Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatikai Tanszék

Ramács Gábor

DEPRESSZIÓ EGYNYELVŰ ÉS TÖBBNYELVŰ FELISMERÉSÉNEK LEHETŐSÉGEI X-VEKTOR MÓDSZERREL

TDK dolgozat

KONZULENS

Dr. Kiss Gábor

BUDAPEST, 2022

Tartalomjegyzék

Kivonat

A beszéd az emberi kommunikáció legáltalánosabb eszköze, ami nyelvi kódot közvetít. Ez a kommunikációs forma rendkívül összetett folyamatok összessége, amit számos környezeti, illetve szervezeti tényező befolyásol. Nemcsak érzelmi hatásokra változik meg a beszéd, hanem bizonyos fiziológiai, illetve pszichiátriai tényezők, akár betegségek jelenlétében is. A beszéd mikrofonnal történő felvételével és a beszédproduktum mélyebb vizsgálatával lehetőség nyílhat különböző betegségek felismerésére.

A depresszió korunk egyik leggyakoribb betegsége, amely hatással van a betegségben szenvedő egyén életminőségére és munkavégző képességére is. Bizonyított, hogy a depresszió képes megváltoztatni az egyén beszédproduktumát, emiatt jelenünk egy fontos kutatási területe a depresszió beszéd alapú automatikus felismerése.

Dolgozatomban a főként beszélőfelismerésre használt x-vektor technikát alkalmazva vizsgálom a depresszió beszéd alapú felismerésének lehetőségeit. A rendelkezésemre álló, beszédproduktumokból álló adatbázisok adatpontjaiból egy mély neurális háló segítségével hozok létre x-vektorokat. Ezen x-vektorok alkotják a tanuló eljárásom bemenetét, amely egy szupport vektor regresszió. Vizsgálataim során létre hozok olyan modelleket is, amelyek tanítóhalmazának és teszthalmazának nyelve azonos és olyan modelleket is, amelyeknél különbözőek. Célom az eljárás nyelvfüggőségének vizsgálata ezen modellek eredményeinek összehasonlítása által.

Abstract

Speech is the most common means of human communication, which conveys a language code. This form of communication is a set of extremely complex processes, which are influenced by many environmental and physical factors. Speech changes not only due to emotional influences, but also in the presence of certain physiological or psychiatric factors, or even diseases. By recording the speech with a microphone and examining the speech product in depth, it is possible to recognize various diseases.

Depression is one of the most common diseases of our time, which affects the quality of life and work performance of the individual suffering from disease. It has been proven that depression can change an individual's speech product, which is why an important area of our present research is speech-based automatic recognition of depression.

In my thesis, using the x-vector technique, which is mainly used for speaker recognition, I investigate the possibilities of speech-based recognition of depression. I use a deep neural network to create x-vectors from the data points of the speech product databases at my disposal. These x-vectors form the input of my learning process, which is a support vector regression. During my research, I create models whose language of the training set and test set is the same, as well as models where they are different. My aim is to examine the language dependence of the procedure by comparing the results of these models.

1 Bevezetés

A beszéd az emberi kommunikáció legáltalánosabb eszköze, ami nyelvi kódot közvetít. Ez a kommunikációs forma rendkívül összetett folyamatok összessége, amit számos környezeti, illetve szervezeti tényező befolyásol [1]. Nemcsak érzelmi hatásokra változik meg a beszéd, hanem bizonyos fiziológiai, illetve pszichiátriai tényezők, akár betegségek jelenlétében is. A beszéd mikrofonnal történő felvételével és a beszédprodukt mélyebb vizsgálatával lehetőség nyílhat különböző betegségek felismerésére. Ezzel a beszélő már betegségének akár korai stádiumában megfelelő kezelésben részesülhet, amivel életminőségének romlása visszafordítható [2].

A depresszió korunk egyik leggyakoribb betegsége, amely hatással van a betegségben szenvedő egyén életminőségére és munkavégző képességére is. A WHO becslései szerint 2030-ra az unipoláris depresszió a három legsúlyosabb betegség között lesz nyilvántartva, a HIV/AIDS és a szívproblémák mellett [3]. Súlyos depresszió hatására az egyén elveszítheti motiváltságát, mely hatására a napi teendők elvégzése is problémát okozhat számára, munkájának teljesítésére is alkalmatlanná válhat. Ennek okán a depresszió nem csak az egyén számára jelent problémát, de komoly gazdasági károkat is okoz [4].

Bizonyított, hogy a depresszió képes megváltoztatni az egyén beszédprodukt umát, amely változás informatikai eszközök segítségével számszerűsíthető és mérhető [2, 5]. Különböző beszédleíró jellemzők kinyerése által a beszédprodukt um jellemezhető, különböző beszélők beszédprodukt umai összehasonlíthatók [6]. Ennek okán lehetőség van a depresszió gépi alapú automatikus felismerésére, depresszióval küzdő emberek azonosítására [7]. A depresszió fent már felsorolt káros hatásai miatt jelenünk egy fontos kutatási területe a depresszió beszéd alapú automatikus felismerése.

Mind a beszédfeldolgozás, a jellemzőkinyerés és az osztályozás során is gyakorta alkalmaznak gépi tanuló algoritmusokat [5]. Használatuk nagyban egyszerűsíti és akár hatékonyabbá is teszi a depresszió felismerés, mint probléma megoldását.

Az elmúlt években számos kutatás készült, amelyek vizsgálatainak fókuszában olyan különböző depressziós beszédadatbázisokat felhasználó gépi tanuló eljárások voltak, amelyek célja a depresszió beszéd alapú felismerése volt [5, 8]. Ezen kutatások nagyrésze

egynyelvű modellek felállításával és vizsgálatával foglalkozott, ami azt jelenti, hogy a gépi tanuló eljárások tanítására és tesztelésére is azonos nyelvű beszédadathalmazt használtak. Miközben ez természetesen egy fontos vizsgálati terület, az elmúlt években kutatók elkezdtek olyan eljárásokat is vizsgálni, amelyeknél a tanító- és a teszthalmaz nyelve különböző [9, 10]. Ezen eljárások nagy előnye, hogy egy meglévő eljárás újabb nyelvre való alkalmazásához nem szükséges új minták gyűjtése, új beszédadatbázis készítése. Ez könnyebbé és olcsóbbá tenné jól működő depresszió felismerő eljárások új nyelvek felé való kiterjesztését.

Dolgozatomban bemutatok egy depresszió beszéd alapú felismerésére megtervezett eljárást, amelyet alkalmazok is négy, különböző nyelvű beszédadatbázison. Az általam készített egynyelvű és többnyelvű depresszió felismerő modellek segítségével vizsgálom a depresszió beszéd alapú automatikus felismerésének lehetőségeit.

1.1 Depresszió

A depresszió, mint betegség definíciója megtalálható a The Diagnostic and Statistical Manual of Mental Disorders (DSM) kiadványban. Eszerint egy egyén esetében akkor beszélhetünk depresszióról, ha mély szomorúság, levertség mellett az alábbi tünetek közül legalább 4 fennáll nála legalább 2 héten át [11]:

- fáradékonyság, gyakori energiaszegény napok
- étvágyváltozás, jelentős fogyás vagy hízás
- alvászavar, inszomnia vagy hiperszomnia
- jelentős negatív önértékelés vagy alaptalan túlzott bűntudat
- koncentrációs képesség csökkenése, határozatlanság
- pszichomotoros gátlás vagy remegés
- gyakori gondolatok a halálról vagy gyakori öngyilkos gondolatok

Súlyos depresszió hatására az egyén elveszítheti motiváltságát, mely hatására a napi teendők elvégzése is problémát okozhat számára, munkájának teljesítésére is alkalmatlanná válhat. Ennek okán a depresszió nem csak az egyén számára jelent problémát, de komoly gazdasági károkat is okoz. Egy 2010-es tanulmányban kimutatták, hogy Európa 30 országát vizsgálva összesen 92 milliárd euró veszteség hozható összefüggésbe a depresszióval, amelyből 54 milliárd termelési kiesés következtében született [12]. Ennél is súlyosabb következménye azonban, hogy a depresszió súlyosbodásával megnő az öngyilkosság kockázata is. Évente csaknem 800 000 ember veszti életét öngyilkosság miatt [13].

1.1.1 Depressziós súlyossági osztályok, skálák

A depresszió súlyosságának mérésére több széleskörűen alkalmazott skála is létezik. Létezik olyan, amely értékének meghatározása szaktudással rendelkező pszichiáter feladata, ám elterjedtek az önkitöltős kérdőívek pontértékein alapuló skálák is. A továbbiakban bemutatok néhány elterjedt és a világban depresszió klinikai diagnózisa során használt skálát.

1.1.1.1 Beck Depression Inventory II

A Beck Depression Inventory II (BDI-II) skála a korábbi, Beck Depression Inventory skála egy módosított változata, amely mára egyértelműen kiszorította az eredeti skála használatát [14]. A skála alapját egy 21 kérdésből álló önkitöltős kérdőív nyújtja. Minden kérdésre a [0; 3] tartományból kapható pontszám, amelyek összege lesz a válaszoló végső pontszáma. Ez alapján a skála teljes tartománya a [0; 63] intervallum. A skála értéktartományaihoz hozzárendelt súlyossági osztályok az 1.1. táblázatban láthatóak.

1.1.1.2 Hamilton Rating Scale for Depression

A Hamilton Rating Scale for Depression (HAM-D) skála értéke a BDI-II-vel szemben csak szakértő által határozható meg [15]. Az érték megállapítása során 17 tünetet vesznek vizsgálat alá, amely tünetekre a pszichiáter a tünet fontosságának függvényében a [0; 2], [0; 3] vagy [0; 4] tartományban ad értéket. A kapott pontértékek összege lesz a vizsgált beteg végső HAM-D skála szerinti értéke. Ezen skála egyes tartományainak és a depressziós súlyossági osztályok összerendelése az 1.1. táblázatban látható.

1.1.1.3 Patient Health Questionnarie for Depression

A Patient Health Questionnarie for Depression (PHQ-8) skála alapja egy 8 kérdésből álló önkitöltős kérdőív, amely a depresszió különböző tüneteiről szóló kérdéseket tartalmaz [16]. Létezik egy módosított változata is ennek a kérdőívnek, amely plusz egy, a kitöltő önbántásra való hajlandóságáról szóló kérdést tartalmaz és a PHQ-9 skála alapját jelenti.

Az egyes kérdésekre annak függvényében kap pontokat a kitöltő a [0; 3] intervallumból, hogy a kérdésben megfogalmazott tüneteket milyen intenzitással érezte magán a kitöltés előtti 2 hétben. Ennek megfelelően a skála teljes tartománya a [0; 24] intervallum. A skála egyes tartományainak depresszió súlyossági osztályokra való leképzése az 1.1 táblázatban található.

Kategória	BDI-II	HAM-D	PHQ-8
nem depressziós	[0; 13]	[0; 7]	[0; 4]
enyhe depresszió	[14; 19]	[8; 13]	[5; 9]
közepes depresszió	[20; 28]	[14; 18]	[10; 14]
súlyos depresszió	[29; 63]	[19; 23]	[15; 19]
nagyon súlyos depresszió	-	[23; 48]	[20; 24]

1.1. táblázat Depressziós súlyossági osztályok határértékei különböző skálák szerint

1.2 Beszédproduktumot leíró jellemzők

A beszéd hanghullámok révén (akusztikai úton) közvetíti az információt. A beszéd összefoglaló neve mindannak, amit egy nyelvi közösség tagjai, vagyis az ugyanazon nyelven beszélő emberek szóbeli érintkezésük során hangos közlésként mondanak. A beszédkeltés alapvető fiziológiai szervei a tüdő, a légcső, a gége, a garat, a száj- és orrüreg, amelyek működését az agy irányítja [17]. Ezek alapján a beszédjel feldolgozása által vagy a beszédkeltésért felelős szervekben való elváltozásokat (például: megfázás, tumor), vagy az agy működésében fellelhető elváltozásokat (például: depresszió, Parkinson-kór) van lehetőség detektálni [18]. Ilyen detekció végrehajtásához definiálhatóak úgynevezett alacsony szintű és magas szintű jellemzők, amelyek különböző aspektusokból képesek jellemezni a vizsgált beszédproduktumot.

A beszéd egy időben folyamatosan változó jel. Ilyen időben folytonos jelek - mint a beszéd - vizsgálata hatékonyan elvégezhető egy meghatározott időablak alkalmazásával. Megfelelően kicsi időtartományú ablak választásával az időtartományon belül közel állandónak tekinthető a beszédjel, amely így lehetővé teszi a jel különböző paraméterei mentén való mérését, jellemzését. Több tanulmány is kimutatta, hogy ilyen időablakos módszerrel több olyan jellemző is meghatározható, amelyek hasznosak lehetnek egy beszédalapú depresszió felismerő eljárás elkészítésekor. Ilyen jellemzők lehetnek az energia, formánsfrekvenciák, MFCC (Mel Frequency Cepstrum Coefficients) értékek, alapfrekvencia, jitter és shimmer értékek. Ezeket az értékeket szokás alacsony szintű jellemzőknek is nevezni [5, 19].

Ugyanakkor magas szintű (statisztikai) jellemzők alatt olyan, a beszédproduktumot számszerűsítő leírókat értünk, amelyeket nem hangokon, hanem nagyobb nyelvi elemeken értelmezünk, például mondatokon vagy hosszabb beszédszakaszokon. Ilyen jellemzők az artikulációs sebesség, a beszédtempó, a tranziens arány vagy a relatív szünethossz [20].

1.2.1 Mel skála

A hangmagasság a hang azon tulajdonsága, amely alapján magasnak vagy mélynek ítélünk egy hangot. A hangmagasság egy szubjektív érzet, függ a befogadótól. Egy hang hangmagasságát a kiváltó hang számos különböző fizikai jellemzője befolyásolja, mint például a hang frekvenciája, hangossága, sprektuma, tartóssága. Leírására több, az érzékelést alapul vevő hangmagasság skálát is létre hoztak [21].

A mel skála egy pszichofizikai hangmagasságskála, mértékegysége a mel. Alapja az a kísérlet, amely szerint, ha egy átlagos hallgató meghallgat egy 4000 Hz-es hangot, amelyet aztán egy jóval alacsonyabb követ, majd a hallgató feladata, hogy behangoljon egy oszcillátort a két hallott hang felére, az alanyok legnagyobb valószínűséggel 1000 Hz környékére fogják állítani az eszközt. Emiatt a mel hangmagasságskálán az 1000 Hz-et a skálát definiálók a 0 Hz és a 4000 Hz között éppen félúton helyezték el. Ezzel együtt az egész skálára igaz, hogy az értékeit úgy állították be rajta, hogy amennyiben egy hang frekvenciája megduplázódik, duplázódjon meg az érzékelt hangmagasság is [22].

Adott hang frekvenciájából a hang mel skála szerinti hangmagasság értéke kiszámolható az alábbiak szerint:

$$m = 2595 * \log_{10}(1 + \frac{f}{700}) \tag{1.1}$$

ahol f a kiváltó hang frekvenciája, m a mel skála szerinti hangmagassága.

Ekkor a 0–16 kHz közötti frekvenciatartományt a 0–2400 mel értéksor jellemzi, így 100 mel felett 24 mel-sávot lehet mérni.

1.2.2 Mel Frekvenciás Kepsztrum Együtthatók

A jelfeldolgozásban gyakori a jelek idő- és frekvenciatartományban való elemzése. Létezik azonban egy olyan fogalom, a kepsztrum, amely ugyancsak használatos bizonyos jelek feldolgozásában; eredetileg szeizmikus jeleknél alkalmazták annak érdekében, hogy szétválasszák a jelet komponenseire (elválasszák a jelet a visszhangjától). Vizsgálatok során azonban kiderült, hogy ezen módszer nem csak a szeizmikus jelek, de a beszédfeldolgozás területén is hasznosítható [24].

Az MFC rövidítés a Mel Frekvenciás Kepsztrum (Mel Frequency Cepstrum) kifejezésnek felel meg, azaz az MFC-együtthatók mel frekvenciás kepsztrális együtthatók (Mel Frequency Cepstrum Coefficients, MFCC). Az MFC-együtthatók kiszámítása egy úgynevezett lényegkiemelési eljárás, amelyet igen széles körben alkalmaznak a beszédtechnológiában. Lényegkiemelés alatt azt értjük, hogy a beszédjelből az információtartalom szempontjából releváns paramétereket kiemeljük, a többi, lényeges információt nem hordozó vagy redundáns jellemzőt pedig eldobjuk, lényegében tehát beszédtömörítés történik. Az eljárás alapja, hogy feltételezzük, ha az emberi hallást kellő pontossággal közelítő algoritmust használunk a beszéd tömörítésére, akkor lényegi információt nem veszítünk [21].

Egy beszédjelből a beszédjelet leíró mel-sávos energia értékeket úgy lehet megkapni, hogy a beszédjel egy rövid szeletét Fourier-transzformáljuk, majd rajta szűrősoros elemzést végzünk el, azaz az összetevőket mel-sávok szerint összegezzük. Az így kapott összegek alkotják a beszédjel mel sprektumát. Ezen összegek logaritmusának diszkrét koszinusz transzformációját véve kapjuk meg az MFC-együtthatókat [25]:

$$X_k = \sum_{n=0}^{N-1} x_n * \cos\left(\frac{\pi}{N} * \left(n + \frac{1}{2}\right) * k\right) \tag{1.2}$$

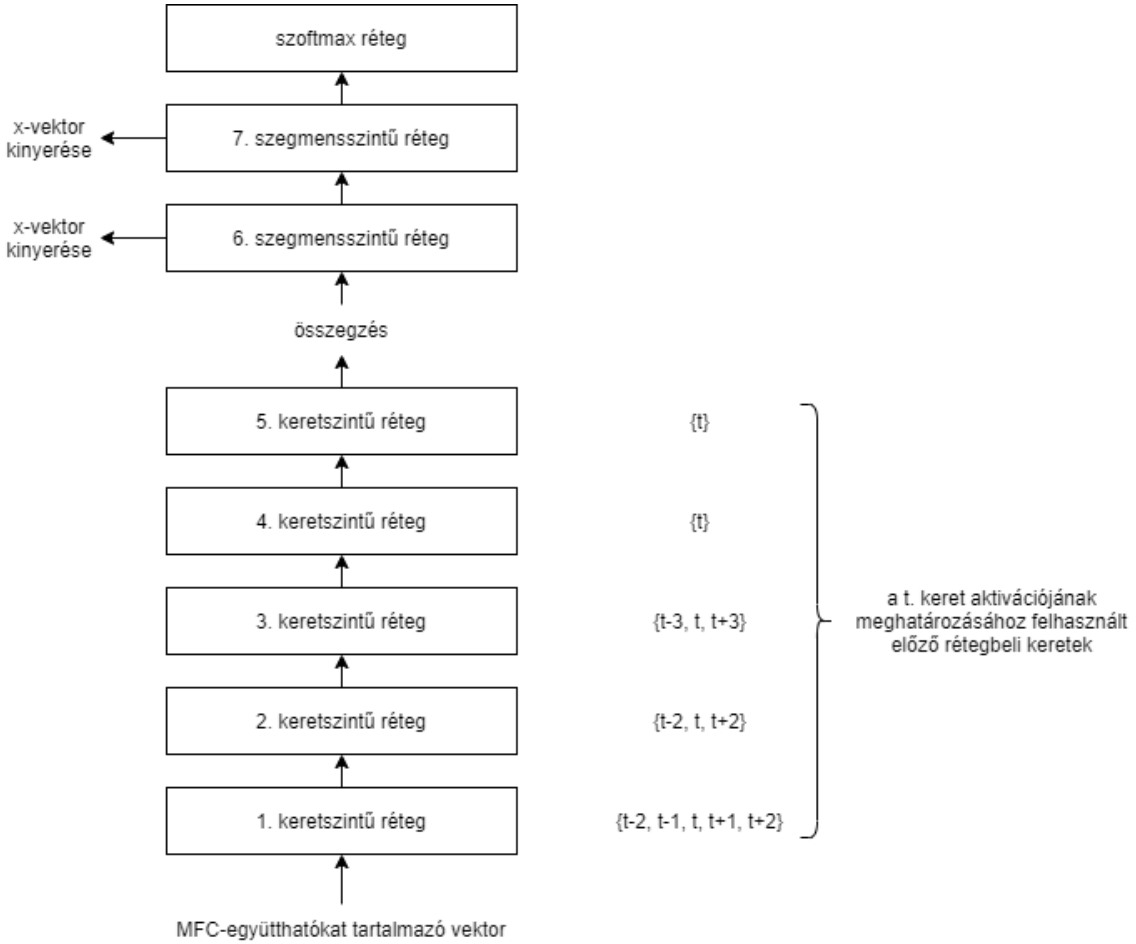
ahol X_k a k-adik transzformált MFC-együttható $(k = 0, 1, \dots)$, x_n az n-edik eredeti együttható, N a szűrők száma.

2 Módszerek

2.1 X-vektor technika

Az x-vektor technika egy neurálisháló-alapú jellemzőkinyerő eljárás, mely a változó hosszú hangfelvételeket x dimenziószámú jellemzőtérbe képzi. Technikailag egy mély neurális hálóról (Deep Neural, Network, DNN) van szó, melynek bemenetei keretszintű vektorok, a mélyebben elhelyezkedő rejtett rétegei keretszintűek, magasabb rejtett rétegei pedig szegmensszintűek [26].

A bemeneti keretszintű vektorok sok esetben az egyes beszélőkre számított MFC-együtthatókat tartalmazó vektorok, melyeket felhasználva a DNN képes végrehajtani a jellemzőkinyerést. A hálóból kinyerhető x-vektorok (más néven beágyazások) a szegmensszintű rétegek aktivációit jelentik.



2.1. ábra A Snyder és munkatársai által bevezetett architektúra szemléltetése

Az egyik legelterjedtebb x-vektor architektúrát Snyder és munkatársai vezették be, melyet beszélő-felismerésre fejlesztettek. Az architektúra felépítése a következő: az első 5 réteg a hálóban keretszintű, majd egy összegzést követően 2 szegmensszintű réteg, végül egy szoftmax réteg [26].

A keretszintű rétegek egy úgynevezett időeltolásos módszert valósítanak meg. Ennek lényege, hogy a t. keret aktivációjának kiszámításához egy adott rétegben nem csak a t. kerethez tartozó kimenetét használjuk fel az előző rétegnek, hanem annak valamekkora környezetét is. Például, az első keretszintű rétegben a $\{t-2, t-1, t, t+1, t+2\}$ kereteket használja fel a háló. Hasonlóan a második rétegben a t. keret aktivációja az előző réteg $\{t-2, t, t+2\}$ kereteitől, míg a harmadik rétegben a $\{t-3, t, t+3\}$ keretek aktivációitól függ. Látható, hogy ezáltal a t. keret aktivációja nem csak önmagától, hanem a rétegek előrehaladtával az egyre bővülő környezetétől is függ. A negyedik és az ötödik rétegek bár még mindig keretszintűek, de már nem bővítik tovább a t. keret aktivációját meghatározó környezetének méretét.

Az összegzés során a háló az ötödik keretszintű réteg kimenetének átlagát és szórását számolja ki, ez lesz az első szegmensszintű réteg bemenete. A szegmensszintű rétegek és a szoftmax réteg már egy teljesen összekötött hálót alkotnak. A szoftmax réteg a háló tanításához használt halmazban lévő beszélők számával azonos darab neuront tartalmaz. A teljes háló felépítését a 2.1. ábra szemlélteti.

A háló tanításához a fent bemutatott struktúrának köszönhetően elegendő szegmensszinten címkézett adathalmaz. Egy már betanított háló esetében az x-vektorok mindkét (hatodik és hetedik) szegmensszintű rétegből kinyerhetők.

2.1.1 X-vektor használata depresszió felismerésben

Az előzőekben bemutatott architektúrát eredetileg beszélő felismerésre dolgozták ki, ám hipotézisünk szerint az architektúra alkalmas lehet a betegség felismerés problémájának megoldására is. Alkalmazták már például az x-vektor technikát Parkinson-kór automatikus felismerésére [27].

Az x-vektorok használata a depresszió-felismerésben egy jelenleg is kutatásra érdemes, nem teljes mértékben felderített terület. A korábbiakban a Magyar Depressziós Beszédadatbázis hanganyagaiból generált x-vektorok felhasználása által sikerült jó

eredményeket elérni a depresszió-felismerés területén [28]. Az x-vektorok előállításához két DNN-t is alkalmaztak, az egyik egy angol nyelvű adatbázison előre tanított háló volt, míg a második a Magyar Spontán Beszédadatbázis (BEA korpusz) adatain tanított. A második esetben az adatok diverzitásának és zajtűrő képességének növelése érdekében az eredeti adathalmaz mellett kipróbálták a BEA korpusz augmentált verzióját is. Keret szintű reprezentációként ugyancsak két megközelítést teszteltek: egy 23 MFC-együtthatóból álló reprezentációt, valamint egy 40 szűrőbankból álló reprezentációt. A becslés elvégzéséhez SVR modellt alkalmaztak. A legjobb eredményeket az augmentált BEA korpuszon tanított háló x-vektorinak felhasználásával sikerült elérniük, a szűrőbankokból álló reprezentáció használata mellett. A modell által elért RMSE: 9,54, Pearson korreláció: 0,68.

2.2 Gépi tanuló eljárások

2.2.1 Szupport Vektor Gép

A Szupport Vektor Gép (Support Vector Machine, SVM) egy felügyelt tanulást megvalósító gépi tanuló eljárás, melyet széleskörűen alkalmaznak osztályozási problémák megoldására [29]. Az eljárás célja, hogy egy jellemzővektorokból álló minta elemeit az eljárás képes legyen osztályokhoz rendelni. Alapvetően az eljárás bináris osztályozásra alkalmas, ám kiterjeszthető n osztályú osztályozásra is. A kétosztályú osztályozás esetén hagyományosan a két osztályt megfeleltetjük a -1 és +1 értékekkel.

2.2.1.1 Lineárisan szeparálható eset

Az osztályozás alapgondolata, hogy a modell a jellemzővektorok által kifeszített térben keres egy olyan hipersíkot, amely hiba nélkül képes kettéválasztani a jellemzővektorokat osztálycímkéiknek megfelelően. A két osztályba eső vektorok közül azokat, amelyek a legközelebb esnek a hipersíkhoz, szupport vektoroknak nevezzük. Abban az esetben, ha több olyan hipersík is létezne, amely kettéválasztja a mintáinkat, a szupport vektoroktól számított lehető legnagyobb margóval rendelkező hipersíkot választja a modell, ezzel törekedve a lehető legnagyobb általánosítóképesség elérésére.

Legyen a tanítóhalmazban mérete n, a benne lévő x jellemzővektorok dimenziója D, y a hozzájuk tartozó osztálycímke! Ekkor tehát a tanítóhalmaz felfogható az x jellemzővektorok és az y osztálycímkék rendezett ketteseinek halmazaként:

$$\{\{x_i;\ y_i\} | i \in \{1, 2, \dots, n\}, x_i \in \mathbb{R}^D, y_i \in \{-1, 1\}\} \quad (2.1)$$

Ekkor a keresett hipersík a $w \cdot x + b = 0$ egyenlettel írható le, ahol w egy a hipersíkra merőleges vektor, $\frac{b}{\|w\|}$ pedig a hipersík origótól vett távolsága.

Legyen J a szupport vektorok indexeinek halmaza! Ekkor a maximális margójú hipersík megtalálása lényegében a w és a b értékének meghatározása az alábbi feltételek mellett:

$$w \cdot x_j + b \geq 1, ha\ y_j = 1, j \in J \quad (2.2)$$

$$w \cdot x_j + b \leq -1, ha\ y_j = -1, j \in J \quad (2.3)$$

Ekkor a hipersíkhoz tartozó margó szélessége $\frac{1}{\|w\|}$, ezt szükséges maximalizálni. Ezen kifejezés értéke ott maximális, ahol $\|w\|$ minimális. Ezen minimalizálással pedig egyenértékű az $\frac{1}{2}\|w\|^2$ kifejezés minimalizálása.

Az előző két feltételt egyesítve és a fenti logikát alkalmazva így az alábbi feltételrendszer és feladat adódik:

$$y_j * (w \cdot x_j + b) - 1 \geq 0, \forall j \in J \quad (2.4)$$

$$min_{w,b}\left(\frac{1}{2}\|w\|^2\right) \quad (2.5)$$

Ez a feladat egy konvex optimalizálási feladat, amely megoldható Lagrange-multiplikátorok használatával.

2.2.1.2 Lineárisan nem szeparálható eset

Gyakori eset, hogy a megoldani kívánt osztályozási probléma a fentiekkel ellentétben nem lineárisan elválasztható, azaz nem létezik olyan hipersík a jellemzővektorok által kifeszített térben, amely hiba nélkül elválasztaná a tanító halmaz elemeit az osztálycímkéiknek megfelelően. Az SVM képes kezelni az ilyen problémákat is, mely esetekre bevezeti a kernel függvény fogalmát.

A kernel függvény segítségével lehetőség van arra, hogy a jellemzővektorokat áttranszformáljuk az eredeti terükből egy olyan, magas dimenziójú térbe, ahol már lineárisan szeparálhatóvá válnak, azaz található elválasztó hipersík hozzájuk. A legelterjedtebb kernel függvények az alábbiak:

- lineáris: $K\left(\theta(x_i),\theta(x_j)\right) = x_i^T x_j$ (2.6)
- polinomiális: $K\left(\theta(x_i),\theta(x_j)\right) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (2.7)
- radiális bázisfüggvény (RBF): $K\left(\theta(x_i),\theta(x_j)\right) = e^{-\gamma\|x_i - x_j\|^2}, \gamma > 0$ (2.8)
- szigmoid: $K\left(\theta(x_i),\theta(x_j)\right) = \tanh(\gamma x_i^T x_j + r), \gamma > 0$ (2.9)

ahol d, γ és r a kernel függvények hiperparaméterei.

Fontos megjegyezni azonban, hogy elronthatja a modell általánosító képességét az, ha túl magas dimenzióba transzformál a kernel függvény, így bevett módszer egy ξ gyengítő paraméter bevezetése, amely értéke a következő tartományokba eshet:

- ha az adott vektor a hipersík megfelelő oldalán helyezkedik el a margón kívül: $\xi = 0$
- ha az adott vektor a hipersík nem megfelelő oldalán helyezkedik el: $\xi = 1$
- ha az adott vektor a hipersík megfelelő oldalán helyezkedik el, de a margón belül: $0 < \xi < 1$

Ennek megfelelően mind a feltételek, mind a minimalizálni kívánt kifejezés módosul az alábbiak szerint:

$$y * \left(w \cdot x_j + b\right) - 1 + \xi_j \geq 0, \forall j \in J \quad (2.10)$$

$$min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right) \quad (2.11)$$

ahol n a tanítóhalmaz mérete, $C \geq 0$ hiperparaméter szabályozza a ξ gyengítő paraméter okozta büntetés mértékét.

2.2.2 Szupport Vektor Regresszió

A Szupport Vektor Regresszió (Support Vector Regression, SVR) egy felügyelt gépi tanuló eljárás, amely működési elve hasonlít a Szupport Vektor Gép algoritmus alapgondolatához, de a tanítómintákhoz ebben az esetben nem osztálycímkék rendeltek, hanem egy y valós szám, az algoritmus feladat pedig ezen szám becslése [29]. Ehhez a SVR egy olyan $f(x)$ függvényt keres, amely kellően jól közelíti az y célértéket, miközben minimális a függvény változása. Ez az optimum megtalálása az algoritmus legnagyobb

feladata, kompromisszumot kötni az y értékek maximum ε hibájú közelítése és a függvény általánosító képessége között.

A SVR döntési függvénye teljesen hasonló a SVM által keresett hipersík egyenletéhez:

$$f(x) = w \cdot x + b \tag{2.12}$$

Ahhoz, hogy a függvény változása minél kisebb legyen, az $\frac{1}{2}\|w\|^2$ kifejezés értékét kell minimalizálni az alábbi feltételrendszer mellett:

$$|y_j - f(x_j)| \leq \varepsilon, \forall j \in J \tag{2.13}$$

Hasonlóan a SVM esetéhez, itt is előfordulhat, hogy ezen feltételrendszer nem kielégíthető. Ekkor itt is alkalmazható kernel függvény, amely segítségével magasabb dimenziójú térbe transzformálható a probléma, valamint bevezethető egy gyengítő ξ paraméter, amely nyomán az alábbiak szerint módosul a feltételrendszer és a feladat:

$$|y_j - f(x_j)| \leq \varepsilon + \xi_j, \forall j \in J \tag{2.14}$$

$$min_{w,b}(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i) \tag{2.15}$$

ahol $C \geq 0$ hiperparaméter szabályozza a ξ gyengítő paraméter okozta büntetés mértékét.

2.2.3 Skálázás

Gépi tanuló eljárások alkalmazása során gyakori, hogy különböző adatelőkészítő lépések végrehajtása ajánlatos a különböző modellek alkalmazása előtt. Ilyen lépés lehet többek között a kiugró értékek elhagyása, hiányzó értékek pótlása vagy a numerikus adatok skálázása is [30].

A skálázás célja, hogy az adathalmaz numerikus oszlopainak értékeit úgy módosítsa, hogy azok közös skálát használjanak anélkül, hogy torzítaná az értéktartományok különbségeit. Ez az átalakítás számos formában megvalósítható, ezek közül szeretnék bemutatni néhányat. Az alábbiakban egy X minta x_i elemének különböző skálázási lehetőségeit mutatom be, ahol $\widehat{x_i}$ a skálázott érték:

- Z-skálázás: $\widehat{x_i} = \frac{x_i - \bar{x}}{\sigma}$ (2.16)
- 0-1 skálázás: $\widehat{x_i} = \frac{x_i - \min(X)}{\max(X) - \min(X)}$ (2.17)
- Maximum abszolút skálázás: $\widehat{x_i} = \frac{x_i}{\max(X')}$ (2.18)

ahol $\bar{x}$ a minta számtani közepe, σ a szórása, $\min(X)$ a minimuma és $\max(X)$ a maximuma, X' pedig páronként megegyezik az X minta abszolútértékeivel, azaz $\max(X')$ a minta legnagyobb abszolút értékű eleme.

2.2.4 Kiértékelési metrikák

Egy gépi tanuló eljárás alkalmazása során az első lépések egyike az, hogy a rendelkezésre álló mintát fel kell osztani két disztjunkt halmazra, melyek közül az egyik tanítóhalmaznak, a másikat teszthalmaznak nevezzük. Erre azért van szükség, mert a kiértékelés során egy már betanított modellt egy olyan mintán kell kiértékelni, amely minta elemeit nem használtuk fel a modell tanítása és optimalizálása során. Így az eljárás a következő: a választott gépi tanuló eljárást betanítjuk a tanítóhalmaz elemein, majd kiértékeljük azt a teszthalmaz elemein.

Legyen a teszthalmaz mérete n, x_i az i. mintapont a teszthalmazban, y_i az i. mintaponthoz tartozó becsülni kívánt érték, $\widehat{y_i}$ az i. mintaponthoz tartozó becsült érték ($1 \leq i \leq n,\ i \in \mathbb{N}$)! Ekkor egy regressziós feladat során leggyakrabban alkalmazott kiértékelési metrikák a következők szerint definiálhatók [32]:

- Mean Absolute Error (MAE):

$$\frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n} \tag{2.19}$$

- Root Mean Squeared Error (RMSE):

$$\frac{\sqrt{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}}{n} \tag{2.20}$$

- Pearson korreláció (CORREL):

$$\frac{\sum_{i=1}^{n}(y_i-\bar{y})*(\hat{y_i}-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}*\sqrt{\sum_{i=1}^{n}(\hat{y_i}-\bar{\hat{y}})^2}} \tag{2.21}$$

Annak érdekében, hogy olyan modellek is összehasonlíthatóak legyenek, amelyek különböző skálát használó címkék becslésének feladatára hoztak létre, szokás használni a RMSE egy módosított változatát is. A Normalized Root Mean Squerd Error (NRMSE). metrikának többféle, különböző definiciója is használatos, dolgozatomban én az alábbit használom:

$$NRMSE = \frac{RMSE}{\sigma} \tag{2.22}$$

ahol σ a modell tanításához használt minta címkéinek szórása, RMSE pedig a korábbiakban már definiált kiértékelési metrika értéke a vizsgált modellen. A NRMSE ilyen formában való definiálásával össze lehet hasonlítani különböző tanítóhalmazon tanított, különböző skálát használó címkék becslésének minőségét.

Osztályozási feladatok során – hasonlóan a regressziós feladatokhoz – több metrika is használható, ezek közül szeretnék bemutatni néhányat az alábbiakban. Ezen metrikák bemutatásához azonban szükséges néhány további fogalom leírása:

- Álnegatív (False Negative, FN): Ha a mintapont osztálya eredetileg pozitív (esetünkben beteg vagyis depressziós), de a predikció negatív (nem beteg).
- Valós negatív (True Negative, TN): ha a mintapont osztálya eredetileg negatív és a predikció is negatív.
- Álpozitív (False Positive, FP): ha a mintapont osztálya eredetileg negatív, de a predikció pozitív.
- Valós pozitív (True Positive, TP), ha a mintapont osztálya eredetileg pozitív és a predikció is pozitív.

Osztályzási algoritmusok kiértékeléséhez használható metrikák [33]:

- Pontosság (Accuracy): A helyesen osztályozott mintapontok hányadosa a teljes mintahalmaz méretéhez képest.

$$\frac{TP+TN}{FN+TN+FP+TP} \tag{2.23}$$

- Szenzitvitás (Sensitivity): A helyesen pozitívnak osztályozott mintapontok hányadosa az összes valós pozitív mintapontok számához képest.

$$\frac{TP}{FN+TP} \tag{2.24}$$

- Specificitás (Specificity): A helyesen negatívnak osztályozott mintapontok hányadosa az összes valós negatív mintapontok számához képest.

$$\frac{TN}{TN+FP} \tag{2.25}$$

- Tévesztési mátrix (Confusion Matrix): A fent definiált 4 értéket (TN, FP, FN, TP) szokás egy kétszer kettes mátrixban bemutatni, amit tévesztési mátrixnak nevezünk (2.1. táblázat).

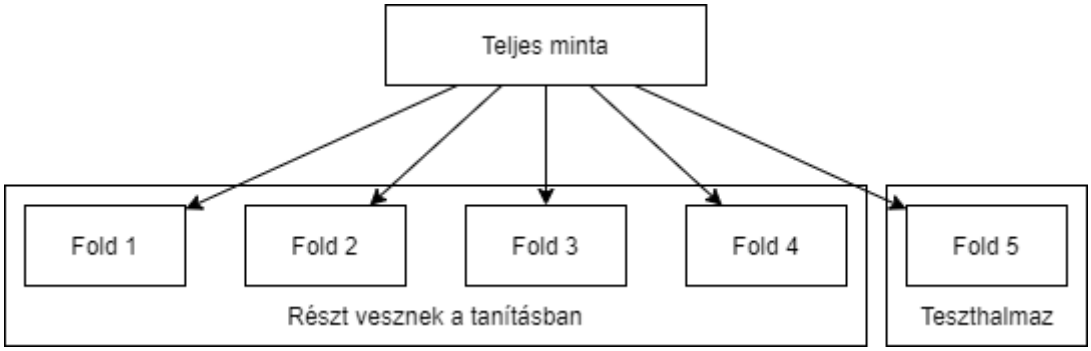
	Valóság		
Predikció		pozitív	negatív
	pozitív	TP	FP
	negatív	FN	TN

2.1. táblázat A tévesztési mátrix szemléltetése

2.2.5 Keresztvalidáció

Gépi tanuló eljárások alkalmazása során – amennyiben a rendelkezésre álló adatok mennyisége túl kevésnek bizonyul – gyakori a keresztvalidáció technika (Cross Validation, CV) használata [31]. A keresztvalidáció azt jelenti, hogy a folyamat legelején nem kerül felosztásra a teljes rendelkezésre álló minta egy tanuló- és egy teszthalmazra, hanem (k-fold CV esetén) a minta k darab disztjunkt halmazra kerül bontásra. A modell

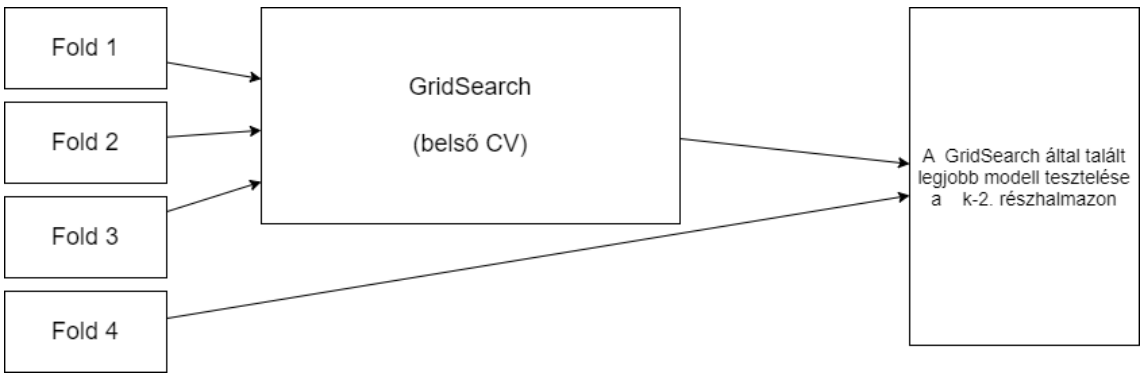
alkalmazása során k iterációt hajtunk végre, minden egyes iterációban $k-1$ darab halmaz unióját, mint tanítóhalmazt használva a tanításhoz az egy kimaradt halmazt használva a teszteléshez. Egy iterációjának felosztását mutatja meg a 2.2 ábra.



2.2. ábra Egy 5-fold-os klasszikus keresztvalidáció egy iterációja

A beágyazott keresztvalidáció (Nested CV) egy olyan kiértékelési módszer, amely során egy keresztvalidáción belül egy másikat is alkalmazunk. A belső ciklus úgy működik, mint egy klasszikus CV, a külső ciklus pedig a teszthalmazán teszteli a belső ciklusban optimalizált legjobb hiperparaméterekkel rendelkező modellt.

A két CV dolgozhat azonos és különböző számú fold-okkal is. Egy k-fold külső keresztvalidációt és l-fold belső keresztvalidációt megvalósító beágyazott keresztvalidáció során a külső keresztvalidáció 1 iterációjában $k-1$ részhalmaz uniója kerül át a belső keresztvalidációba, amely ezt a részhalmazt bontja $l-1$ részre és hajt rajuk végre egy klasszikus CV-t. Így összesen $k*l$ tanítást hajt végre ez a hiperparaméter-optimalizáló algoritmus. Az algoritmus egy (külső) iterációját szemlélteti a 2.3. ábra. A beágyazott keresztvalidáció lehetőséget nyújt arra, hogy a belső CV eredményeit és a külső ciklus eredményeit is mérve képet alkothassunk a modell általánosító képességéről.



2.3. ábra Egy 4-fold-os beágyazott keresztvalidáció egy (külső) iterációja

3 Adatbázisok bemutatása

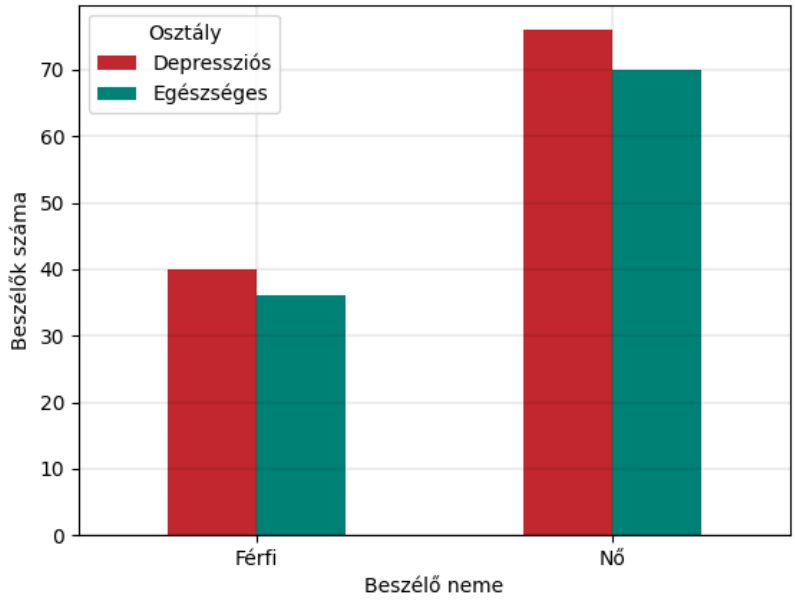
Vizsgálataim során négy, különböző nyelvű depressziós beszédadatbázist használtam fel, az alábbiakban ezeket szeretném egyesével bemutatni. Mindegyik beszédadatbázisra igaz, hogy a bennük található hangmintákat a későbbi feldolgozás érdekében egységesen egycsatornás, 8000 Hz-en mintavételezett wav formátumú fájlokká konvertáltam. A különböző nyelvű beszédadatbázisok áttekintésére nyújt lehetőséget a 3.1. táblázat.

Adatbázis	Nyelv	Méret	Depressziós súlyossági skála	Átlag	Szórás
DBA	magyar	222	BDI-II	16,02	13,08
AVID-Corpus	német	127	BDI-II	15,04	12,04
E-DAIC	angol	274	PHQ-8	6,95	6,11
MODMA	kínai	52	PHQ-9	9,4	8,41

3.1. táblázat A különböző nyelvű adatbázisok áttekintése

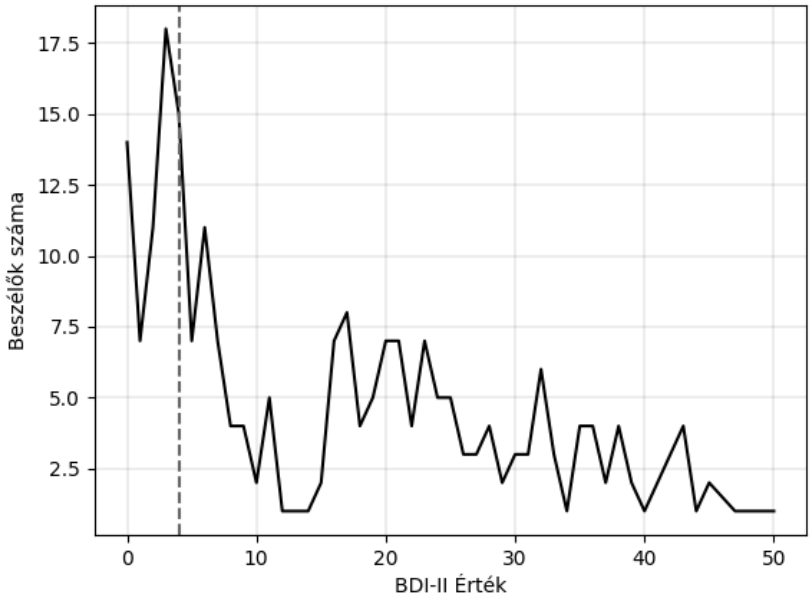
3.1 Magyar nyelvű depressziós beszédadatbázis

A Magyar Depressziós Beszédadatbázis (DBA) egy magyar nyelvű, egészséges és depressziós beszélőktől származó hanganyagokat tartalmazó adatbázis, amelyet a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszékének munkatársai készítettek a Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinikájának munkatársai segítségével. Az adatbázisban a beszédproduktumok mellett a beszélőkről megtalálhatóak az alábbi adatok: életkor, nem, depresszió súlyosságát leíró BDI-II skála szerinti érték.



3.1. ábra A DBA adatbázis nemek szerinti eloszlása a depressziós és az egészséges minták között

A beszélőktől származó hangminták előre rögzített szöveg („Az északi szél és a Nap” rövid mese) felolvasása által keletkeztek. Az adatbázisban 146 női és 76 férfi beszélőtől származó adatsor szerepel. A mintában szereplő nők 52,05%-a depressziós, a férfiak 52,63%-a. A minták nemek szerinti eloszlását és a BDI-II skála szerinti értékeik eloszlását a 3.1 és a 3.2 ábra mutatja be.

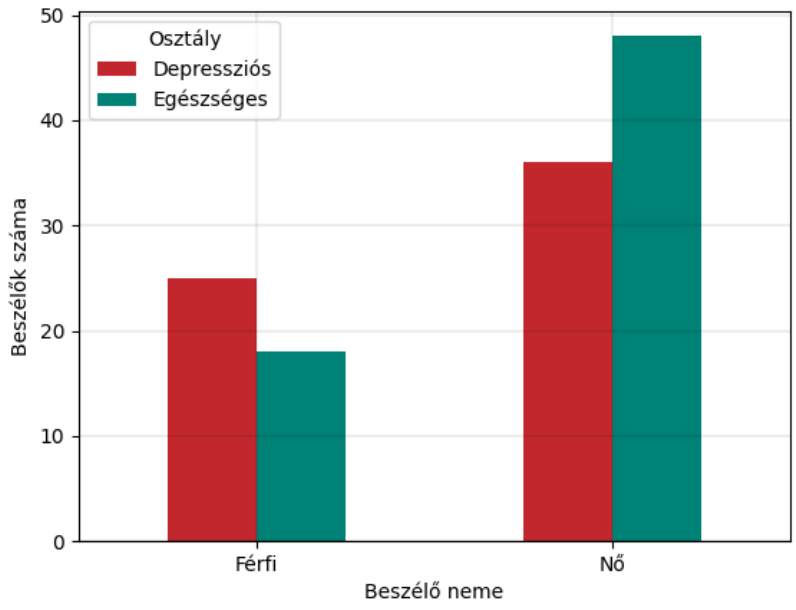


3.2. ábra A DBA adatbázis BDI-II skála szerinti eloszlása

A mintában található BDI-II skála szerinti értékek minimuma 0, maximuma 50, átlaga 16,02, szórása 13,08.

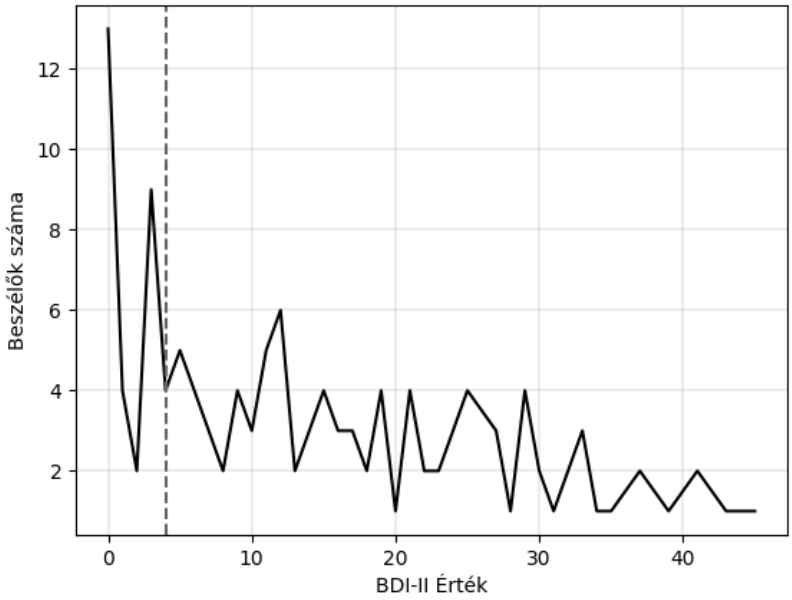
3.2 Német nyelvű depressziós beszédadatbázis

Az Anonymized Videos from Diverse Countries (AVID) egy német nyelvű adatbázis, amelynek én egy egészséges és depressziós beszélőktől származó hanganyagokat tartalmazó részhalmazát használtam fel kutatásaim során [8]. A hanganyagokon túl az adatbázis tartalmazza még a beszélők nemét és a BDI-II skálán elért értékét. Az adatbázishoz a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszéke által nyerhettem hozzáférést. A minta nemi eloszlása a 3.3. diagramon látható.



3.3. ábra Az AVID adatbázis nemek szerinti eloszlása a depressziós és az egészséges minták között

A beszélőktől származó hangminták előre rögzített szöveg („Az északi szél és a Nap” rövid mese) felolvasása által keletkeztek, hasonlóan a magyar adatbázishoz. Az adatbázisban 61 depressziós és 66 egészséges beszélőtől szerepelnek hanganyagok. A mintában található BDI-II skála szerinti értékek minimuma 0, maximuma 45, átlaga 15,04, szórása 12,04, eloszlása az 3.4. diagramon látható.

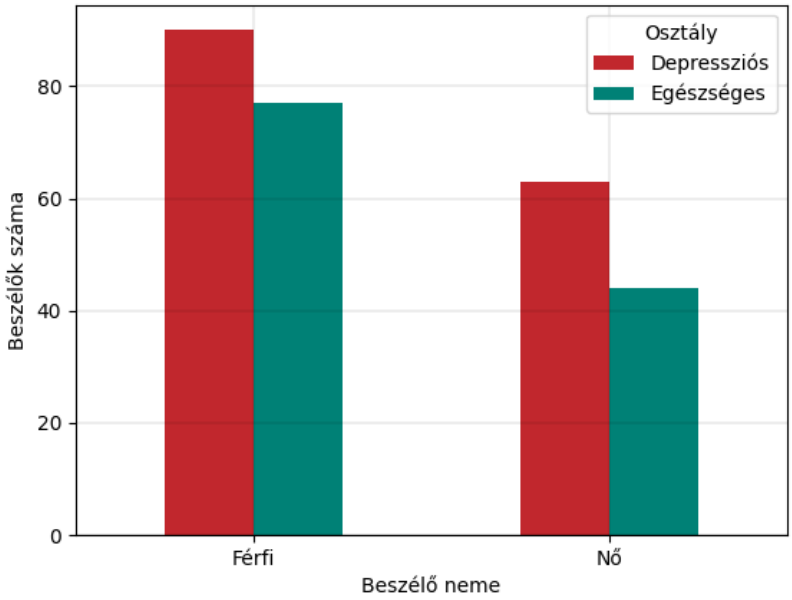


3.4. ábra Az AVID adatbázis BDI-II skála szerinti eloszlása

3.3 Angol nyelvű depressziós beszédadatbázis

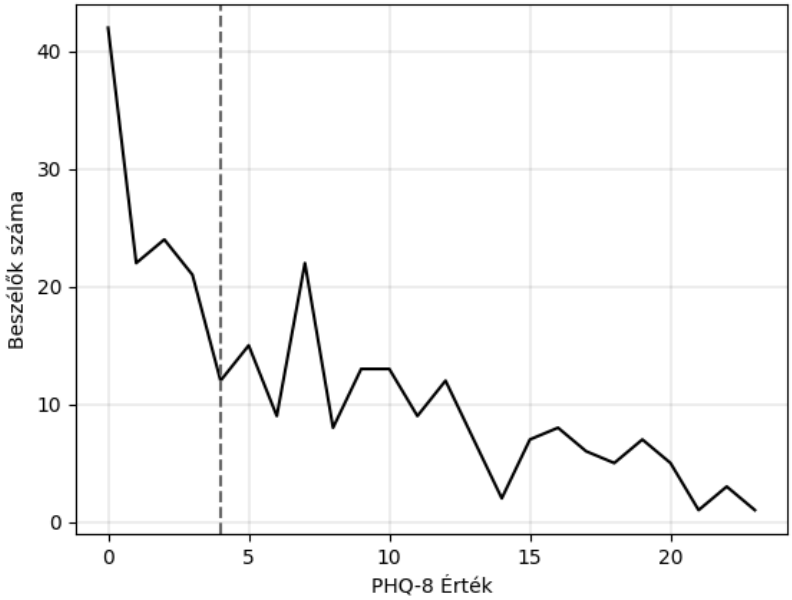
Az Extended Distress Analysis Interview Corpus (E-DAIC) egy angol nyelvű adatbázis, amely klinikai interjúk hanganyagait és videófelvételeit tartalmazza. Bár az adatbázis tartalmaz a depressziós és egészséges hangmintákon kívül egyéb betegséggel küzdő beszélőktől is hanganyagokat (poszttraumás stressz szindróma, túlzott szorongás), azonban kutatásaim során én csak a depressziós és egészséges beszélők hangmintáit használtam. Az interjúkat egy számítógépes ágens segítségével rögzítették a Dél-kaliforniai Egyetemen (University of Southern California, USC) [34].

Ezen adatbázis esetében szükséges volt még az x-vektorok generálása előtt egy adatelőkészítő lépést is elvégeznem: mivel az interjú közben sok szünet van a hanganyagokban, ezért az adatbázis mellé kapott leíró szövegek alapján kivágtam a szüneteket a hanganyagokból.



3.5. ábra Az E-DAIC adatbázis nemek szerinti eloszlása a depressziós és az egészséges minták között

Az adatbázisban 153 depressziós és 121 egészséges beszélőtől szerepelnek hanganyagok, nemi eloszlásukat a 3.5 ábra szemlélteti. A mintában található PHQ-8 skála szerinti értékek minimuma 0, maximuma 23, átlaga 6,95, szórása 6,11, eloszlása az 3.6. diagramon látható.

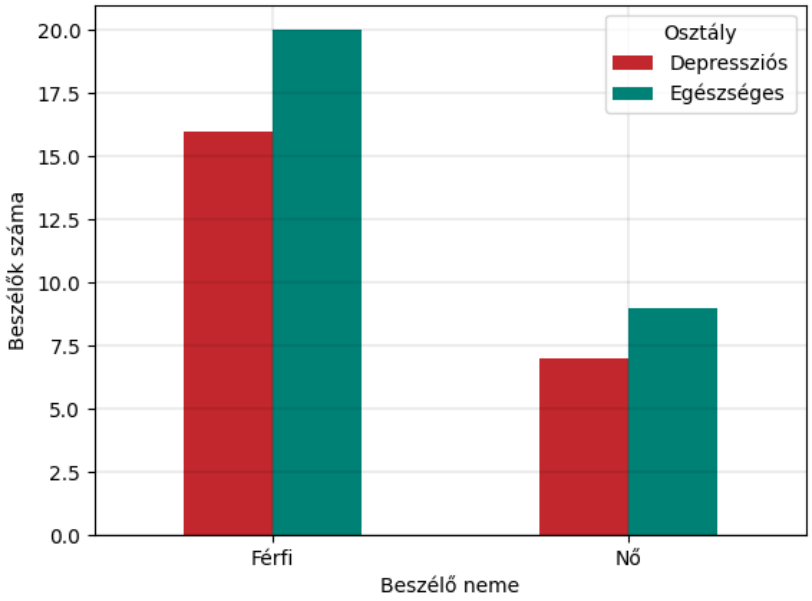


3.6. ábra Az E-DAIC adatbázis PHQ-8 skála szerinti eloszlása

3.4 Kínai nyelvű depressziós beszédadatbázis

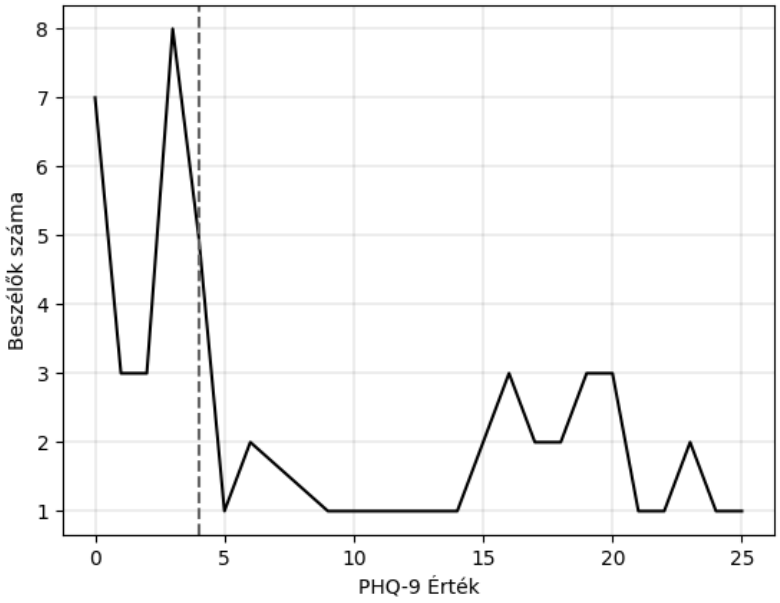
A Multi-modal Open Dataset for Mental disorder Analysis (MODMA) adatbázis depressziós és egészséges személyektől tartalmaz szabad szöveges hanganyagokat és EEG jeleket. A hanganyagok kínai nyelvűek és jelen kutatásom során én csak ezeket használtam fel dolgozatomhoz. A mintákat a Lanzhou Egyetemen gyűjtötték. A hanganyagok tartalmaznak interjút, felolvasást (folytonos szöveg és szójegyzék) és képleírást is. [35] A minták nemek szerinti eloszlását mutatja be a 3.7. diagram.

Mivel ezen adatbázis esetében egy-egy beszélőhöz több minta is tartozik, így fontos megjegyezni, hogy a hanganyagok tanítóhalmazba és teszthalmazba sorolásakor természetesen nem a hanganyagokat, hanem a beszélőket osztottam két disztjunkt halmazra, így minden esetben igaz volt, hogy nem szerepelt adott beszélőtől hangminta a tanítóhalmazban és a teszthalmazban is. Egy adott beszélőre vonatkozó végső predikció a teszthalmazban lévő, adott beszélőtől származó összes hanganyag predikcióinak számtani átlaga.



3.7. ábra A MODMA adatbázis nemek szerinti eloszlása a depressziós és az egészséges minták között

Az adatbázisban 23 depressziós és 29 egészséges beszélőtől szerepelnek hanganyagok. A mintában található PHQ-9 skála szerinti értékek minimuma 0, maximuma 25, átlaga 9,40, szórása 8,41, eloszlása az 3.8. diagramon látható.



3.8. ábra A MODMA adatbázis PHQ-9 skála szerinti eloszlása

4 Eredmények

4.1 Modellek bemutatása

Munkám során célom megvizsgálni a depresszió-felismerés SVR modell általi lehetőségeit. A modelljeim bemenete a különböző nyelvű hangadatbázisokból a Snyder és munkatársai által készített architektúrával előállított x-vektorok, kimenete az adott hangadatbázishoz tartozó depressziós súlyossági skála szerinti értékekhez (továbbiakban címke) előállított becslések.

A modelljeim tanításához és teszteléséhez felhasznált minták nyelve alapján megkülönböztetem az alábbi eseteket:

- Egynyelvű modellek: mind a tanításhoz mind a teszteléshez azonos nyelvű mintákat használok csak fel
- Többnyelvű modellek: Egy vagy több nyelv mintáit használom tanításhoz és egy másik, tanításhoz nem használt nyelv mintáit használom teszteléshez

A modellek tanításakor többféle skálázást is teszteltem a modellek bemeneti x-vektorjain és a becsülni kívánt értékeken egyaránt. Mind ν-SVR-t, mind ε-SVR-t alkalmaztam kutatásaim során. A tanítások alatt kipróbáltam mind lineáris, mind rbf kernel alkalmazását. Lineáris kernel esetében a C, míg rbf kernel esetében a C és γ hiperparaméter optimalizációját végeztem el. Mindkét paraméter esetében korábbi vizsgálataim alapján a legjobb értékeket a $[2^{-10}; 2^{10}]$ intervallumba eső 2 hatványok között kerestem. A modellek hiperparaméter-optimalizációit a modellek által elért RMSE kiértékelési metrika szerint végeztem el.

A tanítás után a teszthalmazon a kiértékelést is elvégzem, a RMSE mellett a MEA és a Pearson korreláció értékeket is vizsgálva. A regresszió becsült értékei alapján osztályzást is végeztem a mintáimon, amely által lehetőségem volt kiértékelni az osztályzás eredményességét is. Ezt a következő metrikák mellett végeztem el: pontosság, szenzitivitás, specificitás. Azt osztályozást minden nyelv esetén a nyelvhez rendelkezésre álló depressziós súlyossági skála szerint végeztem el, az alábbiak szerint:

- BDI-II skálát használó adatbázis esetében, ha a predikció értéke nagyobb, mint 13, akkor a predikált osztály Depressziós (DE), különben Egészséges (HE).

- PHQ-8 és PHQ-9 skálát használó adatbázisok esetében, ha a predikció értéke nagyobb, mint 4, akkor a predikált osztály Depressziós (DE), különben Egészséges (HE).

A létrehozott modelljeimet a tesztelésükhöz használt hangadatbázisoknak megfelelően neveztem el, az egynyelvűeket a „_MONO", a többnyelvűeket a „_MULTI" posztfixekkel ellátva.

4.2 Egynyelvű modellek

Az egynyelvű modellek esetében, mivel semelyik rendelkezésemre álló adatbázis mérete sem teljes mértékben kielégítő, így beágyazott keresztvalidációt alkalmaztam a modellek létrehozása során.

4.2.1 DBA_MONO modell

A magyar nyelvű adatokon a legjobb eredményeket standard skálázás mellett értem el, ν-SVR-t alkalmazva, rbf kernelt használva. Mind a beágyazott, mind a belső keresztvalidáció fold-számát 10-nek választottam. A hiperparaméter-optimalizáció során talált értékeket a 4.1. táblázat tartalmazza.

C	2^4	2^4	2^5	2^4	2^5	2^8	2^7	2^4	2^4	2^4
γ	2^{-9}	2^{-9}	2^{-10}	2^{-9}	2^{-10}	2^{-8}	2^{-8}	2^{-9}	2^{-9}	2^{-9}

4.1. táblázat A DBA_MONO modell optimalizációja során legjobbaknak talált hiperparaméter értékek

A kiértékelés során a regresszió és az azt követő osztályozás eredményeit a 4.2. táblázat tartalmazza.

RMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
9,70	7,67	0,67	86%	91%	79%

4.2. táblázat A DBA_MONO modell eredményei

Ezen az adatbázison korábbi kutatásokban 10,4 [36], illetve 10,1 [37] értékeket értek el, olyan beszédleíró jellemzők felhasználása mellett, mint a jitter, shimmer, intenzitás, formánsfrekvenciák, artikulációs sebesség, beszédtempó, tranziens arány.

Korábbi kutatásokkal összevetve megállapítható, hogy a modellem RMSE értéke kisebb (jobb) azokénál, pontosság és kifejezetten a szenzitivitás tekintetében pedig olyan

eredményeket ér el, amelyek alapján akár klinikai környezetben is alkalmazhatónak tűnik ez az eljárás. Bár a specificitás alacsonyabb a pontosság és a szenzitivitás értékénél, speciálisan a depresszió felismerés területén fontosabb a magas szenzitivitás értéke egy akár alacsonyabb specificitás érték mellet is, hiszen fontosabb a valódi pozitív esetek megtalálása, mint hogy minél kevesebb álpozitív jelzése legyen a modellnek.

4.2.2 AVID_MONO modell

A német adatbázis esetében a legjobb eredményeket ν-SVR-t alkalmazva, rbf kernelt használva értem el. Az előfeldolgozás során Z-skálázást végeztem az adatokon, mind az x-vektorokon, mind a célváltozón. A beágyazott keresztvalidáció során 20-nak választottam mind a belső, mind a külső keresztvalidáció fold-számát. A modell által elért eredményeket a 4.3. táblázat foglalja össze.

RMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
9,27	7,41	0,64	65%	77%	53%

4.3. táblázat Az AVID_MONO modell eredményei

Ugyanezen adatbázis egy részét alkalmazták a 2013-as Audio-Visual Emotion Recognition Challenge (AVEC) Depression Recognition Sub-Challenge (DSC) nevű versenyében is [8]. A versenyben való részvétel feltételeként a nevezetteknek legalább 14,12 RMSE értéket kellett elérniük, míg a győztes modell eredménye 8,68 RMSE érték lett. Egy másik, ezt az adatbázist feldolgozó kutatásban különböző alacsony és magas (statisztikai) szintű jellemzőket felhasználva ugyanezen adathalmazon 8,55 RMSE értéket sikerült elérniük a kutatás készítőinek [9]. Ezekhez viszonyítva az általam alkalmazott x-vektor módszer eredményei elérik ezen modellek performanciáját. Az modellem becslései alapján készített osztályozás rosszabb eredményének okai között lehetnek az adatbázis kisebb mérete és a hanganyagok rosszabb minősége is.

4.2.3 E-DAIC_MONO modell

Az angol E-DAIC adatbázist felhasználva, hasonlóan a magyarhoz, a legjobb eredményeket ν-SVR-t alkalmazva, rbf kernelt használva értem el. Az előfeldolgozás során maximum abszolút skálázást végeztem az adatokon, mind az x-vektorokon, mind a célváltozón. A beágyazott keresztvalidáció során 20-nak választottam mind a belső, mind

a külső keresztvalidáció fold-számát. A modell által elért eredményeket a 4.4. táblázat foglalja össze.

RMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
6,18	5,10	-0,09	56%	100%	0,01%

4.4. táblázat Az E-DAIC_MONO modell eredményei

Ez az adatbázis a 2019-es AVEC-DSC versenyében alkalmazták [38]. A belépési küszöbe a versenynek 6,37 RMSE érték volt, míg a győztes modell eredménye: 5,11 RMSE érték, ha csak a hanganyagokat felhasználó modellt veszem figyelembe [39]. Ezekkel az általam elért eredményeket összevetve láthatjuk, hogy bár a modellem nem ért el jó eredményeket sem a Pearson korreláció, sem az osztályozás bármely metrikáját tekintve, pusztán a RMSE értékét tekintve hasonló eredményeket ért el, mint más kutatócsoportok modelljei.

4.2.4 MODMA_MONO modell

A kínai MODMA adatbázis esetében ν-SVR-t alkalmaztam rbf kernellel. Az előfeldolgozás során nem végeztem skálázást az adatokon. A beágyazott keresztvalidáció során 20-nak választottam mind a belső, mind a külső keresztvalidáció fold-számát. A modell által elért eredményeket a 4.5. táblázat foglalja össze.

RMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
7,75	6,72	0.40	52%	100%	3%

4.5. táblázat A MODMA_MONO modell eredményei

A predikciókat megvizsgálva megértettem, hogy 0,4 értékű Pearson korreláció mellett miért ér el mégis ennyire rossz osztályzást a modell: bár korrelál a hanganyagok valós PHQ-9 értékeivel a predikcióm, de mindössze egy esetben becsül a modell 4 alatti értéket, így az egészséges mintákat félreosztályozza a modellem. Ennek megoldására bevezettem egy utófeldolgozási lépést, az alábbiak szerint:

Felosztottam a teszthalmazomat véletlenszerűen két disztjunkt részhalmazra, egy kisebb T_1 (a teljes teszthalmaz 15%-a) és egy nagyobb T_2 (a teljes teszthalmaz 85%-a) halmazokra. Jelölje a T_1 halmaz címkéit y_1, az ezekhez tartozó predikciókat $\overline{y_1}$! Az y_1

átlagát jelölje A_1, szórását σ_1, $\overline{y_1}$ átlagát $\overline{A_1}$, szórását $\overline{\sigma_1}$! A T_2 halmaz $\overline{y_2}$ predikcióit skáláztam az alábbiak szerint:

$$\widehat{\overline{y_2}} = \left(\overline{y_2} - \overline{A_1}\right) * \frac{\sigma_1}{\overline{\sigma_1}} + A_1 \tag{4.1}$$

ahol $\widehat{\overline{y_2}}$ a T_2 halmaz skálázott predikciói.

Ezáltal a T_2 halmaz predikcióinak átlaga és szórása megegyezik a T_1 valódi címkéinek átlagával és szórásával. A 4.6. táblázatból leolvasható, hogy az utólagos skálázás segít az osztályozás eredményességén, de ront a regresszió eredményein.

RMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
10,43	8,41	0.47	71%	72%	70%

4.6. táblázat A MODMA_MONO modell eredményei az utófeldolgozási skálázás után

Zhenyu Liu és társai kutatásukban ugyanezen az adathalmazon SVM modell használata mellett 54%-62%-os pontosságot értek el, amelyben a 4 különböző típusú hanganyagokat (interjú, folytonos szöveg olvasás, szójegyzék olvasás és képleírás), külön-külön dolgozták fel. Xin Chen és Zhigeng Pana a teljes adathalmazon döntési fák alkalmazása mellett 83%-os pontosságot, 77%-os szenzitivitást és 89%-os specificitást sikerült elérniük. Látható, hogy az általam alkalmazott x-vektor technika használata mellett elért eredmények jobbak az SVM modell által elértektől, de elmaradnak a döntési fákat alkalmazó kutatásban szereplőktől.

4.3 Többnyelvű modellek

A többnyelvű modellek esetében minden esetben 5-fold-os (klasszikus) keresztvalidációt alkalmaztam a hiperparaméterek optimalizációja során.

4.3.1 Adatbázisok előfeldolgozása

Annak érdekében, hogy a különböző depressziós súlyossági skálákat alkalmazó nyelvek hangadatbázisait felhasználhassam egymás modelljeinek tanítására, szükséges volt, hogy azonos skálára alakítsam a címkéiket. Ennek érdekében a DBA és az AVID adatbázisokat a BDI-II skála szerinti címkéiket skáláztam át az alábbiak szerint:

Legyen X egy BDI-II depressziós súlyossági skála szerinti címkék halmaza, amelyet szeretnénk átalakítani a PHQ-8 depressziós súlyossági skála szerinti címkékre!

Ekkor felosztom X elemeit a 4 depressziós súlyossági osztály (nem depressziós, enyhe depresszió, közepes depressziós, súlyos depresszió) szerint, ezeket jelölje rendre X_1, X_2, X_3, X_4! Jelölje az i. súlyossági osztály BDI-II szerinti határait $\min(BDI)_i$ és $\max(BDI)_i$, PHQ-8 szerinti határait $\min(PHQ)_i$ és $\max(PHQ)_i$! Ekkor minden X_i-re az alábbi 2 lépést végeztem el:

$$Z_i = X_i - \frac{\max(BDI)_i + \min(BDI)_i}{2} \div \max(BDI)_i - \min(BDI)_i \quad (4.2)$$

$$\widehat{X_i} = Z_i * (\max(PHQ)_i - \min(PHQ)_i) + \frac{\max(PHQ)_i + \min(PHQ)_i}{2} \quad (4.3)$$

ahol $\widehat{X_i}$ az X_i skálázott elemeit tartalmazza.

A többnyelvű modellek esetében a DBA és az AVID adatbázis tekintetében csak a PHQ-8 depressziós súlyossági skálára átalakított címkékkel dolgoztam.

Annak érdekében, hogy a fent bemutatott skálázás után is összehasonlíthatók maradjanak az egynyelvű és többnyelvű modellek eredményei, a továbbiakban minden modell esetében kiszámítom a NRMSE metrika értékét is az eddig általam használt kiértékelési metrikák mellett.

A többnyelvű modellek esetében minden alkalommal a modell eredményeit a korábbiakban már bemutatott, azonos teszthalmazú, egynyelvű modell eredményeivel fogom összevetni. Az egynyelvű modellek NRMSE értékeit a 4.7. táblázat mutatja be.

DBA_MONO	AVID_MONO	E-DAIC_MONO	MODMA_MONO
0,74	0,77	1,01	0,92

4.7. táblázat Egynyelvű modellek NRMSE értékei

4.3.2 DBA_MULTI modell

A modell tanításakor az AVID, az E-DAIC és a MODMA adathalmazokat használtam fel. ν-SVR-t alkalmaztam rbf kernellel, az előfeldolgozás során nem végeztem külön skálázást az adatokon.

C	2^0
γ	2^{-10}

4.8. táblázat A DBA_MULTI modell optimalizációja során legjobbaknak talált hiperparaméter értékek

A (klasszikus) 5-fold-os keresztvalidáció során a 4.8. táblázatban talált hiperparaméterekkel értem el a legjobb eredményt, amelyet a 4.9. táblázat foglal magába.

RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
6,35	0,79	5,65	0,32	52%	100%	0%

4.9. táblázat A DBA_MULTI modell eredményei

A DBA_MONO NRMSE értéke 0,74, amelyhez képest a többnyelvű modell eredménye, habár rosszabb, de nem számottevően. Azonban a MODMA_MONO modellhez hasonlóan az osztályozás rossz eredményei miatt újra alkalmaztam a korábban már bemutatott utófeldolgozási skálázás módszert, amellyel a teszthalmaz egy kisebb T_1, véletlenszerűen kiválasztott részének felhasználásával skálázom a teszthalmaz nagyobb részét (T_2), ezzel beállítva T_2 predikcióinak számtani átlagát és szórását T_1 címkéi számtani átlagára és szórására.

RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
7,23	0,90	5,56	0,31	64%	63%	65%

4.10. táblázat A DBA_MULTI (skálázott) modell eredményei

A 4.10. táblázatból látható, hogy bár a skálázás ront a NRMSE értékén, de nagyban javítja az osztályozás minőségét.

4.3.3 AVID_MULTI modell

A modell tanításakor az DBA, az E-DAIC és a MODMA adathalmazokat használtam fel. ν-SVR-t alkalmaztam rbf kernellel, az előfeldolgozás során nem végeztem külön skálázást az adatokon.

C	2^{-1}
γ	2^{-10}

4.11. táblázat Az AVID_MULTI modell optimalizációja során legjobbaknak talált hiperparaméter értékek

A (klasszikus) 5-fold-os keresztvalidáció során a 4.11. táblázatban talált hiperparaméterekkel értem el a legjobb eredményt, amelyet a 4.12. táblázat foglal magába.

RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
6,32	0,79	5,66	0,03	48%	100%	0%

4.12. táblázat Az AVID_MULTI modell eredményei

A korábbiakban bemutatott AVD_MONO NRMSE értéke 0,77, ezzel összehasonlítva az AVID_MULTI modell eredménye nagyon hasonló. Azonban az osztályozás minősősége ebben az esetben is indokolttá teszi az utófeldolgozási skálázás alkalmazását.

RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
8,61	1,08	6,68	-0,02	46%	49%	43%

4.13. táblázat Az AVID_MULTI (skálázott) modell eredményei

Az utófeldolgozási skálázás sem igazán segített az osztályozás javításán, amely az alacsony Pearson korreláció tekintetében érthető. A modell eredményeit összefoglalja a 4.13. táblázat.

4.3.4 E-DAIC_MULTI modell

A modell tanításakor a DBA, az AVID és a MODMA adathalmazokat használtam fel. ν-SVR-t alkalmaztam rbf kernellel, az előfeldolgozás során itt sem végeztem külön skálázást az adatokon.

C	2^{10}
γ	2^{-9}

4.14. táblázat Az E-DAIC_MULTI modell optimalizációja során legjobbaknak talált hiperparaméter értékek

A (klasszikus) 5-fold-os keresztvalidáció során a 4.14. táblázatban talált hiperparaméterekkel értem el a legjobb eredményt, amelyet a 4.15. táblázat foglal magába.

RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
6,26	0,77	5,12	0,06	56%	97%	3%

4.15. táblázat Az E-DAIC_MULTI modell eredményei

A korábbiakban bemutatott E-DAIC_MONO NRMSE értéke 1,01, ezzel összehasonlítva a jelenleg vizsgált modell eredménye meggyőző, jobbnak értékelhető. Viszont az osztályozásban elért eredményei okén ennél a modellnél is alkalmaztam a korábban már bemutatott utófeldolgozási skálázást. A skálázás utáni eredményeket a 4.16. táblázat foglalja össze.

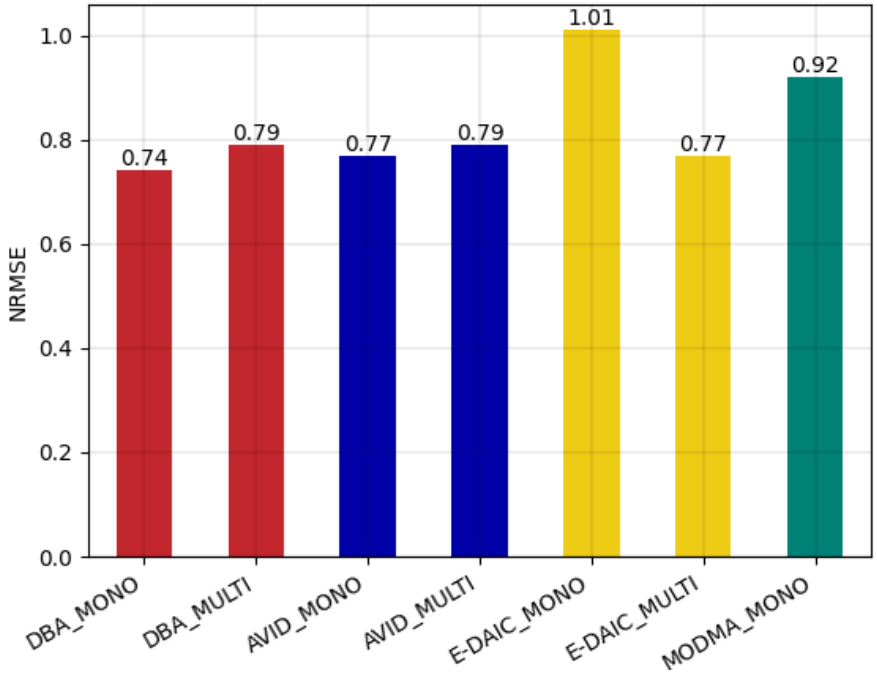
RMSE	NRMSE	MAE	CORREL	pontosság	szenzitivitás	specificitás
9,04	1,12	7,10	0,08	53%	64%	39%

4.16. táblázat Az E-DAIC_MULTI (skálázott) modell eredményei

A skálázás segített némileg az osztályozás eredményein, nőtt a specificitás, miközben érdemben nem csökkent a pontosság.

4.3.5 Összesített eredmények áttekintése

Minden többnyelvű modell esetében elmondható, hogy a referenciaként használt azonos, de egynyelvű modellekhez képest a performanciájuk ugyan elmarad, de nem jelentős mértékben az NRMSE metrika tekintetében. Az osztályozás során azonban rendre rossz eredményeket értek el a többnyelvű modellek, amelynek oka lehet a különböző depressziós súlyossági skálák használata, amelyek miatt be kellett vezetnem egy plusz előfeldolgozási skálázást, valamint a különböző nyelvek azon nyelvspecifikus tulajdonságai, amelyeket tartalmaznak az x-vektorok, de függetlenek a beszélő egészségügyi helyzetétől. Az összes bemutatott modell NRMSE értékeit foglalja össze a 4.1 diagram.



4.1. ábra A létrehozott modellek NRMSE értékeinek összefoglalása

Összefoglalás

A depresszió korunk egyik leggyakoribb pszichiátriai megbetegedése. A tünetek súlyosságától függően hatással van a beteg életminőségére és munkaképességére is. A depresszió időben való diagnosztizálása kiemelt jelentőségű a beteg gyógyulásának érdekében, azonban mivel az orvosi diagnosztizálás idő- és költségigényes, fontos kutatási terület, hogyan lehet támogatni ezt a folyamatot automata rendszerekkel.

Írásomban bemutattam a beszéd egy alacsony szintű jellemzőjét, a mel frekvenciás kepsztrális együtthatókat, amelyek használata széles körben elterjedt az automatikus beszédfeldolgozásban. Ismertettem a depresszió klinikai diagnózisához szükséges feltételeket és három, különböző depressziós súlyossági skálát is. Ugyancsak bemutattam egy olyan gépi tanuló eljárást, a szupport vektor gépet, amelynek egy módosított változatával az iparban is gyakran oldanak meg regressziós feladatokat.

A dolgozatom készítése folyamán egy gépi tanuló eljárást alkalmazva több megoldást is kipróbáltam a depresszió beszédproduktum alapú felismerésének lehetőségeit vizsgálva. Reprezentációként x-vektorokat alkalmaztam, amely mély neurális háló alapú technikával lehetőség van a változó hosszú hangfelvételeket fix dimenziószámú jellemzőtérbe képezni. Az x-vektorokat egy SVR alapú depresszió felismerő modell bemeneteként alkalmaztam, amely eredményein több lépésben megkíséreltem javítani is. Többféle kernelt is kipróbáltam, hiperparaméter-optimalizációt végeztem, előfeldolgozási lépésként skáláztam a bemeneti x-vektorokat és a célváltozót is, keresztvalidációt alkalmaztam a kiértékelés során. Egyes esetekben bevezettem egy utófeldolgozási lépést is, amely a becsült érétkek egy újabb skálázását jelentette. A kapott predikciókat több, széles körben bevett kiértékelési metrika szerit is kiértékeltem, valamint az értékeik alapján osztályzást is végeztem.

Rendelkezésemre állt egy magyar, egy német, egy angol és egy kínai nyelvű depressziós beszédadatbázis, amelyek felhasználásával egynyelvű és többnyelvű modelleket is alkottam. Az egynyelvű modelljeim eredményeit összevetettem egyéb, ezen a területen végzett kutatások eredményeivel. Az magyar adatbázis esetében a hiperparaméter-optimalizáció után 9,7 RMSE értéket sikerült elérnem, amely jobb, mint az ugyanezt az adatbázist feldolgozó egyéb kutatások eredményei. Hasonlóan a német

adatbázis esetében elért 9,27 RMSE érték olyan, amellyel érdemben képes lettem volna versenyezni a 2013-as AVEC-DSC versenyében.

A többnyelvű modelljeim eredményeit azon egynyelvű modelljeim eredményeimmel vetettem össze, amelyeknél a teszthalmaz nyelve megegyezik az egynyelvű modell nyelvével. Mind a DBA_MULTI, mind az AVID_MULTI modellek esetében sikerült közel azonos NRMSE értékű modelleket alkotnom, mind egynyelvű párjaik, sőt az E-DAIC_MULTI modell jobb NRMSE értéket is ért el az E-DAIC_MONO modellhez képest.

A munkám során megismert és az általam elért eredmények alapján a depresszió beszéd alapú automatikus felismerésének területe egy fontos, de egyelőre gyakorlatban alkalmazott megoldásokat még nélkülöző kutatási terület. A kutatások fő hátráltató tényezője a rendelkezésre álló adatok alacsony száma, ennek megfelelően a modellek által elért eredmények további adatbővítésekkel feltehetőleg javíthatóak. A kutatások mögötti motiváció, a depresszió minél korábbi, gépi eljárásokkal támogatott felismerése miatt ezen kísérletek folytatása egy kihívásokkal teli, de érdekes és törődésre érdemes munka.

5 Irodalomjegyzék

[1] K. Géza. (2010) A beszéd és az információs társadalom. In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó, 3-7.

[2] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech communication, 71, 10-49.

[3] Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS medicine, 3(11), e442.

[4] World Health Organization. (2017). Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). World Health Organization.

[5] 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.

[6] Verde, L., De Pietro, G., & Sannino, G. (2018). Voice disorder identification by using machine learning techniques. IEEE access, 6, 16246-16255.

[7] Dávid, S., Gábor, K., Gábriel, T. M., & Klára, V. Betegségek automatikus szétválasztása időben eltolt akusztikai jellemzők korrelációs struktúrája alapján.

[8] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., ... & Pantic, M. (2013, October). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (pp. 3-10).

[9] Kiss, G., & Vicsi, K. (2017). Mono-and multi-lingual depression prediction based on speech processing. International Journal of Speech Technology, 20(4), 919-935.

[10] Alghowinem, S., Goecke, R., Epps, J., Wagner, M., & Cohn, J. F. (2016, September). Cross-cultural depression recognition from vocal biomarkers. In Interspeech (pp. 1943-1947).

[11] James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., ... & Briggs, A. M. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 392(10159), 1789-1858.

[12] Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jönsson, B., CDBE2010 Study Group, & European Brain Council. (2012). The economic cost of brain disorders in Europe. European journal of neurology, 19(1), 155-162.

[13] Hawton, K., i Comabella, C. C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: a systematic review. Journal of affective disorders, 147(1-3), 17-28..

[14] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. Archives of general psychiatry, 4(6), 561-571.

[15] Hamilton, M. (1960). A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry, 23(1), 56.

[16] Moyer, M. (2020). Comparison of PHQ-8 and Beck Depression Inventory II in the Army National Guard (Doctoral dissertation, Rutgers University-School of Nursing-RBHS).

[17] O. Gábor, (2010) A beszédképzés folyamata. In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó. 19-27.

[18] Afshan, A., Guo, J., Park, S. J., Ravi, V., Flint, J., & Alwan, A. (2018). Effectiveness of voice quality features in detecting depression. Interspeech 2018.

[19] O. G. (szerk) Németh Géza. (2010) A beszéd komplex szerkezete. In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó, 9-10.

[20] Kiss, G. (2019). A depressziós beszéd akusztikai-fonetikai jellemzőinek vizsgálata.

[21] V. Klára, (2010) Pszichofizikai tényezők In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó, 56-69.

[22] Pedersen, P. (1965). The mel scale. Journal of Music Theory, 9(2), 295-308.

[23] Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech communication, 54(4), 543-565.

[24] S. György, (2010) Kepsztrum In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó, 239-240.

[25] N. G. –. O. G. (szerk.) (2010) MFCC-paraméterek In A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Budapest, Akadémai Kiadó, 240-242.

[26] Snyder, D. (2020). X-Vectors: Robust neural embeddings for speaker recognition (Doctoral dissertation, Johns Hopkins University).

[27] Moro-Velazquez, L., Villalba, J., & Dehak, N. (2020, May). Using x-vectors to automatically detect parkinson's disease from speech. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1155-1159). IEEE.

[28] Egas-López, J. V., Kiss, G., Sztahó, D., & Gosztolya, G. (2022, May). Automatic Assessment of the Degree of Clinical Depression from Speech Using X-Vectors. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8502-8506). IEEE.

[29] Russell, Stuart J. (Stuart Jonathan). (2010). Artificial intelligence : a modern approach. Upper Saddle River, N.J. :Prentice Hall.

[30] Mjolsness, E., Sharp, D. H., & Alpert, B. K. (1989). Scaling, machine learning, and genetic neural nets. Advances in applied mathematics, 10(2), 137-163.

[31] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

[32] Novak, E. (2006). Deterministic and stochastic error bounds in numerical analysis (Vol. 1349). Springer.

[33] Michael. Steinbach, Vipin Kumar, Pang-Ning Tan, (2011) Egy osztályozó teljesítményének a kiértékelése In Bevezetés az adatbányászatba, Budapest, Panem Könyvkiadó Kft.

[34] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. (2014). The distress analysis interview corpus of human and computer interviews. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

[35] Cai, H., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., ... & Hu, B. (2020). Modma dataset: a multi-modal open dataset for mental-disorder analysis. arXiv preprint arXiv:2002.09283.

[36] Hajduska-Dér, B., Kiss, G., Sztahó, D., Vicsi, K., & Simon, L. (2022). The applicability of the Beck Depression Inventory and Hamilton Depression Scale in the automatic recognition of depression based on speech signal processing. Frontiers in Psychiatry, 1767.

[37] Kiss, G., Sztahó, D., & Tulics, M. G. (2021). Application for Detecting Depression, Parkinson's Disease and Dysphonic Speech. In Interspeech (pp. 956-957).

[38] F Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., ... & Pantic, M. (2019, October). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop (pp. 3-12).

[39] Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019, October). Multi-level attention network using text, audio and video for depression prediction. In Proceedings of the 9th international on audio/visual emotion challenge and workshop (pp. 81-88).

[40] Liu, Z., Wang, D., Zhang, L., & Hu, B. (2020). A novel decision tree for depression recognition in speech. arXiv preprint arXiv:2002.12759.

[41] Chen, X., & Pan, Z. (2021). A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. International Journal of Environmental Research and Public Health, 18(12), 6441.