A depresszió hang alapú felismerésének optimalizációja hangfelvétel hossz alapján

Pašić Azra[1], Kiss Gábor[2], Sztahó Dávid[2]

[1]Karlsruhe Institute of Technology
[2]Budapest Műszaki és Gazdaságtudományi Egyetem
uwhvy@student.kit.edu, {kiss.gabor, sztaho}@tmit.bme.hu

Kivonat A depresszió komoly hangulatzavar, amely világszerte már a lakosság több mint 3%-át érinti, és ez a szám feltehetően tovább fog nőni az elkövetkezendő években, évtizedekben. A depresszió diagnosztizálása maga is egy komoly feladat, amely jelenleg kizárólag a terület szakembereire hárul, akikből pedig egész bizonyosan nincs elég. Ebben a helyzetben nagy jelentőséggel bírhat egy olyan automatizált depresszió felismerési rendszer bevezetése, amely nagymértékben asszisztálni tudná a szakemberek munkáját a diagnosztizálás során. E cikkben bemutatunk egy, a depresszió osztályozására fejlesztett hang-alapú felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, a jellemzőkiválasztást és a szupport vektor gépek hiperparaméter-optimalizációját. Természetesen, a hang-alapú modellhez szükséges egy optimális hangfelvétel hossz meghatározása is, mely kompromisszumot jelent a felismerőrendszer igényei és a páciensek kényelme között. A modell hatékonyságát különböző hosszúságú felvételeken vizsgáltuk, hogy belátást nyerjünk abba, hogy a felvétel-hossz miként és milyen mértékben befolyásolja a felismerés pontosságát.
Kulcsszavak: depresszió, beszédjel alapú detektálás, osztályozás, szupport vektor klasszifikáció

1. Bevezetés

A súlyos depresszív zavar (legtöbbször csak „depresszió”) olyan mentális zavar, amely a levertség, reménytelenség, szorongás és kitartó szomorúság tüneteivel jár (Association et al., 2013) (Cummins et al., 2015). Világszerte már a lakosság több mint 3%-át érinti (Andrade et al., 2003), és ez a szám feltehetően tovább fog nőni az elkövetkezendő években, évtizedekben. A betegség hatása az érintettek életminőségére olyan krónikus megbetegedésekhez lett hasonlítva mint a cukorbetegség és a magas vérnyomás (Hays et al., 1995). Ezen kívül pedig a depressziós betegeknél húszszor nagyobb az esély az öngyilkosságra mint az egészséges lakosságnál (Lépine and Briley, 2011). Mindezek ellenére a depresszió nagyon is kezelhető betegségnek számít, de ehhez szükséges az időszerű felismerés. Gyógyulás után is érdemes a korábbi betegekkel foglalkozni, mivel a visszaesés veszélye nagy, és az első depressziós epizódtól szenvedők 80%-a legalább még egyet tapasztal élete folyamán (Lépine and Briley, 2011).

Mivel a depresszió diagnosztizálása és szűrése is szakemberekhez van kötve, folytonos a pszichológus és pszichiáter hiány, ami ahhoz is vezet, hogy a depressziós betegek nagy része nem is kerül felismerésre (Lépine and Briley, 2011) — annak ellenére, hogy a kezelés elmaradása megötszöröz az öngyilkosság esélyét (Strakowski and Nelson, 2015). Ebben a helyzetben nagy jelentőséggel bírhat egy olyan automatizált depresszió felismerési rendszer bevezetése, amely nagymértékben asszisztálni tudná a szakemberek munkáját a diagnosztizálás során. A diagnosztikai eljárásban az orvos megfigyeli a betegnek a kinézetén, a viselkedésén és a hangulatán kívül a beszédét — ezen belül pedig a hangját, hangzását is (Association et al., 2013). Ebből kifolyólag a depresszió automatikus felismerése hang alapján sokat ígérő ötlet. A depresszió és a beszéd kapcsolata már az 1980-as évektől kutatott, és több akusztikai illetve fonetikai paramétert kapcsolatba hoztak a depresszióval (Nilsonne, 1988).

E cikkben bemutatunk egy, a depresszió osztályozására fejlesztett hang-alapú felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, a jellemzőkiválasztást és a szupport vektor gépek hiperparaméter-optimalizációját. Természetesen, a hang-alapú modellhez szükséges egy optimális hangfelvétel hossz meghatározása is, mely kompromisszumot jelent a felismerő-rendszer igényei és a páciensek kényelme között (Rutowski et al., 2019). A modell hatékonyságát különböző hosszúságú felvételeken vizsgáltuk, hogy belátást nyerjünk abba, hogy a felvétel-hossz miként és milyen mértékben befolyásolja a felismerés pontosságát. Gépi tanulással kétféle felismerés valósítható meg: az osztályozás, amely a depressziós állapotot becsüli meg, és a regresszió, amely annak a súlyosságáról kísérel információt adni. Ebben a cikkben az osztályozást használtuk, melyet szupport vektor gépekkel valósítottunk meg.

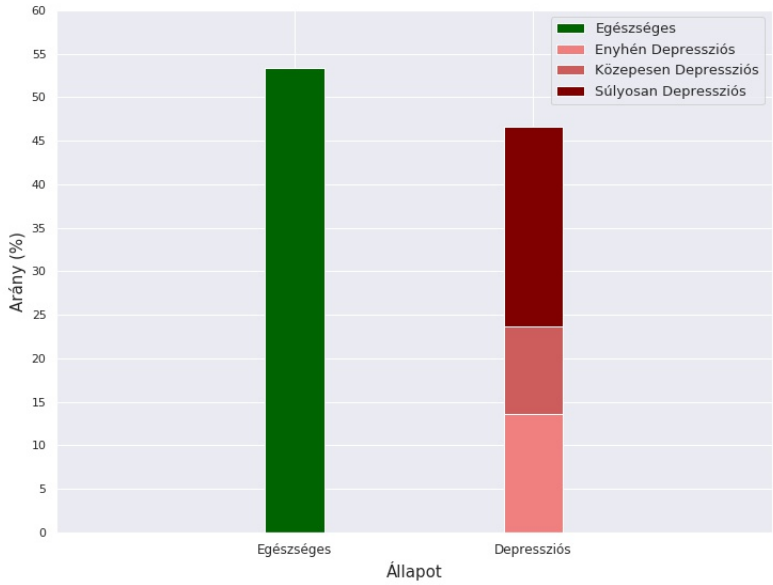
A cikk bevezetés utáni felépítése a következő: először bemutatjuk a beszéd adatbázist amivel dolgoztunk, majd a kutatásban felhasznált módszereinket – ezen belül az előfeldolgozást, a jellemző kinyerést, az osztályozást és a tesztelést is. Ezután következik az eredmények bemutatása és tárgylása, valamint az összegzés és a konklúzió.

2. Adatbázis

A beszédminták gyűjtése a Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinikájával együtt lett végezve. A beszélők lefedik a depresszió súlyosságának különböző fokozatait, az egészséges állapottól az egészen súlyos depresszióig. A vizsgált személyek egy fonetikusan kiegyensúlyozott mesét („Az északi szél és a nap") olvastak fel, amely széles körben elterjedt a hasonló vizsgálatokban. A felvételek csendes helyiségben kerültek rögzítésre, 44.1 kHz mintavételi frekvenciával. Az adatbázisba gyűjtött felvételekhez el lett készítve a fonéma szintű szegmentálás, a labor által fejlesztett automatikus szegmentáló program segítségével (Kiss et al., 2013).

A depresszió súlyossága is minden esetben rögzítésre került – a két legelterjedtebb skála a Hamilton Rating Scale for Depression (HAMD) (Williams, 1988) és a Beck Depression Index (BDI) (Beck et al., 1996). Mi a BDI továbbfejlesztett

változatát használtuk, a BDI-II skálát (Beck et al., 1996). A BDI-II skála pontszámaihoz a következő besorolás adott: 0-13 egészséges, 14-19 enyhe depresszió, 20-28 közepes depresszió, 29-63 súlyos depresszió. A BDI pontszámok 0-tól 50-ig fordultak elő az adatbázisban. Az adatbázis 118 hangfelvételt tartalmazott, ebből 55 depressziós és 63 egészséges mintát. A különböző súlyosságok előfordulása az 1. ábrában adott. A vizsgált személyek átlagéletkora 42,5 év (min.: 20; max.: 70; std: 14,5).



1. ábra: Az egészséges és depressziós minták eloszlása az adatbázisban

3. Módszerek

A jellemző-kinyerés Python 2.7 programmal lett végezve (Python, 2007). A librosa és soundfile csomagok a felvételek kezeléséhez és az akusztikai jellemzők kinyeréséhez lettek felhasználva (McFee et al., 2015). További jellemzők a parselmouth (Jadoul et al., 2018) csomaggal kerültek kinyerésre, amely a Praat program C++ kódjából kinyert Python változata (Boersma et al., 2002). A parselmouth-tal együtt lett használva a tgt csomag, amely a Praat által generált Textgrid fájlok (ezek tartalmazzák a szegmentálást) kezeléséhez volt szükséges (Buschmeier and Wlodarczak, 2013). A különböző klasszifikációs modellek a

LibSVM könyvtárral lettek felépítve (Chang and Lin, 2011). A hiperparaméter optimalizáció Grid Search algoritmussal lett végezve, amely a lehető paraméterkombinációkból a legjobbat választja ki.

3.1. Előfeldolgozás

A felvételek először 16 kHz-en újra lettek mintavételezve. A BDI-II pontszámuk alapján a minták a depressziós és egészséges csoportokba lettek sorolva és az alapján felcímkézve. Ezt követően a szegmentálás segítségével a felvételek három részre lettek osztva, majd ezekből lett képezve a három vizsgált hossz – az egy harmad, két harmad és egész felvétel – méghozzá úgy, hogy csak mondat végén történtek a vágások. Ez azért volt lényeges, mert az idő alapú szeparáció amely nem veszi figyelembe a mondathatárokat torzította volna az akusztikai jellemzőket. Továbbá ez azt is jelenti, hogy az egy harmad és két harmad nem szó szerint értendő (az egy harmad felvétel valamivel rövidebb, mint a két harmad felvétel fele). A hasonló kutatásokban használatos jellemzők alapján ezek a paraméterek kerültek kiszámításra a felvételeken (Kiss and Vicsi, 2017) (Kiss and Vicsi, 2014) (Cummins et al., 2015) (Alghowinem et al., 2013): formáns frekvenciák (F1, F2, F3), mel-skálás spektrogram, mel-frekvenciás kepsztrális együtthatók (MFCC-k, 10 koefficienssel), chromagram, tonal centroid, valamint különböző intenzitás, frekvencia és hangmagasság értékek a Praat-ból (jitter, shimmer, number of voice breaks, fraction of locally unvoiced frames, degree of voice breaks). A jellemzők -1 és 1 közötti értékekre lettek normalizálva.

3.2. Jellemzők kiválasztása

Az algoritmusok pontosságát nagyban befolyásolja a megfelelő jellemzők kiválasztása, vagyis a lényegtelen jellemzők elhagyása. Ez főleg fontos kis mintahalmaz esetén, mint amilyen a miénk is. Az optimális jellemzők Fast Forward Selection-nel kerültek kiválasztásra. Az eljárás során az i-dik lépésben rendelkezésre áll az algoritmus szerint optimális i-1 hosszú jellemzővektor, amihez ezután egyesével hozzá lesznek adva a még fel nem használt jellemzők és k-fold kereszt validáció alapján (default hiperparaméterekkel) az i hosszúságú jellemzővektor közül ki lesz választva az, amely a legnagyobb pontosságot adta (Mao, 2002). Az eljárás hátránya, hogy ha egy lépésben egy jellemző be lesz választva a jellemzőhalmazba, az minden halmazban benne lesz, viszont a jellemző kiválasztás gyors (Mao, 2002).

3.3. Osztályozás

A szupport vektor gépek alapelve, hogy a címkézett példákat (azaz a training készletet) térbeli pontokként jelenítse meg, oly módon, hogy az osztályok a lehető legjobban el legyenek különítve (Cortes and Vapnik, 1995). Ezt követően az új adatpontokat ugyanabba a térbe térképezi fel, és attól függően, hogy az osztályok közötti rés melyik oldalára esnek, a két kategória egyikébe lesznek sorolva (Cortes and Vapnik, 1995). Lineárisan nem szeparálható problémák esetén kernel

függvény segítségével a probléma nagyobb dimenzitású térbe kerül, amelyben szeparálhatóvá alakul (Cortes and Vapnik, 1995). Különböző kernel függvények léteznek, mint például a polinomiális, a szigmoid és a radiális (Cortes and Vapnik, 1995). A kutatás során c-SVC algoritmust radiális (Radial Basis Function) kernellel használtunk, különböző gamma együtthatókkal és C értékekkel (a C határozza meg az osztályok minél nagyobb elkülönülésének és a hibás oldalra eső minták számának a trade-off-ját). Ezek a hiperparaméterek Grid Search algoritmussal lettek kiválasztva, amely kipróbál minden kombinációt és kiválasztja a legjobban teljesítő hiperparaméter-párt.

3.4. A tesztelési eljárás

Az adatbázis alacsony mintaszáma miatt az ebben az esetben szokásos k-fold keresztvalidáció (k-Fold Cross Validation) (Kohavi et al., 1995) lett használva a tesztelések során (mint ahogy az FFS és a Grid Search során is). A keresztvalidációs eljárás a mintahalmazt k egyenlő részre osztja, majd mindegyik csoportot egyszer teszthalmazként használ, a megmaradó részeket $(k-1)$ pedig tanítóhalmazként. A teszthalmazokon kapott eredmények átlaga jellemzi az egész rendszer pontosságát. A modell jellemzésére tévesztési mátrixokat is bemutatunk, amelyekből kivehető, hogy az egészséges és a depressziós mintákat külön-külön mennyire jól ismeri fel a modellünk.

4. Eredmények

A kísérleteket egy harmad, két harmad és egész felvételeken végeztük, a jellemzőkinyerés és normalizálás után a jellemzővektorokat Fast Forward Selection-nel kaptuk meg, majd ezeken tanítva a Grid Search-et megtaláltuk az optimális hiperparamétereket a szupport vektor osztályozáshoz. A tesztelési eljárás során minden esetben 10 részre osztottuk az adathalmazt, és a teljesítmény értékeléséhez a pontosságot (a helyesen osztályozott minták számának és az összes minta számának hányadosát) használtuk. A következő táblázatban láthatóak a különböző hosszúságú felvételeken elért pontosságok és a hibásan osztályozott minták számának relatív csökkenése (az egy harmad felvételhez képest).

	Egy harmad	Két harmad	Egész felvétel
Pontosság	88%	90%	92%
Hiba relatív csökkenése	-	17%	33%

1. táblázat. Az elért pontosságok és a hiba relatív csökkenése

Egy harmad felvételen a legjobb paramétereknek bizonyultak a C=1 és g = 0.125. A tíz kiválasztott jellemző között voltak koefficiensek az MFCC-ből, a chromagramból, a mel-skálás spektrogramból, a contrastból, valamint a shimmer,

a number of voice breaks (egymást követő impulzusok közötti hosszabb szünetek száma) és a formáns frekvenciák is. A következő táblázatban láthatóak az egy harmad felvételen elért eredmények tévesztési mátrix formájában.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	92.1%	7.9%
Tényleges depressziós	16.4 %	83.6%

2. táblázat. Az egy harmad felvételen kapott tévesztési mátrix

A két harmad felvételen számított hiperparaméterek kevéssel eltérnek az előbbitől: C=2, g = 0.25. A jellemzőknél azonban nagy a hasonlóság – továbbra is a tíz kiválasztott között volt az MFCC, a chromagram, a mel-skálás spektrogram és a number of voice breaks, de ebben az esetben beválasztásra került egy koefficiens a tonal centroid-ból is. A 3-as számú táblázatban láthatóak az eredmények.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	90.5%	9.5%
Tényleges depressziós	10.9 %	89.1%

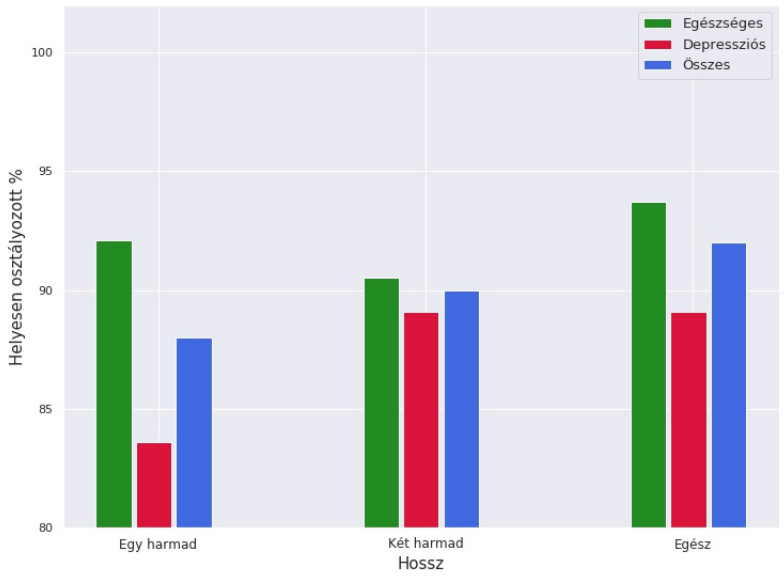
3. táblázat. A két harmad felvételen kapott tévesztési mátrix

Az egész felvételen végzett kísérletnél a C érték 10-nek lett választva az algoritmus által. A kiválasztott jellemzők ugyanazokból a kategóriákból kerültek ki, mint a két harmad felvételen végzett jellemzőválasztás során (MFCC, chromagram, mel-skálás spektrogram, number of voice breaks, tonal centroid). Az eredmények a 4-es számú táblázatban láthatóak.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	93.7%	6.3%
Tényleges depressziós	10.9 %	89.1%

4. táblázat. Az egész felvételen kapott tévesztési mátrix

A különböző tévesztési mátrixokból kivehető, hogy a felismerés pontossága alapvetően javul, ha hosszabb felvételt használunk (ami várható is volt). A két harmad felvételt használva növelni lehetett a depressziósok helyes osztályozását az egy harmad felvételhez képest. Bár megnőtt a hibásan depressziósnak osztályozottak száma, ilyen estekben fontosabb, hogy a ténylegesen betegeket minél jobban felismerjük (továbbá egészében a két harmad felvétel 2%-kal pontosabb volt az egy harmadnál, mint ahogy azt láthattuk az első táblázatban).



2. ábra: Az osztályozás pontossága felvétel-hossz és osztály szerint

Az egészségesek helyes felismerése az egész felvétel használatával javult fel. Az egész felvételen a depressziósok felismerése maradt a két harmad felvétel szintjén, de az egészségeseknek a felismerési pontossága az egy harmadhoz képest is nőtt. Ezekből az adatokból érdekes felvetések is felállíthatóak – bár sok tényező játszik közre, az eredmények alapján feltűnik, hogy a felvételek első és utolsó harmada (eleje és vége) bizonyos okokból kifolyólag az egészségesek felismeréséhez volt fontos, a közepe pedig a depressziósokról rejtett több információt.

Mivel a felvételek egész hossza mindössze 40 másodperc körül mozog és ezzel is 90% körüli pontossággal lehetett következtetni az alanyok állapotára, az egész felvételen elért eredmény az algoritmus és a páciensek igényeit is jó mértékben ötvözi.

5. Összegzés és konklúzió

A cikkben bemutattunk egy, a depresszió osztályozására készített hang-alapú automatikus felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, azoknak a kiválasztását (Fast Forward Selection módszerrel) és a hiperparaméter optimalizációt (Grid Search módszerrel). Az osztályozáshoz szupport vektor klasszifikációt használtunk, radiális kernellel és különböző hiperparaméter-

kombinációkkal. Mindezek az eljárások a kisebb adatbázisokon használatos k-Fold Cross Validation módszerrel lettek becsülve pontosságra.

A kísérletek során azt vizsgáltuk, hogy a felvételek hossza hogyan befolyásolja a rendszerünk teljesítményét. Az adatbázisunkban található eredeti felvételek három részre lettek osztva, mondatok félbeszakítása nélkül, fonéma szegmentálás segítségével. Ebből lettek kialakítva az egy harmad, a két harmad és az egész felvétel csoportjai. A teljesítmények becslésére tévesztési mátrixokat használtunk, amelyek kimutatták a helyesen és hibásan becsült minták százalékát az egészséges és depressziós osztályoknál külön is.

A beválasztott jellemzők alapján a legjobban a mel-skálás spektrogram, a mel-frekvenciás kepsztrális együtthatók, a chromagram, a tonal centroid, valamint a number of voice breaks adja meg a helyes osztályozáshoz szükséges információkat.

A teszt eredmények azt mutatták, hogy minél hosszabb felvételt használtunk, a pontosság teljességében nőtt, két-két százalékkal. A legjobb eredményt az egész felvételen értük el, ahol is 92% pontossággal tudtuk az egészségi állapotot megbecsülni. A két osztály klasszifikációs eredményeit külön-külön tekintve érdekes fejleményeket figyelhettünk meg, miszerint a felvételek eleje és vége leginkább az egészségesek helyes felismeréséhez járult hozzá, a közepe pedig a depressziósokról rejtett több információt. Ennek a felvetésnek a helyességét és esetleges hatását következő munkákban érdemes lehetne komolyabban megvizsgálni.

Köszönetnyilvánítás

A K128568 számú projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, a K pályázati program finanszírozásában valósult meg.

Irodalomjegyzék

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: Detecting depression: a comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7547–7551. IEEE (2013)

Andrade, L., Caraveo-Anduaga, J.J., Berglund, P., Bijl, R.V., Graaf, R.D., Vollebergh, W., Dragomirecka, E., Kohn, R., Keller, M., Kessler, R.C., et al.: The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. International journal of methods in psychiatric research 12(1), 3–21 (2003)

Association, A.P., et al.: Diagnostic and statistical manual of mental disorders. BMC Med 17, 133–137 (2013)

Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F.: Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. Journal of personality assessment 67(3), 588–597 (1996)

Boersma, P., et al.: Praat, a system for doing phonetics by computer. Glot international 5 (2002)

Buschmeier, H., Wlodarczak, M.: TextGridTools: A TextGrid processing and analysis toolkit for Python. In: Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013) (2013)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Communication 71, 10–49 (2015)

Hays, R.D., Wells, K.B., Sherbourne, C.D., Rogers, W., Spritzer, K.: Functioning and well-being outcomes of patients with depression compared with chronic general medical illnesses. Archives of general psychiatry 52(1), 11–19 (1995)

Jadoul, Y., Thompson, B., De Boer, B.: Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics 71, 1–15 (2018)

Kiss, G., Sztahó, D., Vicsi, K.: Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features. In: 2013 IEEE 4th international conference on cognitive infocommunications (CogInfoCom). pp. 579–582. IEEE (2013)

Kiss, G., Vicsi, K.: Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters. In: International conference on statistical language and speech processing. pp. 120–131. Springer (2014)

Kiss, G., Vicsi, K.: Comparison of read and spontaneous speech in case of automatic detection of depression. In: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000213–000218. IEEE (2017)

Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Montreal, Canada (1995)

Lépine, J.P., Briley, M.: The increasing burden of depression. Neuropsychiatric disease and treatment 7(Suppl 1), 3 (2011)

Mao, K.: Fast orthogonal forward selection algorithm for feature subset selection. IEEE Transactions on Neural Networks 13(5), 1218–1224 (2002)

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. vol. 8 (2015)

Nilsonne, A.: Speech characteristics as indicators of depressive illness. Acta Psychiatrica Scandinavica 77(3), 253–263 (1988)

Python, J.: Python programming language. In: USENIX Annual Technical Conference (2007)

Rutowski, T., Harati, A., Lu, Y., Shriberg, E.: Optimizing speech-input length for speaker-independent depression classification. Proc. Interspeech 2019 pp. 3023–3027 (2019)

Strakowski, S., Nelson, E.: Major Depressive Disorder. Oxford University Press (2015)

Williams, J.B.: A structured interview guide for the hamilton depression rating scale. Archives of general psychiatry 45(8), 742–747 (1988)