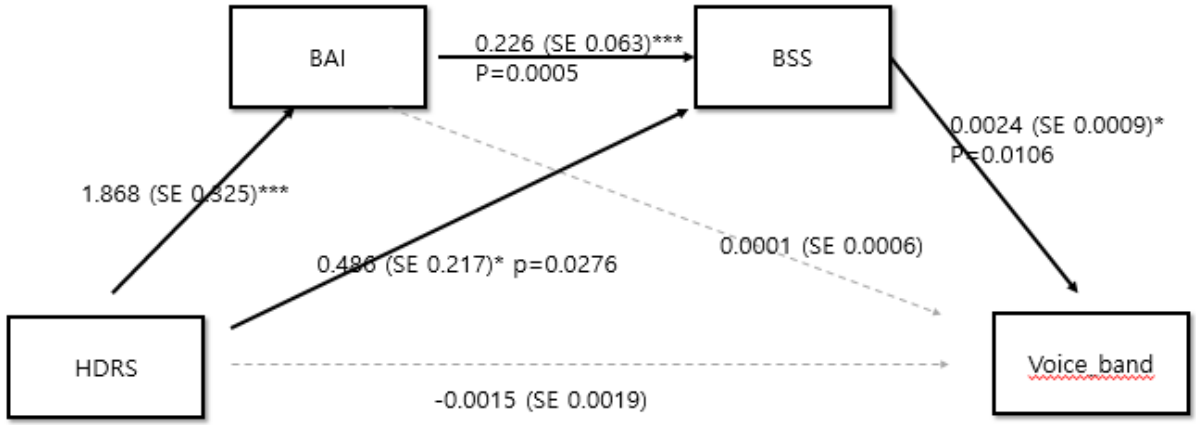


그림 2-5. Mediation analysis between voice and speech feature and clinical characteristics in healthy control group




* p<0.05, ** p<0.01, *** p<0.001

그림 2-6. Mediation analysis between voice and speech feature and clinical characteristics in current depression group




* p<0.05, ** p<0.01, *** p<0.001

그림 3-1. ROC curve of classification DLSR and DHSR by voice and speech features (suicidal risk : BSS)

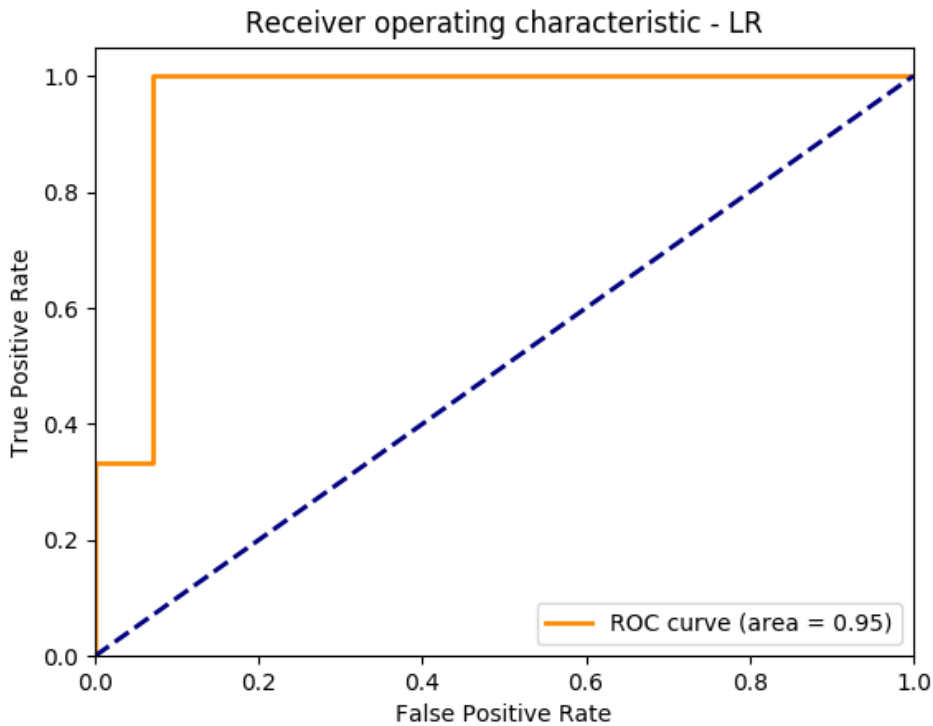


그림 3-2. Classification results of DLSR and DHSR by text and clinical characteristics (suicidal risk : BSS)

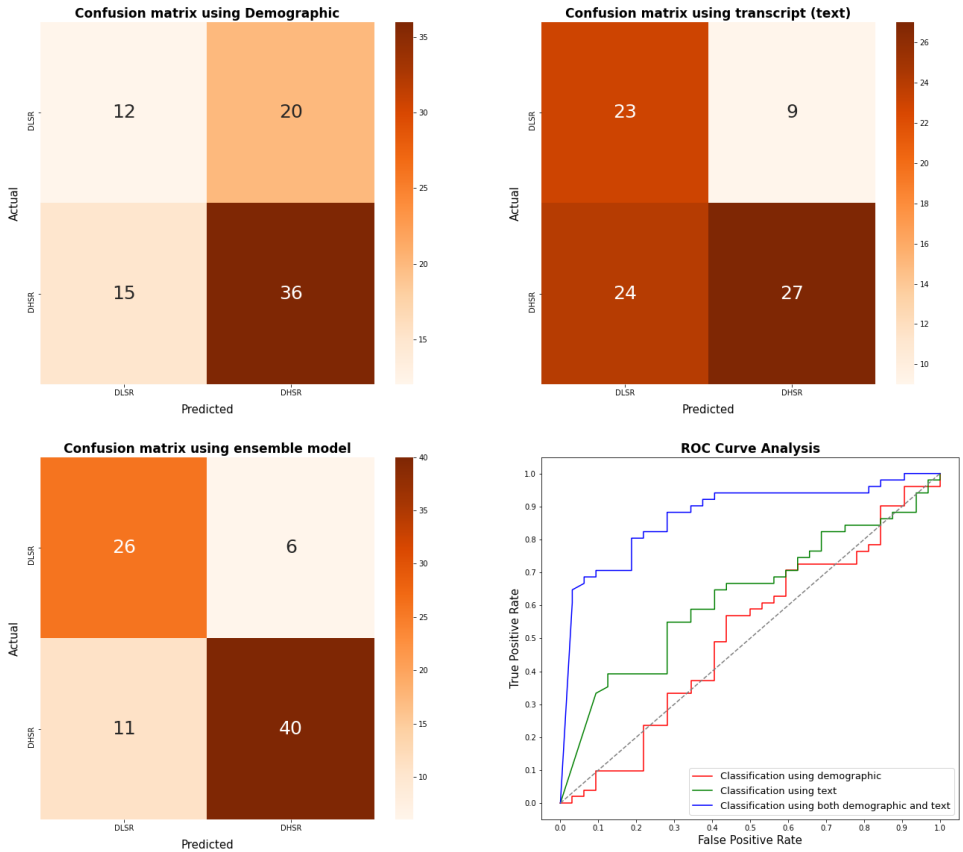


그림 3-3. ROC curve of classification DLSR and DHSR by voice and speech features (suicidal risk : MINI)

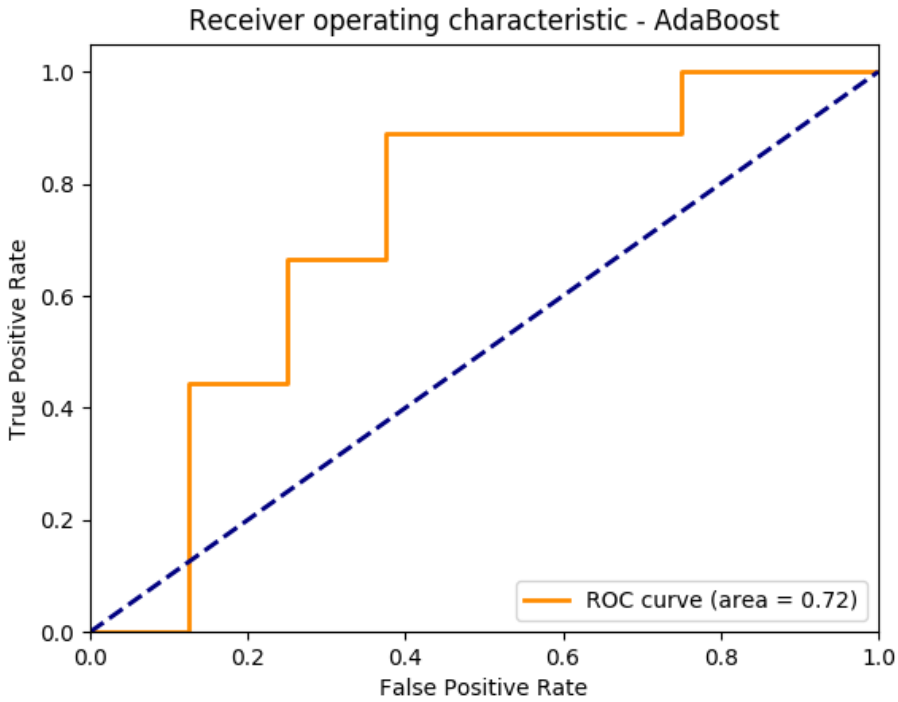
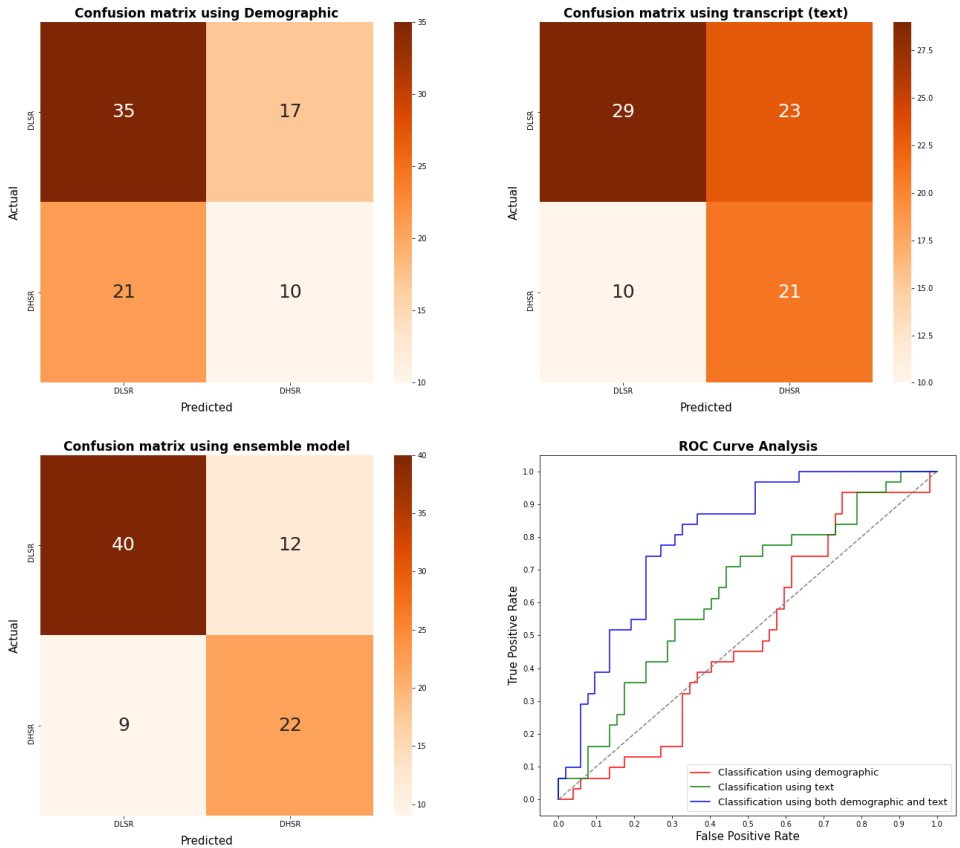


그림 3-4. Classification results of DLSR and DHSR by text and clinical characteristics (suicidal risk : MINI)



# 참고 문헌


1. W. H. O. "Depression and Other common mental disorders: global health estimates." Geneva: World Health Organization (2017): 1-24.
2. Lepine, J.P. and M. Briley, The increasing burden of depression. Neuropsychiatr Dis Treat, 2011. 7(Suppl 1): p. 3-7.
3. Mathers, C.D. and D. Loncar, Updated projections of global mortality and burden of disease, 2002-2030: data sources, methods and results. Geneva: World Health Organization, 2005.
4. TB USTUN. Global burden of depressive disorders in the year 2000. British J Psychiatry. 2004;184:3 8 6-9 2.
5. Wagner CJ, Metzger FG, Sievers C, Marschall U, L'Hoest H, Stollenwerk B, et al. Depression-related treatment and costs in Germany: do they change with comorbidity? A claims data analysis. J Affect Disord. 2016;193:257-66.
6. Kleine-Budde K, Muller R, Kawohl W, Bramesfeld A, Moock J, Rossler W. The cost of depression - a cost analysis from a large database. J Affect Disord. 2013;147(1-3):137-43.
7. Sobocki P, Lekander I, Borgstrom F, Strom O, Runeson B. The economic burden of depression in Sweden from 1997 to 2005. Eur Psychiatry. 2007; 22(3):146-52.
8. Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the

United States (2005 and 2010). J Clin Psychiatry. 2015;76(2):155-62.

9. Chang SM, Hong JP, Cho MJ. Economic burden of depression in South Korea. Soc Psychiatry Psychiatr Epidemiol. 2012;47(5):683 -9.

10. Shin, D., Kim, N. W., Kim, M. J., Rhee, S. J., Park, C. H. K., Kim, H., ... & Ahn, Y. M. (2020). Cost analysis of depression using the national insurance system in South Korea: a comparison of depression and treatment-resistant depression. BMC health services research, 20(1), 1-11.

11. Jamison, K. and F. Goodwin, Suicide in manic depressive illness. Manic-depressive illness, 1990: p. 157-173.

12. World Health Organization. Mental health: Suicide data [Internet]. 2019 [cited 13 October 2019]. [Available from: https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/

13. Naghavi, M. and C. Global Burden of Disease Self-Harm, Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the Global Burden of Disease Study 2016. BMJ, 2019. 364: p. l94.

14. Statistics Korea. Causes of Death Statistics in 2018 [Internet]. 2019 [cited 13 October 2019]. [Available from: http://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01

15. Palmer, C. S., Halpern, M. T., & Hatziandreu, E. J. (1995). The cost of suicide and suicide attempts in the United States. Clinical Neuropharmacology, 18, S25-S33.
16. Association, A.P., Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). 2013, Arlington, VA, US.: American Psychiatric Pub.
17. International Statistical Classification of Diseases and Related Health Problems. Tenth Revision. Vol. 1. 1992, Geneva: World Health Organization.
18. Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, severity, and Comorbidityof 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry. 2005;62:617-709.
19. Kawakamia N, Shimizub H, Haratanic T, Iwatad N. Kitamurae T. lifetime and 6-month prevalence of DSM-III-R psychiatric disorders in an urban community in Japan. Psychiatry Res. 2004;121:293-301.
20. Ohayon MM, Hong SC. Prevalence of major depressive disorder in the general population of South Korea. J Psychiatr Res. 2006;40(1):30-6.
21. Sheehan, D.V., Depression: underdiagnosed, undertreated, underappreciated. Manag Care, 2004. 13(6 Suppl Depression): p. 6-8.
22. Davidson, J.R. and S.E. Meltzer-Brody, The underrecognition and

undertreatment of depression: what is the breadth and depth of the problem? Journal of clinical psychiatry, 1999. 60(7): p. 4-11.

23. Mitchell, A.J.; Vaze, A.; Rao, S. Clinical diagnosis of depression in primary care: A meta-analysis. Lancet 2009, 374, 609-619, doi:10.1016/s0140-6736(09)60879-5.

24. Watson, D., et al., The measurement of suicidal ideation. Crisis, 2001. 22(1): p. 12-4.

25. Ghasemi, P., A. Shaghaghi, and H. Allahverdipour, Measurement Scales of Suicidal Ideation and Attitudes: A Systematic Review Article. Health Promot Perspect, 2015. 5(3): p. 156-68.

26. Im, J. S., et al. (2018). "Cultural Stigma Manifested in Official Suicide Death in South Korea." Omega (Westport) 77(4): 386-403.

27. N. Sands, "Assessing the risk of suicide at triage," Australasian Emergency Nursing Journal, vol. 10, no. 4, pp. 161-163, 2007.

28. D. E. Clarke, A.-M. Brown, and L. Giles-Smith, "Triaging suicidal patients: sifting through the evidence," International Emergency Nursing, vol. 16, no. 3, pp. 165-174, 2008.

29. Sullivan, H.S., The Psychiatric Interview. Psychiatry, 1951. 14(4): p. 361-373.

30. E. Kraepelin, Manic depressive insanity and paranoia. J. Nerv. Ment. Dis., 53 (1921), p. 350

31. Pope, B., et al., Anxiety and depression in speech. J Consult Clin

Psychol, 1970. 35(1): p. 128-33.

32. Szabadi, E., C.M. Bradshaw, and J.A. Besson, Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. Br J Psychiatry, 1976. 129: p. 592-7.

33. A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, D.M. Wilkes, Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk, IEEE Trans. Bio-Eng., 51 (2004), pp. 1530-1540

34. Cummins, N., et al., A review of depression and suicide risk assessment using speech analysis. Speech Communication, 2015. 71: p. 10-49.

35. Little, B., et al., Deep learning-based automated speech detection as a marker of social functioning in late-life depression. Psychol Med, 2021. 51(9): p. 1441-1450.

36. Low, L.S., et al., Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng, 2011. 58(3): p. 574-86.

37. Vazquez-Romero, A. and A. Gallardo-Antolin, Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks. Entropy (Basel), 2020. 22(6).

38. Zhao, Y., et al., Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech. Front Neurorobot, 2021. 15: p. 684037.

39. Hashim, N.W., Wilkes, M., Salomon, R., Meggs, J., 2012. Analysis

of timing pattern of speech as possible indicator for near-term suicidal risk and depression in male patients. In: 2012 International Conference on Signal Processing Systems (ICSPS 2012), pp. 6-13.

40. Sourirajan, V., et al., A machine learning approach to detect suicidal ideation in us veterans based on acoustic and linguistic features of speech. arXiv preprint arXiv:2009.09069, 2020.

41. Bouloudnine, S., et al., [Perception of atypical antipsychotics' side effects through speech analysis of schizophrenic patients. TALK Study]. Encephale, 2011. 37 Suppl 2: p. S143-50.

42. Sinha, P., et al., Predictors of Effect of Atypical Antipsychotics on Speech. Indian J Psychol Med, 2015. 37(4): p. 429-33.

43. 백연숙 외, 우울증 화자 음성의 음향음성학적 특징, 한국음성학회 2012, 4권 1호, p. 91~98.

44. Handelman, L.D. and D. Lester, The content of suicide notes from attempters and completers. Crisis, 2007. 28(2): p. 102-4.

45. Sawhney, R., et al., Robust suicide risk assessment on social media via deep adversarial learning. J Am Med Inform Assoc, 2021. 28(7): p. 1497-1506.

46. Parrott, S., et al., Social Media and Suicide: A Validation of Terms to Help Identify Suicide-related Social Media Posts. J Evid Based Soc Work (2019), 2020. 17(5): p. 624-634.

47. Al Asad, N., Pranto, M. A. M., Afreen, S., & Islam, M. M. (2019, November). Depression detection by analyzing social media posts

of user. In 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON) (pp. 13-17). IEEE.

48. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences, 18, 43-49.

49. Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017, October). Hybrid depression classification and estimation from audio video and text information. In Proceedings of the 7th annual workshop on audio/visual emotion challenge (pp. 45-51).

50. 김은이, & 고은정. (2018). 텍스트 및 영상의 멀티모달분석을 이용한 트위터 사용자의 감성 흐름 모니터링 기술. 한국융합학회논문지, 9(1), 419-431.

51. Amisha, P. M., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. Journal of family medicine and primary care, 8(7), 2328.

52. Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., ... & Lobach, D. (2012). Effect of clinical decision-support systems: a systematic review. Annals of internal medicine, 157(1), 29-43.

53. Kotze, B., & Brdaroska, B. (2004). Clinical decision support systems in psychiatry in the information age. Australasian Psychiatry, 12(4), 361-364.

54. Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC medicine, 11(1), 1-8.
55. Wells, K.B.; Burnam, M.A.; Rogers, W.; Hays, R.; Camp, P. The course of depression in adult outpatients: Results from the Medical Outcomes Study. Arch. Gen. Psychiatry 1992, 49, 788-794.
56. Wagner, H.R.; Burns, B.J.; Broadhead, W.E.; Yarnall, K.S.H.; Sigmon, A.; Gaynes, B. Minor depression in family practice: Functional morbidity, co-morbidity, service utilization and outcomes. Psychol. Med. 2000, 30, 1377-1390, doi:10.1017/s0033291799002998.
57. Davidson, J.R.; Meltzer-Brody, S. The underrecognition and undertreatment of depression: What is the breadth and depth of the problem? J. Clin. Psychiatry 1999, 60, 4-11.
58. Hall, R.C.; Wise, M.G. The Clinical and Financial Burden of Mood Disorders. Psychosomatics 1995, 36, S11-S18, doi:10.1016/s0033-3182(95)71699-1.
59. Cuijpers, P.; Smit, F. Subthreshold depression as a risk indicator for major depressive disorder: A systematic review of prospective studies. Acta Psychiatr. Scand. 2004, 109, 325-331.
60. Rodríguez, M.R.; Nuevo, R.; Chatterji, S.; Ayuso-Mateos, J.L. Definitions and factors associated with subthreshold depressive conditions: A systematic review. BMC Psychiatry 2012, 12, 181,

doi:10.1186/1471-244x-12-181.

61. Rapaport, M.H.; Judd, L.L. Minor depressive disorder and subsyndromal depressive symptoms: Functional impairment and response to treatment. J. Affect. Disord. 1998, 48, 227-232.

62. Morales, M. R., & Levitan, R. (2016, December). Speech vs. text: A comparative analysis of features for depression detection systems. In 2016 IEEE spoken language technology workshop (SLT) (pp. 136-143). IEEE.

63. Jeon HJ. Depression and suicide. J Korean Med Assoc 2011;54:370-375.

64. Isometsa ET. Psychological autopsy studies: a review. Eur Psychiatry 2001;16:379-385.

65. Tiihonen, J.; Mittendorfer-Rutz, E.; Torniainen, M.; Alexanderson, K.; Tanskanen, A. Mortality and Cumulative Exposure to Antipsychotics, Antidepressants, and Benzodiazepines in Patients with Schizophrenia: An Observational Follow-Up Study. Am. J. Psychiatry 2016, 173, 600-606, doi:10.1176/appi.ajp.2015.15050618.

66. Leucht, S.; Samara, M.; Heres, S.; Patel, M.X.; Furukawa, T.; Cipriani, A.; Geddes, J.; Davis, J.M. Dose Equivalents for Second-Generation Antipsychotic Drugs: The Classical Mean Dose Method. Schizophr. Bull. 2015, 41, 1397-1402, doi:10.1093/schbul/sbv037.

67. Sheehan, D.V.; Lecrubier, Y.; Sheehan, K.H.; Amorim, P.; Janavs,

J.; Weiller, E.; Hergueta, T.; Baker, R.; Dunbar, G.C. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 1998, 59.

68. Yoo, S.W.; Kim, Y.S.; Noh, J.S.; Oh, K.S.; Kim, C.H.; NamKoong, K.; Chae, J.H.; Lee, G.C.; Jeon, S.I.; Min, K.J.; et al. Validity of Korean version of the mini-international neuropsychiatric interview. Anxiety Mood 2006, 2, 50-55.

69. Hamilton, M. Development of a Rating Scale for Primary Depressive Illness. Br. J. Soc. Clin. Psychol. 1967, 6, 278-296, doi:10.1111/j.2044-8260.1967.tb00530.x.

70. Bobo, W.V.; Angleró, G.C.; Jenkins, G.; Hall-Flavin, D.K.; Weinshilboum, R.; Biernacka, J.M. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: Analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. Hum. Psychopharmacol. 2016, 31, 185-192.

71. Zimmerman, M.; Martinez, J.H.; Young, D.; Chelminski, I.; Dalrymple, K. Severity classification on the Hamilton depression rating scale. J. Affect. Disord. 2013, 150, 384-388, doi:10.1016/j.jad.2013.04.028.

72. Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. J. Gen. Intern. Med. 2001, 16, 606-613.
73. Hirschtritt, M.E.; Kroenke, K. Screening for Depression. JAMA 2017, 318, 745-746, doi:10.1001/jama.2017.9820.
74. Levis, B.; Benedetti, A.; Thombs, B.D. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. BMJ 2019, 365, l1476, doi:10.1136/bmj.l1476.
75. Beck, A.T.; Steer, R.A. Relationship between the Beck Anxiety Inventory and the Hamilton Anxiety Rating-Scale with Anxious Outpatients. J. Anxiety Disord. 1991, 5, 213-223.
76. Bardhoshi, G.; Duncan, K.; Erford, B.T. Psychometric Meta-Analysis of the English Version of the Beck Anxiety Inventory. J. Couns. Dev. 2016, 94, 356-373, doi:10.1002/jcad.12090.
77. Dolan, M.; Anderson, I.M.; Deakin, J. Relationship between 5-HT function and impulsivity and aggression in male offenders with personality disorders. Br. J. Psychiatry 2001, 178, 352-359.
78. Barratt, E.S. Anxiety and impulsiveness related to psychomotor efficiency. Percept. Mot. Ski. 1959, 9, 191-198.
79. Spinella, M. Normative data and a short form of the barratt impulsiveness scale. Int. J. Neurosci. 2007, 117, 359-368, doi:10.1080/00207450600588881.
80. Lee, S.-R.; Lee, W.-H.; Park, J.-S.; Kim, S.-M.; Kim, J.-W.;

Shim, J.-H. The Study on Reliability and Validity of Korean Version of the Barratt Impulsiveness Scale-11-Revised in Nonclinical Adult Subjects. J. Korean Neuropsychiatr. Assoc. 2012, 51, 378-386, doi:10.4306/jknpa.2012.51.6.378.

81. Moore, E., 2nd; Clements, M.A.; Peifer, J.W.; Weisser, L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans. Biomed. Eng. 2008, 55, 96-107.

82. Belalcázar-Bolaños, E.A.; Orozco-Arroyave, J.R.; Vargas-Bonilla, J.F.; Haderlein, T.; Nöth, E. Glottal Flow Patterns Analyses for Parkinson's Disease Detection: Acoustic and Nonlinear Approaches. In Proceedings of the Transactions on Petri Nets and Other Models of Concurrency XV; Springer: London New York 2016; pp. 400-407.

83. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th python in science conference 2015, Austin, TX, USA, 6-12 July 2015; pp. 18-24, doi:10.25080/majora-7b98e3ed-003.

84. Snell, R.; Milinazzo, F. Formant location from LPC analysis data. IEEE Trans. Speech Audio Process. 1993, 1, 129-134, doi:10.1109/89.222882.

85. Bachu, R.G.; Kopparthi, S.; Adapa, B.; Barkana, B.D. Separation of voiced and unvoiced using zero crossing rate and energy of the

speech signal. In American Society for Engineering Education (ASEE) Zone Conference Proceedings; 2008.

86. Mardia, K.V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā. Indian J. Stat. Ser. B 1974, 115–128.

87. Vapnik, V.; Golowich, S.E.; Smola, A.J. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In Advances in Neural Information Processing Systems 9; Mozer, M.C., Jordan, M.I., Petsche, T., Eds.; MIT Press: Boston, MA, USA, 1997; pp. 281–287.

88. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Internal Representations by Error Propagation; California Univ San Diego La Jolla Inst for Cognitive Science: San Diego, CA, USA, 1985.

89. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

90. Chollet, F. Keras: The python deep learning library. Astrophys. Source Code Libr. 2018, ascl-1806.022.

91. Schubert, E.; Wolfe, J.; Tarnopolsky, A. Spectral centroid and timbre in complex, multiple instrumental textures. In Proceedings of the international conference on music perception and cognition, North Western University, Evanston, IL, USA, 2004.

92. Kim, J.-C. Detection of the Optimum Spectral Roll-off Point using

Violin as a Sound Source. J. Korea Soc. Comput. Inf. 2007, 12, 51–56.

93. Standards Secretariat, A.S.o.A. American National Standard Acoustical Terminology; Acoustical Society of America: Melville, NY, USA, 1994.

94. Gouyon, F.; Pachet, F.; Delerue, O. On the use of zero-crossing rate for an application of classification of percussive sounds. In Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy, 7–9 December 2000.

95. Nilsonne, Å.; Sundberg, J.; Ternström, S.; Askenfelt, A. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. 1988, 83, 716–728.

96. Darby, J.K.; Hollien, H. Vocal and Speech Patterns of Depressive Patients. Folia Phoniatr. Logop. 1977, 29, 279–291, doi:10.1159/000264098.

97. Wang, J.; Zhang, L.; Liu, T.; Pan, W.; Hu, B.; Zhu, T. Acoustic differences between healthy and depressed people: A cross-situation study. BMC Psychiatry 2019, 19, 1–12, doi:10.1186/s12888-019-2300-7.

98. Taguchi, T.; Tachikawa, H.; Nemoto, K.; Suzuki, M.; Nagano, T.; Tachibana, R.; Nishimura, M.; Arai, T. Major depressive disorder discrimination using vocal acoustic features. J. Affect. Disord. 2018, 225, 214–220.

99. Nilsonne, Å.; Sundberg, J.; Ternström, S.; Askenfelt, A. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. 1988, 83, 716–728.

100. Darby, J.K.; Hollien, H. Vocal and Speech Patterns of Depressive Patients. Folia Phoniatr. Logop. 1977, 29, 279–291, doi:10.1159/000264098.

101. Mundt, J.C.; Snyder, P.; Cannizzaro, M.S.; Chappie, K.; Geralts, D.S. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguistics 2007, 20, 50–64, doi:10.1016/j.jneuroling.2006.04.001.

102. Cannizzaro, M.; Harel, B.; Reilly, N.; Chappell, P.; Snyder, P.J. Voice acoustical measurement of the severity of major depression. Brain Cogn. 2004, 56, 30–35, doi:10.1016/j.bandc.2004.05.003.

103. Meeks, T.W.; Vahia, I,V.; Lavretsky, H.; Kulkarni, G.; Jeste, D.V. A tune in "a minor" can "b major": A review of epidemiology, illness course, and public health implications of subthreshold depression in older adults. J. Affect. Disord. 2011, 129, 126–142.

104. Judd, L.L.; Akiskal, H.S.; Maser, J.D.; Zeller, P.J.; Endicott, J.; Coryell, W.; Paulus, M.; Kunovac, J.L.; Leon, A.C.; I Mueller, T.; et al. Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. J. Affect. Disord. 1998, 50, 97–108, doi:10.1016/s0165–

0327(98)00138-4.

105. Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; Stoyanov, V. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); Association for Computational Linguistics (ACL): 2016; pp. 1-18.

106. Van Hee, C.; Lefever, E.; Hoste, V. SemEval-2018 Task 3: Irony Detection in English Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5-6 June 2018; Association for Computational Linguistics: 2018; pp. 39-50.

107. Beck, A.T., M. Kovacs, and A. Weissman, Assessment of suicidal intention: the Scale for Suicide Ideation. Journal of consulting and clinical psychology, 1979. 47(2): p. 343

108. Lai, K., et al., Assessing Suicide Reporting in Top Newspaper Social Media Accounts in China: Content Analysis Study. JMIR Ment Health, 2021. 8(5): p. e26654.

109. Range, L.M. and E.C. Knott, Twenty suicide assessment instruments: evaluation and recommendations. Death Stud, 1997. 21(1): p. 25-58.

110. West, S.G., J.F. Finch, and P.J. Curran, Structural equation models with nonnormal variables: Problems and remedies. 1995.

111. Jiang, L., Cao, Y., Ni, S., Chen, X., Shen, M., Lv, H., & Hu, J. (2020). Association of sedentary behavior with anxiety,

depression, and suicide ideation in college students. Frontiers in psychiatry, 11, 1403.

112. Oon-Arom, A., Wongpakaran, T., Kuntawong, P., & Wongpakaran, N. (2021). Attachment anxiety, depression, and perceived social support: a moderated mediation model of suicide ideation among the elderly. International psychogeriatrics, 33(2), 169-178.

113. "Linear & Quadratic Discriminant Analysis · UC Business Analytics R Programming Guide". uc-r.github.io. Retrieved 2020-03-29.

114. Tharwat, Alaa (2016). "Linear vs. quadratic discriminant analysis classifier: a tutorial". International Journal of Applied Pattern Recognition. 3 (2): 145. doi:10.1504/IJAPR.2016.079050. ISSN 2049-887X.

115. Sun, Y., Liu, Z., Todorovic, S., & Li, J. (2007). Adaptive boosting for SAR automatic target recognition. IEEE Transactions on Aerospace and Electronic Systems, 43(1), 112-125.

116. Tiku, M., A new statistic for testing for normality. Communications in Statistics-Theory and Methods, 1974. 3(3): p. 223-232.

117. Ci, B. and R.-O. Rule, Confidence intervals. Lancet, 1987. 1(8531): p. 494-7.

118. Öztuna, D., A.H. Elhan, and E. Tüccar, Investigation of four different normality tests in terms of type 1 error rate and power

under different distributions. Turkish Journal of Medical Sciences, 2006. 36(3): p. 171-176.

119. Ghasemi, A. and S. Zahediasl, Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 2012. 10(2): p. 486.

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

121. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.

122. Fant, G. (1962). Formant bandwidth data. STL-QPSR, 3, 1-2.

123. Esposito, A., Esposito, A. M., Likforman-Sulem, L., Maldonato, M. N., & Vinciarelli, A. (2016). On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives. In Recent advances in nonlinear speech processing (pp. 73-82). Springer, Cham.

124. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. Journal of clinical psychiatry,

59(20), 22-33.

125. Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In Twelfth Annual Conference of the International Speech Communication Association.

126. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., & Parker, G. (2013, May). A comparative study of different classifiers for detecting depression from spontaneous speech. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8022-8026). IEEE.

127. Wang, Z., Chen, L., Wang, L., & Diao, G. (2020). Recognition of audio depression based on convolutional neural network and generative antagonism network model. IEEE Access, 8, 101181-101191.

128. He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics, 83, 103-111.

129. General, S. (2012). National Strategy for Suicide Prevention: Goals and Objectives for Action. US Department of Health and Human Services Office of the Surgeon General and National Action Alliance for Suicide Prevention; 2012.[Accessed: July 29, 2015].

130. Cochrane-Brink, K. A., Lofchy, J. S., & Sakinofsky, I. (2000). Clinical rating scales in suicide risk assessment. General Hospital

Psychiatry, 22(6), 445-451.

131. Agoub, M., Moussaoui, D., & Kadri, N. (2006). Assessment of suicidality in a Moroccan metropolitan area. Journal of Affective Disorders, 90(2-3), 223-226.

132. Lim, A. Y., Lee, A. R., Hatim, A., Tian-Mei, S., Liu, C. Y., Jeon, H. J., ... & Hong, J. P. (2014). Clinical and sociodemographic correlates of suicidality in patients with major depressive disorder from six Asian countries. BMC psychiatry, 14(1), 1-8.

133. 유재순, 손정우, & 남민선. (2010). 자살 위험성이 높은 청소년을 대상으로 한 우울중재 및 자살예방 프로그램의 효과. 지역사회간호학회지, 21(1), 71-81.

134. Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Frontiers in public health, 5, 307.

135. Hamdan, A.L., et al., Effect of Hormonal Replacement Therapy on Voice. J Voice, 2018. 32(1): p. 116-121.

136. Naushad, N., Dunn, L. B., Muñoz, R. F., & Leykin, Y. (2018). Depression increases subjective stigma of chronic pain. Journal of affective disorders, 229, 456-462.

137. Blazer, D. G., Hughes, D. C., & George, L. K. (1992). Age and impaired subjective support: Predictors of depressive symptoms at one-year follow-up. Journal of Nervous and Mental Disease.

138. Tsai, S. Y., Lee, J. C., & Chen, C. C. (1999). Characteristics and psychosocial problems of patients with bipolar disorder at

high risk for suicide attempt. Journal of Affective Disorders, 52(1-3), 145-152.

139. Mannuzza, S., Aronowitz, B., Chapman, T., Klein, D. F., & Fyer, A. J. (1992). Panic disorder and suicide attempts. Journal of Anxiety Disorders, 6(3), 261-274.

140. Johnson, J., Weissman, M. M., & Klerman, G. L. (1990). Panic disorder, comorbidity, and suicide attempts. Archives of General Psychiatry, 47(9), 805-808.

141. Davidson, C. L., Wingate, L. R., Grant, D. M., Judah, M. R., & Mills, A. C. (2011). Interpersonal suicide risk and ideation: The influence of depression and social anxiety. Journal of Social and Clinical Psychology, 30(8), 842-855.

142. Davidson, C. L., Wingate, L. R., Grant, D. M., Judah, M. R., & Mills, A. C. (2011). Interpersonal suicide risk and ideation: The influence of depression and social anxiety. Journal of Social and Clinical Psychology, 30(8), 842-855.

143. Sanislow, C. A., Grilo, C. M., Fehon, D. C., Axelrod, S. R., & McGlashan, T. H. (2003). Correlates of suicide risk in juvenile detainees and adolescent inpatients. Journal of the American Academy of Child & Adolescent Psychiatry, 42(2), 234-240.

144. Taylor, J., Kemper, T. S., Loney, B. R., & Kistner, J. A. (2006). Classification of severe male juvenile offenders using the MACI clinical and personality scales. Journal of Clinical Child and Adolescent Psychology, 35(1), 90-102.

145. Carli, V., Jovanović, N., Podlešek, A., Roy, A., Rihmer, Z., Maggi, S., ... & Sarchiapone, M. (2010). The role of impulsivity in self-mutilators, suicide ideators and suicide attempters—A study of 1265 male incarcerated individuals. Journal of affective disorders, 123(1-3), 116-122.

146. Barratt, E. S. (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. Psychological reports, 16(2), 547-554.

147. Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. Journal of clinical psychology, 51(6), 768-774.

148. Nicolai, J., Darancó, S., & Moshagen, M. (2016). Effects of mood state on impulsivity in pathological buying. Psychiatry Research, 244, 351-356.

149. Wingrove, J., & Bond, A. J. (1997). Impulsivity: a state as well as trait variable. Does mood awareness explain low correlations between trait and behavioural measures of impulsivity?. Personality and Individual Differences, 22(3), 333-339.

150. Croarkin, P. E., Levinson, A. J., & Daskalakis, Z. J. (2011). Evidence for GABAergic inhibitory deficits in major depressive disorder. Neuroscience & Biobehavioral Reviews, 35(3), 818-825.

151. Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. Psychological bulletin, 99(2), 143.

152. Ozdas, A., Shiavi, R. G., Silverman, S. E., Silverman, M. K., & Wilkes, D. M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. Ieee transactions on Biomedical engineering, 51(9), 1530-1540.

153. Nilsonne, Å., Sundberg, J., Ternström, S., & Askenfelt, A. (1988). Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. The Journal of the Acoustical Society of America, 83(2), 716-728.

154. Breznitz, Z. (1992). Verbal indicators of depression. The Journal of general psychology, 119(4), 351-363.

155. Hönig, F., Batliner, A., Nöth, E., Schnieder, S., & Krajewski, J. (2014). Automatic modelling of depressed speech: relevant features and relevance of gender.

156. France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. IEEE transactions on Biomedical Engineering, 47(7), 829-837.

157. Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. Psychological bulletin, 99(2), 143.

158. Gyllensten, A. L., Öberg, H., Träskman-Bendz, L., & Ekdahl,

C. (1997). Psychomotor functioning in suicide attempters: An explorative study using the Resource Oriented Body Examination of Bunkan. Nordic Journal of Psychiatry, 51(3), 193-200.

159. McCallie, M. S., Blum, C. M., & Hood, C. J. (2006). Progressive muscle relaxation. Journal of human behavior in the social environment, 13(3), 51-66.

160. Baumeister, R. F. (1990). Suicide as escape from self. Psychological review, 97(1), 90.

161. Jansson, L., & Nordgaard, J. (2016). The psychiatric interview for differential diagnosis (Vol. 270). Switzerland: springer.

162. Courts, F. A. (1942). Relations between muscular tension and performance. Psychological Bulletin, 39(6), 347.

163. Lee, J. M., Cho, B. H., Ku, J. H., Kim, J. S., Lee, J. H., Kim, I. Y., & Kim, S. I. (2001, October). A study on the system for treatment of ADHD using virtual reality. In 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Vol. 4, pp. 3754-3757). IEEE.

164. Stawarz, K., Preist, C., Tallon, D., Thomas, L., Turner, K., Wiles, N., ... & Coyle, D. (2020, April). Integrating the Digital and the Traditional to Deliver Therapy for Depression: Lessons from a Pragmatic Study. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Abstract

A study on artificial intelligence-based clinical decision support system to evaluate depression and suicide risk using voice and text

Daun Shin

Department of Medicine (Psychiatry)

The Graduate School

Seoul National University

**Introduction:** The incidence of depression and suicide continues to increase worldwide, and the resulting socio-economic loss is also enormous. However, in diagnosing depression and suicide, it is difficult to effectively intervene due to a misdiagnosis in situations where it is difficult for the patient to report symptoms with reduced symptoms or conduct an in-depth interview because the diagnosis must be made through the patient's subjective answer. In addition, voices and text used by participants in interviews are traditionally known for their clinical significance in the Department of Psychiatry. In the past, the voice and the subject's utterance content had to be used for diagnosis based on the clinical experience accumulated by clinicians. As technologies for extracting various indicators of voice and extracting spoken words have been developed recently, differences in voices and differences in spoken words according to the risk of depression and suicide are being revealed. In addition, as

machine learning is applied to the medical field, it has been able to detect these subtle differences and help make decisions. However, previous studies have limitations in that the number of subjects is not yet sufficient and consideration of various clinical situations such as drugs is insufficient. Therefore, this study aims to overcome the existing limitations and build an AI-based clinical decision support system that evaluates depression and suicide risk groups based on the subject's voice and words used during the interview.

**Method:** A patient group complaining of depressive symptoms and a normal control group were recruited, respectively, and a Mini International Neuropsychiatric Interview was performed on all subjects and the interview was recorded. After extracting only the section uttered by the subject from among the recorded interview files, various indicators of voice and text data were extracted from each section. In Study I, the initial data were used to compare and analyze the normal group, the mild depression group, and the major depression group. In Study II, the final initial evaluation data was used to identify voice indicators that can distinguish normal and depression, and a diagnostic algorithm using text was constructed. In Study III, an algorithm for diagnosing the high-risk group for suicide was established by defining the high-risk group for depression through the Beck's suicide ideation scale and the suicide-related module of the mini international neuropsychiatric interview.

**Results:** In Study I, 7 voice and speech indicators that can distinguish

33 normal group, 26 mild depression group, and 34 major depression group were extracted. When the accuracy of various models was confirmed based on the voice and speech features, it was confirmed that the performance of the multilayer perceptron was the best. In Study II, the speech features and text data of 83 normal and 83 depressed patients were analyzed, and the area under the curve of the machine learning algorithm based on each was 0.806 and 0.905. In Study III, 83 people in the depression group were compared with both the Beck's suicide ideation scale and the classification method through a mini international neuropsychiatric interview. However, in the case of the algorithm for predicting the suicide high risk group based on voice, the maximum sensitivity was 0.535, and the best performance was the average accuracy of 0.495 in the model using the logistic regression formula. On the other hand, the algorithm for predicting suicide risk based on text also had an area under the curve of only 0.632, but the ensemble model built by integrating text data and sociodemographic information was able to confirm the diagnostic usefulness with an area under the curve of 0.800.

Conclusion: This study established an algorithm for diagnosing depression and suicide risk by extracting voice and speech features and text data from the subject's utterance section in a structured interview. Both data showed excellent performance in diagnosing depression, but insufficient for diagnosing suicide risk. In the case of text data, although there are limitations as data obtained through

structured interviews, when it was integrated with sociodemographic information using statistical techniques, it showed better performance than predicted by sociodemographic information alone, showing utility value. This study is the first study in South Korea to prove the objective diagnostic value of voice and text data, and it is a challenge to a new field of psychiatry as a digital objective diagnostic tool. In the future, additional research is needed on data from more diverse regions and environments in the field.

Keywords : depression, suicide risk, voice, text analysis, machine learning, clinical decision support system
Student Number : 2018-25300