Обрада природног језика и откривање раних знакова деменције

Ученик: Милена Пантић *Ментор:* Анђелка Зечевић

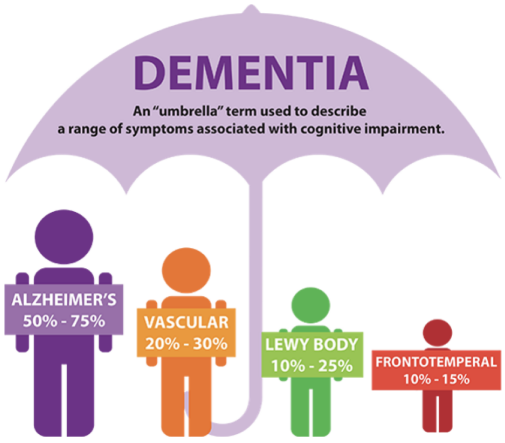
Београд, мај 2026.

Садржај

1 Увод

Деменција подразумева скуп неуродегенеративних болести које утичу на памћење, мишљење и друге когнитивне способности. Ове болести временом уништавају нервне ћелије и тиме доводе до оштећења и слабљења функција мозга. Неретко деменцију прате и промене у расположењу, понашању и губитак контроле емоција. Према изворима Светске здравствене организације, у 2021. години забележено је 57 милиона особа са деменцијом. Број оболелих се сваке године повећа за око 10 милиона, што је већински проузроковано старењем популација, а процењује се да ће у 2050. години чак 152 милиона људи имати деменцију.

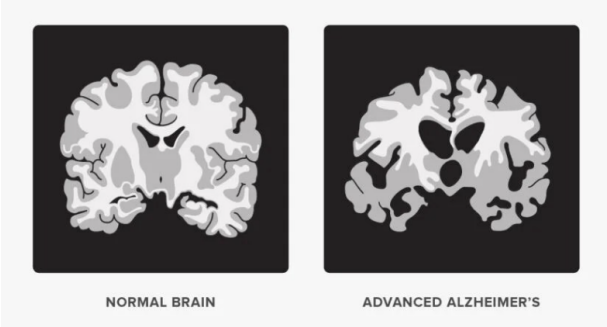
Деменција се најчешће среће код особа старости преко 65 година. Како је болест прогресивна, оболелима су временом све чешће потребни неговатељи који ће помагати при свакодневним активностима. У каснијим фазама болести особе са деменцијом могу и да заборављају лица, да се тешко покрећу, постану агресивнији и теже обављају основне биолошке функције. Тренутно не постоји универзални лек за деменцију. Ипак, постоје лекови (као што су инхибитори холинестеразе, антагонисти НМДА рецептора, селективни инхибитори преузимања серотонина - ССРИ) који помажу око специфичних симптома који могу да се јаве током болести. Поред лекова, у успоравању симптома деменције помаже одржавање здравља и квалитета живота редовним кретањем, социјалном активношћу и другим здравим навикама. Слично, особа може да смањи ризик добијања деменције тако што ће бити физички, когнитивно и социјално активна и водити здрав живот.



Слика 1: Различити типови деменције

Постоје различите врсте деменције, од којих су најпознатије васкуларна, деменција са Левијевим телима, фронтотемпорална деменција и Алцхајмерова болест. **Алцхајмерова болест** је најчешћи узрочник деменције и чини између 60% и 80% свих случајева деменције. Њоме ћемо се и бавити у овом раду. Постоје и такозване друге деменције (енг. other dementia) које укључују неке ређе облике деменције и мешовите деменције (енг. mixed dementia) које имају више узрочника.

Код Алцхајмерове болести долази до озбиљних промена у мозгу много пре него што се јаве први симптоми, чак и 10 до 20 година пре првих когнитивних поремећаја. Неурони обамиру чиме се изазива губитак памћења и долази до промена у понашању особе. Није откривено који је тачно узрок ове болести, али познато је да су две структуре, плакови и замршења или клупка, распрострањеније код особа са Алцхајмеровом болешћу, док се у мањој мери јављају нормалним процесом старења. Плакови (енг. plaques) су наслаге протеина бета-амилоида које се јављају у просторима између неурона и која доводе до ометања комуникације између неурона, упалних процеса и временом одумирања нервних ћелија. Замршења или клупка (енг. tangles) су увијени тау протеини накупљени у ћелијама, који ометају пренос хранљивих материја нервних ћелија што такође временом доводи до њиховог одумирања. Новији лекови на тржишту као што су *Donanemab* и *Lecanemab* се ослањају на ове механизме и тиме побољшавају стање пацијената.



Слика 2: Утицај Алцхајмерове болести на мождане ћелије

Типични рани знак Алцхајмерове болести је оштећено краткорочно памћење. Како се болест прогресивно шири од хипокампуса до осталих делова мозга, временом долази до дезоријентације, конфузије, промена расположења, а касније тешкоћа са говором, гутањем, и ходањем. Препознавање раних знакова деменције може бити веома значајно у побољшању квалитета живота појединца због могућности финијег планирања тока лечења и више опција третмана. У ране симптоме болести спадају и заборављање, губљење ствари, губљење у простору и времену, проблеми са конверзацијама, тешкоће при доношењу одлука, мањак емпатије и социјална изолација. Симптоми Алцхајмерове болести крећу се од поремећаја говора, резоновања, планирања до халуцинација, промена понашања и поремећаја кретања (нпр. падање).

2 Рана детекција деменције

Претклиничка фаза и фаза благог когнитивног поремећаја (енг. mild cognitive impairment, MCI) су фазе деменције које се јављају од 10 до 20 година пре првих значајних когнитивних поремећаја и кључне су за рано идентификовање Алцхајмерове болести. У овим фазама се постижу најбољи резултати лечења постојећим терапијама јер је патологија у мозгу најмања. Такође, промене начина живота у овом периоду могу спречити чак 40 % случајева деменције.

Тренутно се Алцхајмерова болест дијагностикује тек након појаве меморијских оштећења. За процену озбиљности оштећења се користи званичне медицинске смернице (нпр. The Diagnostic and Statistical Manual of Mental Disorders) које дефинишу и процедуре и критеријуме за успостављања дијагнозе. Обављају се клиничке процене, прикупља се историја когнитивних симптома, врше се физички прегледи и когнитивни тестови, анализе крви (због искључивања других симптома који могу да имитирају деменцију) и скенирање мозга. Када је могуће, ради се и анализа течности кичмене мождине и генетски скрининг. Процес дијагностике је дуг, скуп и донекле субјективан. На пример, пацијент ће бити дијагностикован у зависности од тога како лекар процени резултате когнитивног теста или теста конверзације. Између 29% и 76% особа које имају деменцију нису препознати као дементни у данашњој клиничкој пракси.

Когнитивни тестови који се обично срећу у пракси детекције деменције су MMSE, ACE-R, MoCA и тест цртања сата, а данас су неки доступни и преко мобилних телефона. MMSE (акроним од Mini Mental State Examination) тест мери меморију, оријентацију, језик, пажњу и даје резултат од укупно 30 бодова. Средња вредност резултата пацијената са демецијом је 9.7, док је за здраве испитанике значајно виша. ACE-R (акроним од Addenbroke's Cognitive Examination-Revised) тестира пажњу, меморију, језик и визуелно-просторне функције и даје резултат од укупно 100 бодова. MoCA (акроним од Montreal Cognitive Assessment) мери краткорочно памћење, пажњу, језик, фонемичку течност и визуоспацијалну функцију и враћа вредност резултата од укупно 30 бодова. Тест цртања сата (енг. clock drawing test) мери колико тачно је особа нацртала сат, бројеве и казаљке на сату за задато време.

Сви наведени тестови захтевају присуство особе која би интерпретирала резултате и администрирала тестирање. Ниједан од ова четири теста не мери паузе и време које треба особи да се сети тачног одговора, што може бити битан индикатор болести. Такође, тест није довољно сензитиван тј. особа, поготово у ранијим фазама болести, може бити процењена као да није дементна иако то јесте случај.

Вештачка интелигенција (енг. artificial intelligence, AI) би могла да допринесе бржој, прецизнијој, јефтинијој и приступачној дијагностици деменције. Вештачка интелигенција је грана рачунарских наука која омогућава машинама да обављају задатке за које је обично потребна људска интелигенција препознавањем образаца из велике количине података,

без експлицитног програмирања свих могућих случајева. Ова област обухвата неколико кључних грана: машинско учење (енг. Machine Learning, ML), дубоко учење (енг. Deep Learning, DL), обраду природног језика (енг. Natural Language Processing, NLP) и компјутерски вид (енг. Computer Vision, CV).

Машинско учење је грана вештачке интелигенције која се бави анализом алгоритама учења из података, методологијом дизајнирања процеса учења и његовом евалуацијом. Неки од популарних задатака машинског учења су класификација, регресија и кластеринг, док су неки од популарних алгоритама метод потпорних вектора (енг. support vector machine, SVM), случајне шуме (енг. random forest), метод најближих суседа (енг. k-nearest neighbour, kNN), логистичка регресија (енг. logistic regresion), наивни Бајес (енг. naive Bayes), к-средина (енг. k-means) и остали. У раду ћемо се бавити алгоритмима класификације, логистичком регресијом и стаблима одлучивања. Дубоко учење (енг. deep learning) је грана машинског учења које подразумева методе засноване на неуронским мрежама као што су конволутивне неуронске мреже (енг. convolutional neural networks, CNN), рекурентне неуронске мреже (енг. recurrent neural networks) и тренутно популарни трансформери (енг. transformers). **Обрада природног језика** омогућава рачунарима да интерпретирају и генеришу језик, говор и текст, док компјутерски вид омогућава интерпретацију визуелних информација попут слика и видеа.

У случају деменције, приступи засновани на вештачкој интелигенцији се сусрећу код дизајнирања и анализе когнитивних тестова на компјутеру, интерпретације слика као што су скенови мозга или снимци мрежњаче, тестовима покрета и говора и специфичним медицинским тестовима (EEG, процена нивоа амилоид-бета и тау у CSF протеинима или процена нивоа протеина у цереброспиналној течности и крви). Коришћење вештачке интелигенције даје објективније, јефтиније и често прецизније резултате јер пружа могућност да детектујемо додатне индикаторе болести у великим скуповима података. Ови индикатори се често називају и дигитални биомаркери јер се на основу њих и њихових вредности може дијагностиковати болест. Сви поменути примери би требали да олакшају лекарима да утврде постојање деменције и да лакше и поузданије прате ток развоја болести.

Компјутеризовани когнитивни тестови (CogState, CANTAB, компјутеризовани тест цртања сата) поред праћења и мерења когнитивних способности могу да прате и процењују понашање особе током тестирања, као што је, на пример, брзина реакције. Овакви компјутеризовани тестови, засновани на методама вештачке интелигенције, имају побољшану сензитивност (за 4%) тј. способност тачног идентификовања особа са деменцијом и специфичност (за 3%) тј. способност тачног идентификовања особа које нису дементне. Тест CogState мери визуално-моторну функцију, пажњу, меморију и социјалну когницију, а CANTAB тестира меморију, вербалну меморију, пажњу, време обраде информација, способност препознавања емоција, доношења одлука и контролисања одговора. Компјутеризовани тест цртања сата се ради на таблету и тестира визуелно-просторну функцију, памћење и пажњу.

Тестови покрета испитују ход, покрете руке и очију. Користе акцелерометре за мерење разлика у брзини хода и магнетне сензоре за мерење таласних облика куцања прстију. Поред дијагнозе, ови тестови могу да разликују и разне типове деменције. Међутим, постоје потребе за додатном скупом опремом да би се остварили жељени резултати.

Тестови говора, којима ћемо се бавити у раду, користе обраду природног језика да извуку биомаркере спонтаног говора попут брзине говора, дужине реченица и синтаксичке сложености. Технике обраде природног језика могу да детектују суптилне промене у говору које су индикатори деменције, а које би иначе могле да промакну стручним лицима.

Интерпретације скенова мозга могу користити методе машинског и дубоког учења за класификацију пацијената по томе да ли и који тип деменције имају да би спречили субјективност и евентуалне грешке код дијагностике раних симптома које карактерише слабија атрофија и суптилнији знаци болести. Код Алцхајмерове болести, скенирања мозга техникама MRI и CT могу указати на смањени хипокампус, док PET скенер, који је скупљи и мање доступан, пружа информацију о протоку крви и нивоу амилоида. Због комплексности информација које неуроимаџинг чини доступним, овакве методама се теже издвајају специфични фактори.

Од 58% до 79% ризика Алцхајмерове болести је наследно. Неке методе вештачке интелигенције, попут метода потпорних вектора, могу да се користе у сврху идентификовања гена који су одговорни за наслеђивање деменције. Алцхајмерова болест се дијагностикује и преко **EEG теста** који детектује неправилности у можданим таласима и електричној активности мозга. **Снимање мрежњаче** је новија и јефтинија замена за неуроимиџинг које у комбинацији са конволутивним неуронским мрежама такође даје добру дијагностику Алцхајмерове болести.

Мерење нивоа амилоид-бета и тау у CSF протеинима је релативно успешна метода, међутим веома је скупа и понекад инвазивна. Поред оваквих инвазивних метода, може се мерити **ниво протеина у цереброспиналној течности или крви** уз подршку вештачке интелигенције.

3 Улога технологије у раној детекцији деменције

Могућност да се складиште и обрађују велике количине података заједно са прогресима у свету развоја алгоритама и вештачке интелигенције, довели су до иновативних приступа ране детекције деменције. Упоредо са основним медицинским процедурама које укључују преглед пацијента, увид у електронски здравствени картон, лабораторијске анализе, когнитивне тестове и медицинску дијагностику, могуће је користити системе за подршку одлучивању. **Системи за подршку одлучивању** су посебна софтверска решења која на основу података и специјално дизајнираних алгоритама могу да асистирају лекарима експертима у задацима прецизније дијагностике унапређујући на тај начин квалитет и безбедност здравственог система.

У литератури и прегледним радовима који повезују технологију и деменцију се најчешће сусрећу две групе софтверских алата. Прва група алата помаже лекарима да утврде да ли пацијент има деменцију или не. У основи ових алата су такозвани класификациони алгоритми надгледаног машинског учења о којима ће бити више речи у наставку. То су, пре свега, алгоритми логистичке регресије (енг. logistic regression), метод потпорних вектора (енг. support vector machine, SVM), стабла одлучивања (енг. decision trees) и случајне шуме (енг. random forest) или напреднији алгоритми попут неуронских мрежа. Њихове одлуке могу бити базиране на медицинским сликама, вредностима лабораторијских анализа, генетским тестовима, анализи говора или на комбинацији неких од наведених типова података. Друга група алата помаже лекарима да процене прогрес болести код пацијената којима је већ потврђена деменција. Ови алати користе регресионе алгоритме ненадгледаног машинског учења укрштене са стандардним тестовима за процену когнитивних способности. На основу њих се добијају графици који показују ток развоја болести у времену. Важно је напоменути да је за успешан развој ових алата потребно коришћење квалитетних скупова података, што због сложености болести, потребе за посебном експертизом лекара, трошкова, али и законских оквира, није ни једноставно ни често полазиште. Зато у заједници постоје многе иницијативе којима би постојеће стање могло да се поправи. Такође, с обзиром на могуће ризике примена ових алата, заједница ради и на развоју алгоритама који уз високу прецизност имају и нека друга пожељна својства. На пример, интерпретабилност (могућност алата да предочи начин израчунавања резултата и добијања финалне предикције), робусност (могућност алата да одржи очекивано понашање и када неки од улаза није потпун или је драстично другачији), објашњивост (могућност алата да генерише образложење зашто је неки алат дао баш добијени резултат нпр. могућност да процени како одређена својства улаза утичу на коначни резултат). Овакви алгоритми помажу лекарима да боље разумеју резултате коришћених алата и да информисаније и са више поверења доносе одлуке.

Алгоритми за рад са сликама користе слике добијене помоћу магнетне резонанце (енг. magnetic resonance imaging, MRI), рачунске томографије (енг. computed tomography, CT)

или PET (енг. positron emission tomography) скенера који су новијег датума. Ове слике могу бити у 2D или у 3D формату и на њима се могу видети структурне промене мозга. Неки од познатих скупова података су ADNI који је креирала Иницијатива за неуроимаџинг Алцхајмерове болести (енг. Alzheimer's Disease Neuroimaging Initiative) и OASIS настао у склопу Серије имиџинг студија у отвореном приступу (енг. Open Access Series of Imaging Studies). С обзиром на формат слика и ниво детаља којим располажу, за њихову обраду се најчешће користе неуронске мреже и конволутивне неуронске мреже или визуелни трансформери.

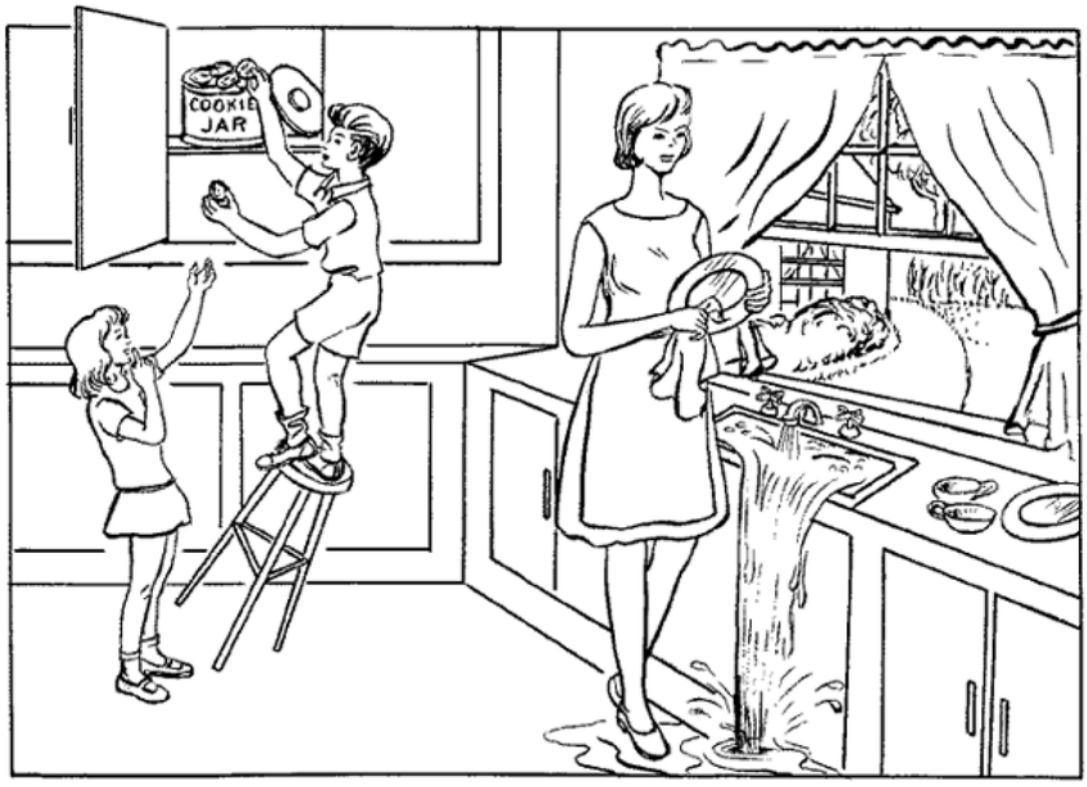
Коришћење медицинских слика за сам процес дијагностике, на жалост, није увек могућ услед различитих организација здравствених система и свеукупних капацитета. Због тога се заједница окреће и коришћењу алтернативних доступнијих и мање инвазивних процедура. Иако су слабљење памћења и когнитивних способности често први приметни симптоми деменције, проблеми са говором су, такође веома, чести. Зато се говор користи и као дигитални биомаркер болести. Код говора се могу анализирати акустички аспект (квалитет звука, брзина говора, прозодија), лингвистички аспект (граматичност, сложеност реченица, богатство вокабулара) или њихова комбинација, када се заправо постижу и најбољи резултати. За ове сврхе се користе скупови података који настају у оквиру пројекта Банка деменције (енг. Dementia Bank) о којем ће бити више речи у наставку. Палета алгоритама која се користи за ове сврхе је изразито богата захваљујући популарности обраде природних језика, подобласти вештачке интелигенције која се бави говором и текстом.

Постоји и велики број пројеката који укључује анализирање генетских података. Ова врста података је важна за разумевање узрока и тока болести, као и за дизајнирање персонализованих лекова. У популарним биоинформатичким базама података као што су DisGeNet, OMIM, DISEASES и друге, могу се наћи информације о генима који се доводе у везу са деменцијом и сродним болестима. Тако, на пример, мутације на генима APP, PSEN1 или PSEN2 готово извесно доводе до ране Алцхајмерове болести. Најпопуларнији скуп података овог типа је NIAGADS који је креирао Национални институт за генетику старења (енг. National Institute on Aging Genetics) а који садржи око 58 хиљада секвенцираних генома.

4 Тестови говора, језика и конверзације

Неки од првих знакова болести током ране фазе деменције јесу поремећаји говора и језика. Примећено је да говор особа са деменцијом карактерише понављање речи, дуже паузе, монотона фреквенција гласа, смањени вокабулар, кратке реченице и просте реченичне конструкције. Да бисмо одредили и ”измерили” карактеристике говора помоћу којих можемо да одредимо да ли особа има деменцију или не, потребно је да особу на одређени начин тестирамо.

Тестови говора, језика и конверзације могу бити тестови описивања слике (content--based) или интервју тестови (content-free). Током **теста описивања слике** учесницима се снима глас док за ограничено време спонтаним говором описују одређену слику - често се користи слика ”крадљивац колачића” (енг. cookie theft) која је и део дијагностичког тест афазије. **Интервју тестови** су снимани разговори неуролога и пацијента. Овакви тестови мере способност пацијента да одговори на питања, одржава спонтани говор и не ”одлута” од теме разговора. Поред поменутих, постоје и **тестови слободне конверзације**, где особа прича о теми коју одабере и **наративни тестови** у којима препричава одређени догађај.



Слика 3: Слика Cookie Theft

5 Екстракција карактеристика говора

Поменута **обрада природног језика**, област вештачке интелигенције која се бави обрадом текста и говора, нам помаже да својства тј. карактеристике текста запишемо нумеричким вредностима које даље модели машинског учења могу да обраде и помоћу њих класификују пацијенте. Индикаторе деменције попут подрхтавање гласа, оклевања у говору, смањеног број слогова по јединици времена, пауза у говору, брзине говора, стопе грешака, записујемо у текстуални или аудио фајл, па на основу њих рачунамо разне метрике које нам дају информацију о томе да ли је особа дементна. Карактеристике издвајамо ручно, применом алгоритама обраде природног језика или, у случају аудио фајлова, применом алгоритама и алата за обраду аудио сигнала. Овакве карактеристике тј. биомаркере делимо на акустичке и лингвистичке. **Акустичке карактеристике** говора описују артикулацију говора и звукове, док **лингвистичке карактеристике** говора подразумевају садржај говора, вокабулар, граматику, синтаксу и садржај. Акустичне карактеристике укључују амплитуду аудио сигнала, фреквенцију гласа и интензитета звука, средњу вредност дужине пауза (на пример, дуже од 250ms), варијацију висине и јачине. Неке од лингвистичких карактеристика су средња вредност броја речи по реченици, број реченица у тексту према укупном броју, број везника према укупном броју речи, број понављаних N-грама према укупном броју речи (где је N=2,3 или 4), број уникатних речи према укупном броју речи и друге.

Након издвајања карактеристика може да се тренира модел класификације машинског учења који ће свакој карактеристици доделити тежину, број који описује колико одређена карактеристика утиче на закључивање модела и његову предикцију. Након тренирања, класификатори машинског или дубоког учења се користе за категоризацију особа по томе да ли имају болест или не. За постизање највеће тачности резултата је битно на који начин бирамо карактеристике тј. биомаркере а најбољи резултати се добијају коришћењем и акустичких и лингвистичких карактеристика. Модели засновани на неуронским мрежама имају могућност учења и издвајања карактеристика.

5.1 Припрема текста и аудио фајлова

Транскрибовани аудио фајлови, добијени ручном транскрипцијом или коришћењем алата за аутоматско препознавање говора, се пре него што се даље обраде, припремају у форму текста која је погодна за препознавање говорних карактеристика. **Припрема** може да подразумева брисање знакова интерпункције, брисање честих речи (стоп речи), претварање великих слова у мала, поделе текста на реченице и речи (токенизација и сегментација) и довођене речи на њихов основни облик (стеминг). Такође, ако је потребно, свакој речи можемо доделити врсту речи (енг. Part-of-Speech, POS) (глагол, именица, придев и слично) коришћењем такозваних **POS тагера**, програма који се развијају за ове сврхе у

области обраде природног језика. Постоје разне врсте POS тагера од којих ћемо овде описати оне засноване на правилима и статистичким прорачунима. **POS заснован на правилима** користи граматичка правила на основу којих разврстава речи у класе. Правила могу описивати неке морфолошке особине речи, попут одређених наставака речи (нпр. наставак -ed за глаголе за прошло време у енглеском језику) или синтаксичке особине као што је улога речи пре одабране речи (нпр. a и the ће означавати именицу у енглеском језику). Међутим, граматичка правила често имају изузетке што је мана овог приступа. **Статистички POS** учи граматичку класификацију речи на посебно припремљеним подацима па онда свакој речи додељује вероватноћу одређене граматичке категорије.

У случају аудио фајлова, припрема подразумева уједначавање фреквенције аудио записа (обично 16 kHz), уклањање шума, сегментирање говора и пауза (деоница звука са тишином).

5.2 Репрезентација текста

Да би текст могао да се обрађује алгоритмима машинског учења, потребно је да се представи у одговарајућој нумеричкој форми. Технике **репрезентације текста** претварају текст у векторе нумеричких вредности и дефинишу начине на основу којих касније можемо да израчунамо неку нумеричку вредност која представља одређену карактеристику говора. Постоје репрезентације **засноване на фреквенцијама речи (BOW, TF-IDF)**, **векторским репрезентацијама речи (Word2Vec, GloVe, FastText)** и **контексту (BERT, GPT, ELMO)**. У овом раду ћемо говорити само репрезентацијама заснованим на фреквенцији (енг. frequency based embeddings).

BOW (енг. bag of words) репрезентација задати текст претвара у неуређени скуп речи (вокабулар). Вокабулар се састоји из свих уникатних речи текста па се свакој од њих додељује вредност у матрици карактеристика која представља колико пута је та реч искоришћена у неком тексту. На овај начин се стиче пуни утисак о присутности речи у неком тексту али се губи информација о граматичкој и семантичкој функцији речи. Да бисмо побољшали BOW модел можемо користити N-граме који чувају ред речи тако што екстракују N узастопних речи уместо појединачних речи. Матрица N-грама може садржати доста нула (енг. sparse матрица) па се обично изостављају превише чести N-грами који означавају стоп речи и превише ретки N-грами који обично означавају грешке у писању. BOW користи фреквенцију речи у документу, што је представљено овом функцијом:

$$BoW(w, d) = count(w \in d)$$

где је w представља реч, а d документ.

TF-IDF (енг. term frequency-inverse document frequency) речима додељује број који представља њихов значај који је у овом случају пропорционалан броју појављивања речи у документу, а обрнуто пропорционалан појављивања речи у читавом скупу докумената. Функција TF-IDF алгоритма је представљена са:

$$tfidf(w,d) = tf(w,d) * idf(w,d)$$

$$tf(w,d) = count(w \in d)$$

$$idf(w) = \ln \frac{1+n}{1+df(w)} + 1$$

где је n број докумената у скупу докумената, а $df(w)$ број докумената који садржи реч w. Приметимо да је TF-IDF функција максимална када се реч често појављује у документу али ретко у свим документима што нам говори да је реч информативнија.

6 Машинско учење и задатак класификације

Машинско учење је област вештачке интелигенције која се бави развојем алгоритама тј. модела који могу да уче на великим скуповима података како би могли да обављају одређени задатак, попут предикције или генерисања садржаја. Машинско учење је велика и динамична област истраживања у којој се посебно издвајају надгледано учење, ненадгледано учење, поткрепљено учење и генеративна вештачка интелигенција. Навешћемо укратко шта подразумева сваки од ова четири типа машинског учења али ћемо се после бавити само надгледаним машинским учењем.

Надгледано машинско учење (енг. supervised machine learning) прави предвиђања користећи податке са већ одређеним ”тачним” одговорима. Користи се приликом решавања задатака регресије и класификације. Проблеми регресије се баве предвиђањем одређене нумеричке вредности (нпр. предвиђање цене), док се проблеми класификације баве одређивањем вероватноће припадности одређеној категорији (нпр. класификација мејлова у spam и не-spam).

Ненадгледано учење (енг. unsupervised machine learning) се бави уочавањем образаца у скупу података и дизајнирањем алгоритама за уочавање кластера - скупова сродних података.

Поткрепљено учење (енг. reinforcement learning) предвиђа одређене резултате тако што модел награђујемо или кажњавамо након добитка задовољавајућих или незадовољавајућих одговора. На основу претходног искуства, модел генерише правило понашања по коме ће имати највећу награду.

Генеративна вештачка интелигенција (енг. generative AI) генерише текст, слике, видее, код или говор помоћу анализирања образаца постојећих података из скупа података које модел даље имитира и тиме генерише нове садржаје.

6.1 Надгледано машинско учење

Као што смо поменули, надгледано машинско учење итеративним тренирањем над скуповима обележених података постепено проучава однос између улазних и излазних података. Модели надгледаног машинског учења укључују линеарну регресију, логистичку регресију, стабла одлучивања, насумичне шуме, SVM, kNN, и наивни Бајес. У наставку ћемо навести основну терминологију коју користимо када говоримо о надгледаном машинском учењу.

Модел је механизам машинског учења чији је задатак да пронађе математичку зависност одређених карактеристика које се у модел уносе, и да накнадно, помоћу тренирања, научи да избацује вредност најближу стварној вредности. Модели машинског учења за тренирање користе податке које се састоје од **карактеристика** (као што су речи и бројеви у табелама, или пиксели и звучни таласи), тј. вредности које модел користи при предвиђању, и **ознака** (класа или бројева), тј. вредности које модел треба да предвиди. Скуп свих података које модел користи приликом тренирања назива се **скуп података (енг. dataset)**. Да би модел био што успешнији, потребно је да наш скуп података буде разноврстан и опсежан, и да садржи релевантне карактеристике, тј. оне које ће битно утицати на предикције резултата. Модел се **тренира**, тј. учи, на делу скупа података који је обележен, док се други део (који није обележен) користи за тестирање модела. Тренирање подразумева да се у свакој итерацији упоређују предикције модела са тачним вредностима (ознакама) скупа за тренирање, након чега се параметри модела ажурирају минимизацијом функције грешке. Након тренинга вршимо **евалуацију** модела, тј. процењујемо колико добро обавља дати задатак над подацима које раније није видео.

6.2 Бинарна класификација

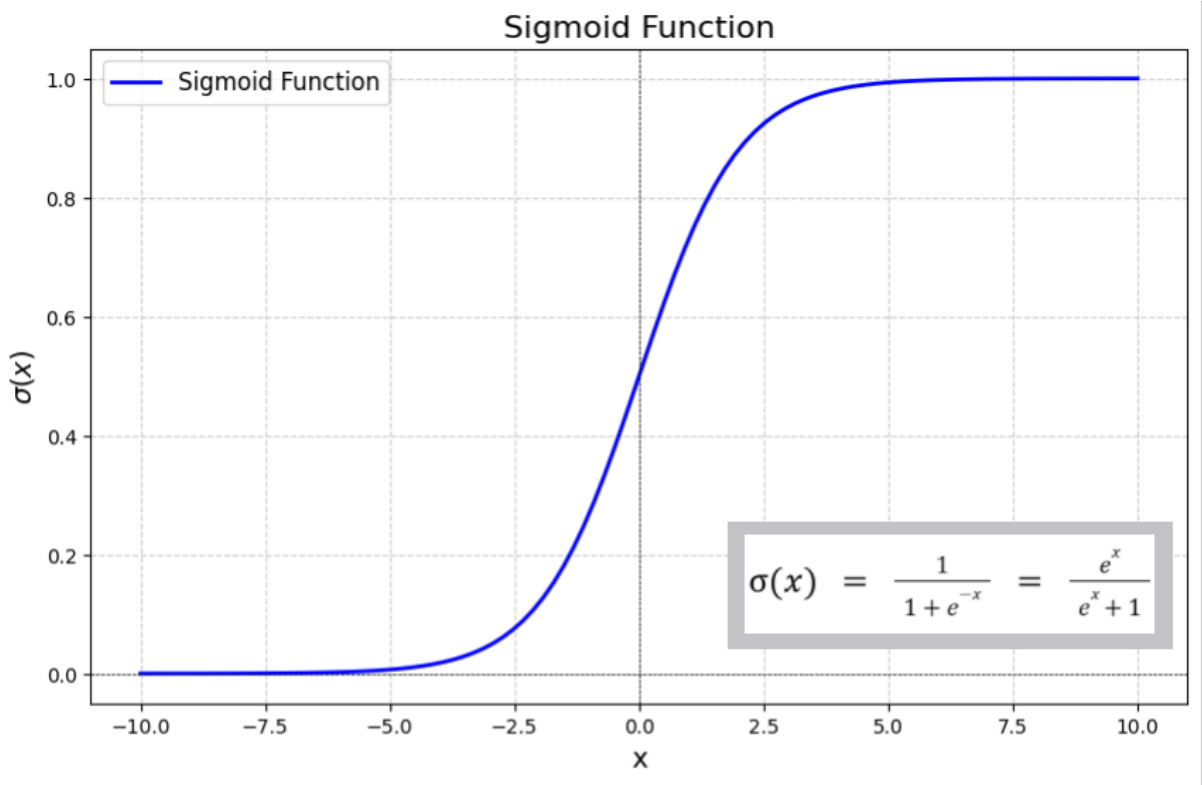
Дијагностика Алцхајмерове болести једна је од задатака **бинарне класификације**, који подразумева предвиђање којој од две категорије тј. групе припада одређени улазни пример. У овом случају се захтева раздвајање пацијената на две групе: пацијенте са Алцхајмеровом болешћу коју ћемо обележавати са АД и пацијенте који немају Алцхајмерову болест коју ћемо обележавати са не-АД. За представљање пацијената ћемо користити карактеристике њиховог говора и текста описане у претходној секцији, док ћемо за саму класификацију користити алгоритме логистичке регресије и стабла одлучивања. Алгоритам логистичке регресије, поред класе АД и не-АД, даје и вероватноћу са којом се пацијент налази у одређеној категорији док алгоритам стабла одлучивања даје поступак извођења закључка у форми стабла. Поред ових алгоритама, за бинарну класификацију се најчешће користе модели случајних шума, модел потпорних вектора, наивни Бајес и разни типови неуронских мрежа.

6.3 Логистичка регресија

Логистичка регресија је алгоритам надгледаног машинског учења помоћу којег се подаци класификују у категорије тако што се предвиђа вероватноћа да одређени податак припадне некој класи на основу неких његових карактеристика. Овај алгоритам се користи најчешће у задатку бинарне класификације.

Општи облик функције логистичке регресије је:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-z}} \tag{1}$$



Слика 4: Сигмоидна функција

Овде је $X = (x_1, x_2, ..., x_k)$ скуп карактеристика улаза за које сматрамо да утичу на вероватноћу, а $z = \omega_1 * x_1 + \omega_2 * x_2 + ... + \omega_n * x_n + b$. Величина z је линеарни предиктор и у њој фигуришу слободни члан b и тежине $\omega_1, \omega_2, ..., \omega_n$ придружене вредностима $x_1, x_2, ..., x_n$.

Сигмоидном функцијом тј. функцијом $\frac{1}{1+e^{-z}}$ се вредности линеарног предиктора трансформишу у вредности у распону $(0, 1)$ које представљају вероватноће припадности класи. Дакле, ако функција узима вредности веће од неке границе коју поставимо, онда улазни пример припада тој класи (позитивној класи), док у супротном не припада класи (припада негативној класи). Ова граница се зове **праг класификације (енг. classification treshold)** или **граница одлучивања (енг. decision boundary)** и бирамо је у зависности од тога колико нам је важно да је модел што ”сигурнији”.

За процену рада модела, у оквиру које поредимо вредности које израчуна модел са правим вредностима у скупу података, користимо **функцију губитка (енг. log-loss)**. Она има следећи облик:

$$\sum_{i=0}^{n} -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i))$$

где су y_i стварне класе, а p_i предикције модела.

Циљ обучавања модела је да минимизујемо ову функцију да би грешка предвиђања била што мања. Можемо да уочимо да се ова функција приближава нули што су вредности предикције ближе тачној вредности, а бесконачности када су најдаље од тачне вредности. Вредности тежина $\omega_1, \omega_2, ..., \omega_n$ и слободног члана b израчунавамо градијентним спустом

(енг. gradient descent) или користећи принцип максималне веродостојности (енг. maximum likelihood estimation). Принцип максималне веродостојности максимизује функцију log--likelihood која је задата формулом:

$$\sum_{i=0}^{n} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

Ова функција се диференцира по траженим параметрима, диференцијали се изједначавају са 0 и решава се добијени систем једначина.

Градијентни спуст је метода оптимизације којом постепено мењамо вредности тежина $\omega_1, \omega_2, ..., \omega_n$ и вредност слободног члана b да би се минимизовала грешка тј. log--loss. Након сваког корака поново рачунамо вредност грешке, па померамо вредности $\omega_1, \omega_2, ..., \omega_n$, b у смеру у коме ће се log-loss смањити. Дакле:

$$w' := w - \eta \frac{\partial f}{\partial w}$$

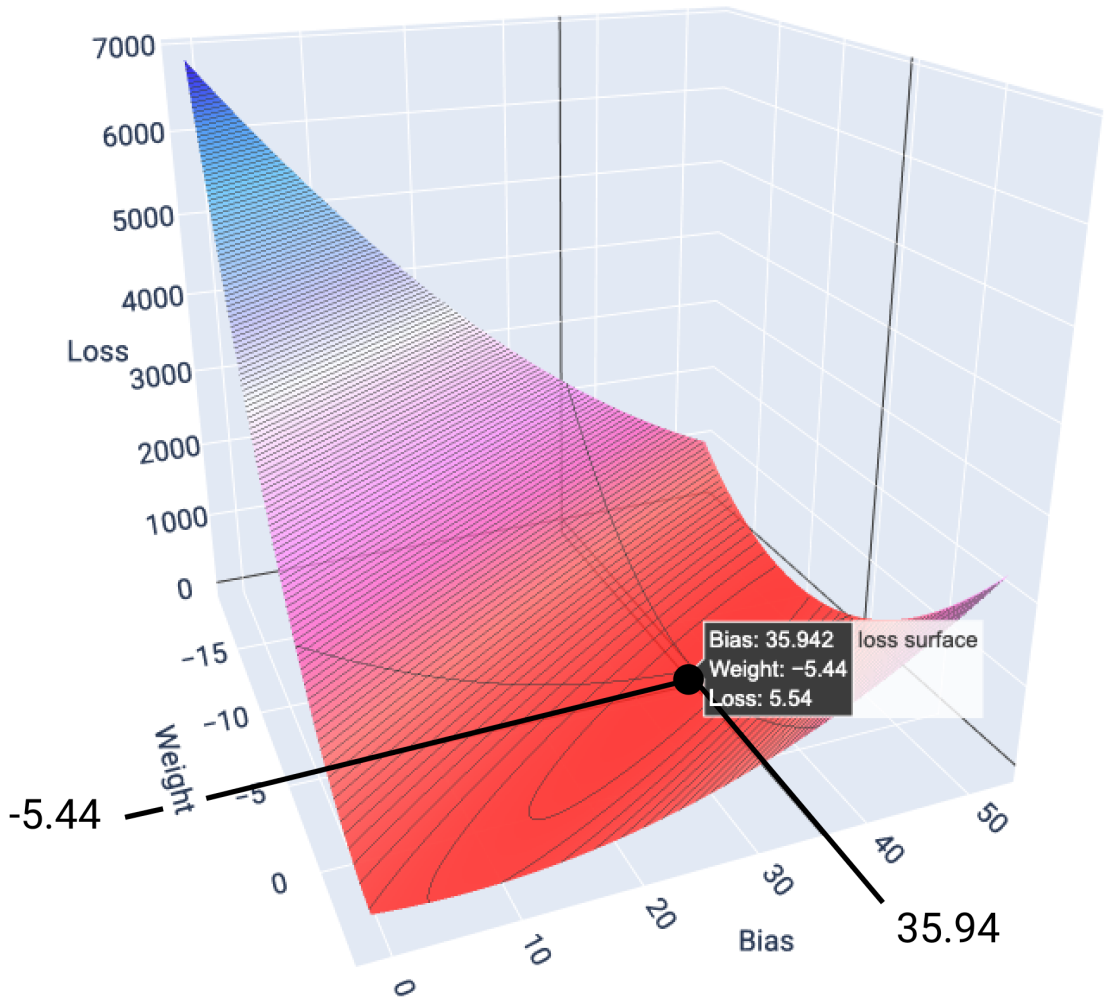
$$b' := b - \eta \frac{\partial f}{\partial b}$$

где је η корак који бирамо пре примењивања методе. Процес понављамо док вредности не конвергирају. Број итерација који је потребан да дође до конвергенције се одређује **кривом губитка**. Пошто је ова функција конвексна, вредности слободног члана и тежина ће бити близу вредности минимума функције. Важно је да изаберемо параметар η тако да вредности конвергирају у коначном броја итерација (не премало η), а да вредности не осцилују (не превелико η).

Регуларизацијом спречавамо да се модел преприлагоди (енг. overfitting) скупу за тренирање. За логистичку регресију користимо *L2* регуларизацију или *рано заустављање* (енг. early stopping). **L2 регуларизација** "кажњава" велике тежине тако што на функцију грешке (log loss) додаје $\lambda \sum \omega$, где је λ стопа регуларизације коју подешавамо у зависности од података и стопе учења. Тако сада гледамо да минимизујемо ову нову функцију. **Рано заустављање** је алтернатива L2 регуларизацији која подразумева заустављање тренирања пре конвергенције (на пример, када функција губитка креће да расте).

6.4 Стабла одлучивања

Стабла одлучивања (енг. decision trees) су алгоритми надгледаног машинског учења који се користе у проблемима регресије и класификације. Састоје се од **кореног чвора** (чвор од кога почињемо итерације), **чворова одлучивања** (чворови који се деле у две гране, тј. одлуке) и **листова** (коначног одговора алгоритма). Алгоритам израчунава критеријум по коме дели податке у сваком кораку, па тако дели цео скуп података на дисјунктне регионе. Чворови одлучивања постављају одређена питања или услове тј. раздвајања по којима деле скуп података у неком кораку. Ови услови могу бити услови који се односе на



Слика 5: Функција губитка је конвексна крива, па можемо наћи њен минимум

једну карактеристику или услови који укључују више карактеристика. Делимо их, такође, на бинарне (деле скуп података на два дела) и не-бинарне (подела скупа података на више делова), али и на услове који зависе од прага тј. границе одлучивања и који су облика *карактеристика>=граница* или *карактеристика=вредност* и *in-set* где карактеристика припада некој колекцији.

Стабла одлучивања се тренирају похлепним подели па владај (енг. greedy divide and conquer) алгоритмом. За сваки чвор се рачуна карактеристика која даје највећу вредност добитка информације (енг. information gain) или Гини индекса (енг. Gini index). Ако таквих карактеристика више нема, онда добијамо лист. Овај алгоритам се зове **ID3 алгоритам**. ID3 креће од корена чвора и у сваком наредном кораку рачуна ентропију за сваку карактеристику. Потом се за ту карактеристику испробавају различите вредности **splitter-а** који дели скуп података на оне који дају максималну **вредност добитка информација**, задате

функцијом:

$$\Delta IG = H_{roditelj} - \frac{1}{N}\sum_{deca} N_{dete} * H_{dete}$$

$$H = \sum_{i=1}^{n} p_i log_2(p_i)$$

где H представљају вредности ентропије у чворовима, p_i вероватноће дешавања догађаја, а N пропорције вредности које припадну одређеној класи. Ова функција представља разлику у ентропији пре и после поделе скупа на два дела. Како ћемо тај скуп поделити, тј. која је вредност splitter-a, рачуна се итеративно као средња вредност свака два суседна елемента скупа (пошто их има коначно), што даје временску сложеност од $O(mnlog^2 n)$, где је m број карактеристика, а број n примера у скупи за тренирање. ID3 алгоритам се завршава када чвор не може више да се подели и тада он постаје лист са ознаком класе којој припада већина вредности које испуњава услове тог чвора. **Ентропија H (Shannon's entropy)** је мера нечистоће, тј. мери количину информација неког догађаја. На пример, H је максимална (најнесигурнија) када су вредности p_i једнаке, а када $H = 0$, тј. сви p_i су 0 осим једног који је једнак 1, онда је догађај потпуно очекиван и ентропија је минимална. У обучавању модела циљ нам је да минимизујемо ентропију тј. максимизујемо вредност добитка информација. Уместо функције добитка информација користи се и **Гини индекс**, функција која има облик:

$$I_G(p) = \sum_{i=1}^{J}(p_i \sum_{k\neq i} p_k) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} p_i^2$$

Од важности нам је и које карактеристике ћемо користити у стаблу одлучивања па ћемо у ту сврху рачунати одређене метрике које нам говоре колико је нека карактеристика битна при класификацији. Сума *split-score* је мера вредности побољшања рада модела након поделе чвора који укључује карактеристику на гране. Такође, користимо број чворова у којима се карактеристика појављује, као и дубину прве појаве карактеристике у чвору.

Стабла одлучивања су лака за интерпретацију, брзо се тренирају и не треба им захтевна припрема података. Међутим, јако су осетљива на мале промене у подацима и лако постају веома комплексна ако се не користи неки алгоритам регуларизације. Овако долази до појаве познате као **преприлагођавање (енг. overfitiing)** када модел постане превише специјализован за коришћени скуп података за тренирање, па не функционише довољно добро на другим скуповима података. Овај проблем решавамо **одсецањем (енг. pruning)** грана или **ограничавањем максималне дубине** и **постављањем минимума вредности које лист мора да садржи**. Хиперпараметре ових процеса (нпр. максимална дубина) добијамо унакрсном валидацијом, тј. тренирањем модела на p различитих група скупова података са различитим параметром и бирања онога за којег проценимо да најбоље ради. Други проблем стабла одлучивања, проблем осетљивости на мале промене, решавамо са **шумама одлучивања**, скупом више независних стабала одлучивања који се тренирају на

варијацијама скупа података. **Случајне шуме (енг. Random Forests, RF)** је модел који ради на принципу усредњавања вредности предикција више стабла одлучивања којима додељујемо случајни шум. Свако стабло одлучивања се тренира на насумичном подскупу скупа за тренирање исте величине. Случајне шуме нису подложне преприлагођавању те не захтевају одсецање.

6.5 Евалуација

Евалуација модела је фаза развоја модела која указује колико успешно модел обавља свој задатак. Евалуација се увек врши на скупу за тестирање и у контексту класификације указује колико су вредности предвиђене моделом близу правим, очекиваним вредностима. У зависности од потреба модела и типа података који се анализирају могу се рачунати и интерпретирати различите мере.

За задатке бинарне класификације најчешће се користе мере као што су тачност, прецизност, одзив и F1 мера. Да би могле да се израчунају, упоређују се очекиване класе за податке из скупа за тестирање и класе које модел који се оцењује предвиђа. Обично се једна класа назива **позитивном** а друга **негативном** (на пример, класа пацијената са Алцхајмеровом болешћу је позитивна а класа пацијената без Алцхајмерове болести негативна). Вредности ових поређења се чувају у такозваној матрици конфузије (енг. confusion matrix), матрици са две врсте и две колоне и пољима:

- **тачно позитивно (енг. true positive, TP)**: број примера који припадају позитивној класи које је модел класификовао као позитивне,
- **тачно негативно (енг. true negative, TN)**: број примера који припадају негативној класи које је модел класификовао као негативне,
- **лажно позитивно (енг. false positive, FP)**: број примера који припадају негативној класи које је модел класификовао као позитивне,
- **лажно негативно (енг. false negative, FN)**: број примера који припадају позитивној класи које је модел класификовао као негативне.

Мере класификације рачунамо преко вредности TP, FP, TN и FN које очитавамо из матрице конфузије и користимо их у складу са особинама скупа података, типа задатка класификације и типа модела. У медицинским скуповима података важно је осврнути се на својство небалансираности. **Небалансиран скуп података (енг. inbalance dataset)** има мало примера позитивне класе у односу на негативну класу, најчешће због природе феномена који се моделује. На пример, мање је особа које имају Алцхајмерову болест него оних које је немају. На балансираности скупа података се може инсистирати и у терминима неког атрибута који описује податке, уколико је то могуће. На пример, може се инсистирати на подједнаком броју пацијената мушког и женског пола или на подједнаком скупу пацијената по старосним групама.

Тачност (енг. accuracy) је мера пропорције тачно класификованих инстанци и рачуна се према формули која је наведена доле.

Слика 6: Матрица конфузије: негативна класа је обично представљена лабелом 0 а позитивна лабелом 1

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Ова мера је корисна када је скуп података балансиран. У случају као што је наш тј. када позитивних инстанци има много мање него негативних, тачност не даје реалну слику о квалитету модела.

Одзив (енг. recall) је мера пропорције тачно класификованих позитивних инстанци према свим заиста позитивним инстанцама.

$$Recall = \frac{TP}{TP + FN}$$

Одзив нам је врло користан у детекцији деменције јер је много опасније класификовати позитивну инстанцу (АД) као негативну (нон-АД) него класификовати негативну (нон-АД) као позитивну (АД).

Удео лажно негативних (енг. false negative rate, FPR) је пропорција стварно негативних инстанци класификованих као позитивне према свим заиста негативним инстанцама.

$$FPR = \frac{FP}{FP + TN}$$

Ову меру користимо када је важније да тачно класификујемо негативне инстанце него позитивне.

Прецизност (енг. precision) је пропорција свих тачно класификованих вредности позитивне класе у односу на све вредности које је модел класификовао као позитивне.

$$Precision = \frac{TP}{TP + FP}$$

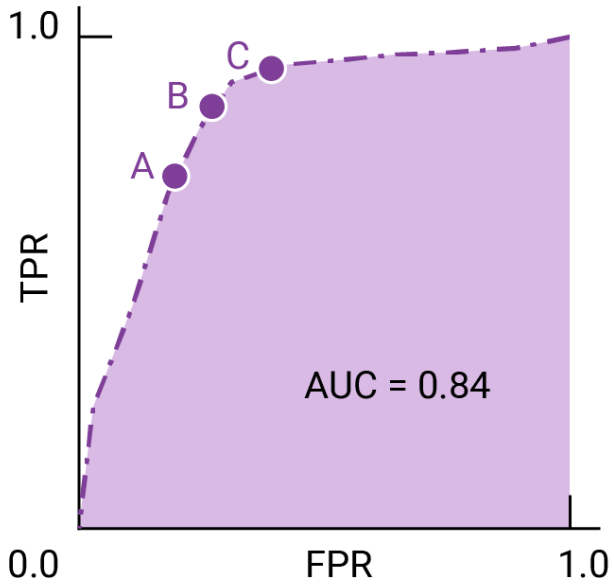
Прецизност користимо када нам је важно да су нам тачне позитивне предикције.

F1 мера (енг. F1 measure) је хармонијска средина одзива и прецизности.

$$F1 = \frac{2TP}{2TP + FP + FN}$$

F1 мера се најчешће користи за оцену класификације модела.

ROC крива (енг. reciever-operating characteristc curve, ROC) је визуелна репрезентација тачности модела. Црта се као одзив у функцији од удела лажно негативних. Површина испод ове криве, такозвани **AUC (енг. area under the curve)**, представља вероватноћу да модел додели позитивној инстанци већу вредност од оне из негативне класе. Ова крива нам је, такође, значајна и за бирање вредности **прагова класификације (енг. threshold)**, који ће бити близу најближе вредности тачки (0,1) која се налази на криви. Ову вредност можемо подесити да буде мало испод или изнад те тачке криве.



Слика 7: ROC крива и површина испод ње (AUC)

7 Класификација пацијената оболелих од Алцхајмерове болести

7.1 Скуп података

Скупови података коришћени приликом детекције деменције се састоје од информација о пацијентима, прикупљених углавном у медицинским установама или центрима за истраживање, на основу којих модел машинског учења може да разликује здраву особу од оболеле особе. Овакви подаци могу укључивати резултате тестирања, биомаркере, податке о старости, полу, навикама и разне друге а потом, када лекари одраде дијагностику, и ознаке (енг. labels) придружене тим подацима.

Скуп података са којим смо радили се зове ***ADReSS*** **(акроним за Alzheimer's Dementia Recognition through Spontaneous Speech)** и саставни је део колекције мултимедијалних скупова података ***DementiaBank***[1] припремљених за проучавање корелације између говорне комуникације и деменције. Скуп података ADReSS је стандардизовани скуп података који се састоји од снимака спонтаног говора и транскрипата снимака описа слике *Cookie Theft* на енглеском језику. Прикупљени су подаци 156 испитаника (78 по свакој групи). Снимци су сегментовани, акустично обрађени и статистички уравнотежени по питању пола и старости, што омогућава истраживачима да пореде различите методе дијагностике које користе говор као индикатор болести, побољшавају постојеће моделе и развијају нове. Такође, ADReSS скуп података је *speaker-independent*, тј. није трениран на истој особи на којој се и тестира, што доприноси већој распрострањености коришћења овог скупа података.

Пример транскрипта АД групе:

well the poor mother's a-doin dishes . there's a boy on a stool . cookie jar . and a girl down below . is that all you wanted to know ? okay . there's a cookie jar . the little boy is standing on a stool and he's ready to go over . the little girl's standing there . the mother's at the sink doing dishes . her water's overflowing . that's about it .

Пример транскрипта не-АД групе:

a boy is taking cookies from the cookie jar giving one to his sister . he's also falling off the stool he is on . the mother is washing dishes . she's letting the sink overflow . it's getting all over the floor . I see nothing further . the girl is saying be quiet if that counts . I don't know what the mother gets by standing in all the water . I don't think that's very important . that's all I see . what have I missed

[1] https://talkbank.org/dementia/

	Train-AD		Test-AD		Train-Non-AD		Test-Non-AD	
Age	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
[50, 55)	1	0	1	0	1	0	1	0
[55, 60)	5	4	2	2	5	4	2	2
[60, 65)	3	6	1	3	3	6	1	3
[65, 70)	6	10	3	4	6	10	3	4
[70, 75)	6	8	3	3	6	8	3	3
[75, 80)	3	2	1	1	3	2	1	1
Total	24	30	11	13	24	30	11	13

Слика 8: ADReSS скуп података, расподела по полу и старости

ADReSS скуп података се углавном користи за две намене: бинарну класификацију пацијената на основу говора и добијање регресионог модела који предвиђа њихове резултате на MMSE тесту. У овом раду ћемо се фокусирати на класификацију пацијената и то по класама АД и не-АД, где АД означава класу пацијената оболелих од Алцхајмерове болести, а не-АД класу пацијената који нису оболели.

База података о пацијентима садржи јединствени идентификатор особе и колоне са карактеристикама као што су старост, пол, резултат на MMSE тесту и ознаке (АД, не-АД). Сваки унос је повезан са аудио снимком нарације и текстуалним транскриптом. Због тога се за класификацију пацијената могу користити модели машинског учења који раде са звуком или сликама (нпр. спектрограмима звука) или са текстуалним подацима. Програми попут openSMILE, ComParE, eGeMAPs i MRCG се конфигуришу и користе тако да могу да издвоје различите карактеристике говора из овог скупа (трајање и брзина говора, број пауза, итд.) и класификују особу по томе да ли има Алцхајмерову болест или не. ADReSS је значајан и по томе што је један од ретких непристрасних скупова података који не користе податке од једне особе више пута. Постигао је тачност класификације од 62.5% , док је за ручно унете транскрипте досегнуо тачност од 76.85%.

Скупови података коришћени за детекцију деменције су и даље оскудни по питању медицинских података и прилично хетерогени али се, уз развитак комплекснијих модела, пре-процесовање података и смањење коришћених карактеристика говора, може очекивати побољшање тачности оваквих модела.

7.2 Окружење KNIME

KNIME [2]**(Konstanz Information Miner)** је алат за анализу, интеграцију и обраду разних врста података који користи једноставно визуелно програмирање и компоненте машинског учења за обављање аналитичких задатака. Састоји се од интерактивног и интуитивног корисничког интерфејса и система Java Database Connectivity. Овај програм нам омогућава да графички представимо модел обраде података помоћу чворова (енг. nodes) који означавају одређену трансформацију над подацима и који се спајају превлачењем и спуштањем (енг. drag-and-drop) у токове података (енг. data pipelines). KNIME је реализовао тим софтверских инжењера на Универзитету Констанз у Немачкој, а прва верзија је избачена 2006. године.

Као што смо навели, KNIME се састоји из мноштва чворова, тј. јединица за обраду и трансформацију података, спојених у токове података сличних графовима, који повезују ове чворове. Сваки чвор је приказан као квадрат са именом, анотацијом (описом чвора), статусом, улазним и излазним портом. Улазни портови чвору дају чвору податке које треба да обради, док излазни избацују резултат чвора. Можемо да спојимо само портове исте врсте, тј. исте боје и облика. Постоје разне врсте портова: data table, database connection, flow variable, tree ensemble model, report и image портови. Статус чвора је приказан ”семафором” испод чвора. По статусу, сваки чвор може бити конфигурисан, неконфигурисан, извршен или са грешком. Такође, сваки чвор има и траку радњи која приказује операције које чвор може да изврши. Трака радњи се користи за конфигурацију, извршавање, обустављање и ресетовање рада чвора.

Чворови обављају разне функције као што су учитавање велике количине података, припрема података и издвајање њихових карактеризација тј. атрибута. Посебно се издвајају чворови за улазно-излазне операције над подацима (ARFF и Hitlist читачи фајлова, конектори база података, CSV читачи фајлова, Hitlist и ARFF уписивачи), манипулације над подацима (филтери, раздвајање података, узорковање података, насумично мешање или сортирање података, спајање и комбиновање података), трансформације података (замена недостајућих вредности, генератори номиналних вредности), учење из података (кластеринг, стабла одлучивања, регресија, неуронске мреже), статистички чворови (средња вредност, стандардна девијација) и чворови визуелизација (хистограми, питице, тачкасти графикони).

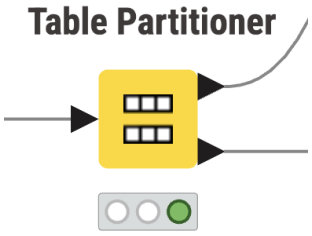
Подаци са којима смо радили се налазе у CSV формату. Чвор **CSV Reader** је предвиђен за читање CSV фајлова. Кроз посебне панеле можемо подесити од ког реда читање креће, како дефинишемо почетак линије или колоне а како вредности.

Чвор **Table Partitioner** дели табелу на два дела по редовима, углавном једну за тре-

[2]https://www.knime.com

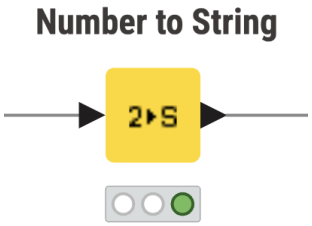
Слика 9: Чвор CSVReader

нирање модела, а другу за тестирање. Партицију можемо дефинисати по пропорцији редова или броју редова који су у првој табели. Подразумева се да остатак редова припада другој табели. Такође, партиција може бити насумична (насумично бира редове за прву табелу) за коју можемо изабрати насумично семе (енг. random seed) тј. почетну вредност која насумичном генератору задаје случајни низ бројева, балансирана (чува дистрибуцију вредности у изабраној колони), линеарна (бира једнако удаљене редове) или партиција методом првих редова (узима узастопне редове почевши од првог реда).



Слика 10: Table Partitioner

Очекује се да у скупу података, циљна променљива тј. колона која садржи вредности класа буде текстуални податак. За претварање нумеричких вредности у текст (стринг) користи се чвор **Number to String**. Да би се остварила стабилност класификације и унапре-



Слика 11: Number to String

дила интерпретабилност резултата користили смо чворове за нормализацију. Чвор **Normalizer** нормализује нумеричке вредности у задатим колонама. Нормализација је стављање бројних вредности на исту скалу тако да нам је лакше да их поредимо. У чвору постоје три врсте нормализације: min-max нормализација, Z-score нормализација и децимално

скалирање. У min-max нормализацији, минимална вредност постаје 0, а највећа 1, док остале рачунамо по формули:

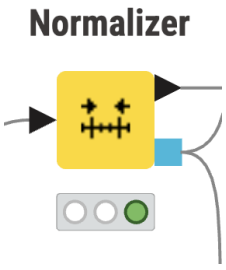
$$value' = \frac{value - value_{min}}{value_{max} - value_{min}}$$

Z-score нормализација подразумева постављање аритметичке средине на 0, док остале вредности рачунамо као

$$value' = \frac{value - average}{\sigma}$$

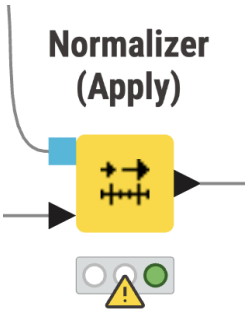
, где је $average$ аритметичка средина $\frac{\sum_{i=1}^{n} x_i}{n}$, а σ стандардна девијација. $\sqrt{\frac{\sum_{i=1}^{n}(x_i - average)^2}{n}}$. Децимално скалирање подразумева рачунање степена j за који је $\frac{value_{max}}{10^j} < 1$ па рачунање осталих вредности по формули:

$$value' = \frac{value}{10^j}$$



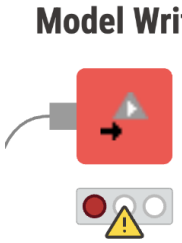
Слика 12: Normalizer

Normalizer(Apply) је често повезан са чвором **Normalizer**. Он нормализује дате нумеричке вредности по истим, већ израчунатим, параметрима нормализације чвора **Normalizer**.



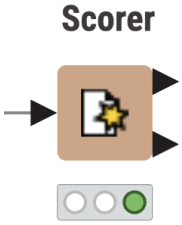
Слика 13: Normalizer(Apply)

Чвор **ModelWriter** преписује модел у фајл који може да се сачува, и касније прочита помоћу чвора **Model Reader**.



Слика 14: Model Writer

Чвор **Scorer** евалуира модел и приказује метрике тачности и матрицу конфузије. Он пореди вредности колоне предикције са тачним вредностима колоне валидационог скупа и затим их обележава као TP, FP, TN, FN. Метрике које рачуна су одзив, тачност, прецизност, сензитивност, специфичност, F1 мера и друге.

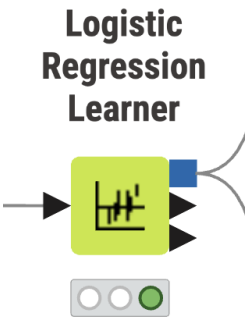


Слика 15: Scorer

Навешћемо сада чворове који су искоришћени у моделима логистичке регресије и стабла одлучивања које користимо за задатак класификације.

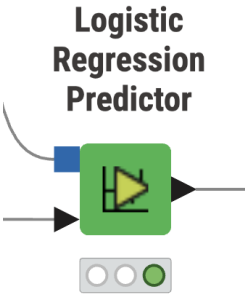
Чвор **Logistic Regression Learner** тренира модел логистичке регресије. За рачунање параметара логистичке регресије може се одабрати један од два оптимизациона алгоритма: метод најмањих квадрата (енг. iteratively reweighted least squares) или метод стохастичког градијентног спуста (енг. stohastic average gradient, SAG). Метод најмањих квадрата користи Фишерово оцењивање $\omega' = \omega + I(\omega)^{-1} * U(\omega)$ (где је I Фишерова информација, U функција грешке, слична функцији градијента, а ω су тежине) и користи се за мање скупове података. Стохастички градијентни спуст користи варијанту градијентног спуста који је погодан за веће скупове података. Корак учења може бити фиксиран (унапред одабран) или одабран алгоритмом линијске претраге којим се проналази оптимална вредност корака. Такође, стохастички градијентни спуст може и не мора да користити регуларизацију - технику која спречава да се модел превише прилагоди (енг. overfitting) тренинг подацима. Може се одабрати *Gauss prior* регуларизација (слична L2 регуларизацији) која на функцију губитка додаје $\lambda \sum \omega^2$ и тако спречава коришћење превеликих тежина. Друга регуларизација је *Laplace prior* регуларизација која на функцију губитка додаје $\lambda \sum |\omega|$ и тако задржава само најважније карактеристике.

Чвор **Logistic Regression Predictor** израчунава (предвиђа) резултате за податке из ску-



Слика 16: Logistic Regression Learner

па тестирања користећи модел логистичке регресије (чвор Logistic Regression Learner) са којим је повезан.



Слика 17: Logistic Regression Predictor

Чвор **Decision Tree Learner** тренира модел стабла одлучивања. Алгоритам приликом сваке поделе може рачунати Гини индекс (енг. Gini index) или количник добитка (енг. gain ratio) по формули $\frac{InformationGain}{SplitInformation}$ где је $SplitInformation = -\sum p_i log_2(p_i)$. За регуларизацију овог алгоритма се може користити техника одсецања (енг. pruning) и/или ограничење на минимални број примера по листу.

Слика 18: Decision Tree Learner

Decision Tree Predictor израчунава (предвиђа) резултате за податке из скупа за тестирање користећи модел стабла одлучивања са којим је повезан (чвор Decision Tree Predictor).

Decision Tree View је интерактивно стабло одлучивања приказано графички преко библиотеке JavaScript. Поред чворова, приказани су и назив и број чланова класе.

Слика 19: Decision Tree Predictor

Слика 20: Decision Tree View

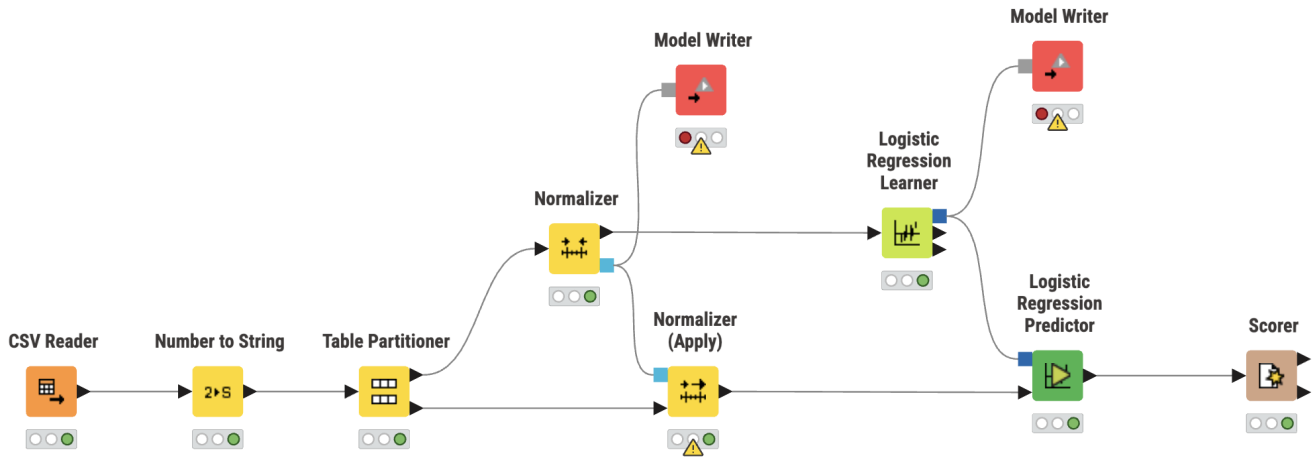
7.3 Класификација наратива

Коришћене карактеристике укључују 34 лингвистичке карактеристике програма CLAN (енг. Computerized Language ANalysis) са универзитета Carnegie Melon и 88 акустичких карактеристика скупа eGeMAPS. Лингивстичке карактеристике које су мерене су: трајање узорка у секундама, број речи у минуту, однос речи отворене (именице, глаголи, придеви, прилози) и затворене класе (заменице, предлози, везници, помоћни глаголи), број речи отворене и затворене класе, број речи, исказа и средња дужина исказа и однос типа и токена (тј. однос уникатних речи према укупном броју речи). Акустичке карактеристике које су посматране укључују: F0 semitone (најнижа фреквенција говора), јачина звука, подрхтавање говора, спектрални флукс (тј. стабилност тона), и Mel-frequency cepstral coefficients (MFCC). Од алгоритама машинског учења користили смо логистичку регресију и стабла одлучивања.

7.3.1 Логистичка регресија

Модел логистичке регресије је имплементиран у окружењу KNIME и представљен је дијаграмом који се налази у наставку. Дијаграм почиње чвором *CSV Reader* у који учитавамо нашу табелу са подацима у форми CSV фајла. Затим користимо чвор *Number to String* да претворимо вредности ознаке АД и не-АД у стринг, да би модел лакше препознао категорије и да им не би доделио ”тежину”. Чвор *Table Partitioner* дели податке у групу за тренирање, коју потом нормализујемо *Normalizer* чвором, и другу за тестирање која се нормализује применом чвора *Normalizer (Apply)*. Однос скупа података је 80%:20% док је за нормализацију коришћен min -max алгоритам. Модел тренирамо на *Logistic Regression Learner* чвору, па тестирамо *Logistic Regression Predictor* чвором који израчунава да ли особа има деменцију или не. На крају евалуирамо модел чвором **Scorer** који рачуна

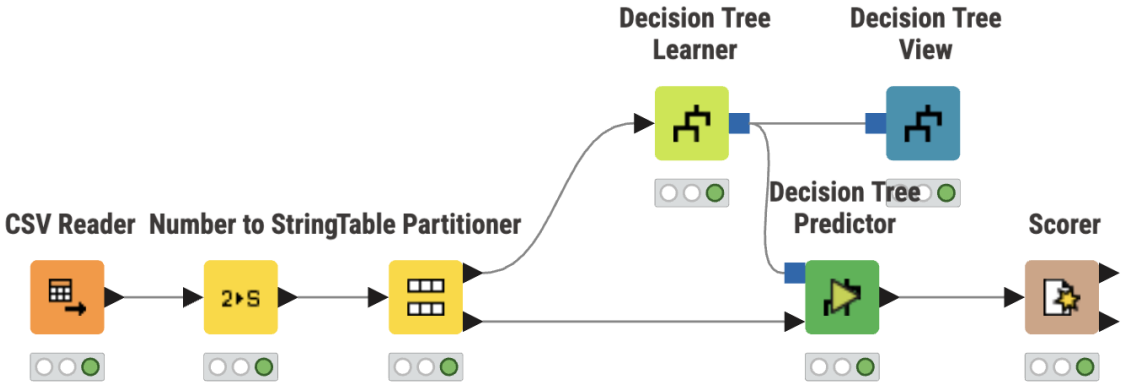
матрицу конфузије и опционо га чувамо помоћу чвора **Model Writer**.



Слика 21: Дијаграм модела логистичке регресије

7.3.2 Стабла одлучивања

Слично као и горњем примеру, податке учитавамо у *CSV Reader* чвору и преводимо вредности лабела у стринг. Податке делимо у групу *Table Partitioner* чвором за тренирање који тренирамо помоћу *Decision Tree Learner* чвора и тестирање који избацује коначне предикције након коришћења *Decision Tree Predictor* чвора. Чвор *Scorer* евалуира наш модел, а чвор *Decision Tree View* нам даје графички приказ нашег стабла одлучивања.



Слика 22: Дијаграм модела стабла одлучивања

7.3.3 Резултати и анализа резултата

Будући на велики број атрибута који се односе на акустичка и лингвистичка својства скупа података, као и број хиперпараметара за подешавање алгоритама, постојао је већи

број експеримената. У свим поменутим експериментима је као мера евалуације коришћена тачност јер је скуп података балансиран. Такође, резултати постојећих експеримената су изражени у терминима тачности па је ово омогућавало финије упоређивање са резултатима заједнице.

Прва група експеримената се односила на коришћење искључиво акустичких својстава. У њима су се најбоље показала својства као што су енергија, MFCC и дескриптори самог гласа као што су логаритам количника хармоније и шума (енг. harmonic-to-noise ratio), параметри квалитета звука (енг. voice quality features) и нека својства јасноће звука (енг. psychoacoustic spectral sharpness). Саме перформансе класификатора су се, у зависности од подешавања хиперпараметара, кретала у распону тачности од 0.6 до 0.7.

Друга група експеримената се односила на коришћење искључиво лингвистичких својстава. Тачност ових експеримената је била већа у односу на акустичка својства и износила је у распону од 0.7 до 0.75 при чему су резултати добијени алгоритмом стабла одлучивања били у просеку већи. Међу важнијим својствима су се истакли: број речи по реченици, број речи у јединици времена и однос броја различитих речи и укупног броја речи у тексту.

Трећа група експеримената је комбиновала акустичка и лингвистичка својства и резултирала је за обе класе алгоритама најбољим тачностима - од 0.7 до 0.8.

Најновији резултати заједнице који користе овај скуп података а који су базирани на неуронским мрежама остварују тачност око 0.9.

8 Закључак

Како Алцхајмерова болест представља све већи изазов у здравству широм света, то је све већи значај проналажења приступачних техника ране дијагностике. По постојећим резултатима рудиментарних модела можемо закључити да су говор и језик доста прецизни, ефикасни, неинвазивни, јефтини и приступачни биомаркери раних когнитивних поремећаја деменције. Међутим, иако доста препознат у научној сфери вештачке интелигенције и обраде природног језика, овакав приступ дијагностици још увек није присутан у широј клиничкој употреби. Примена ове методе у клиничкој пракси захтева превазилажење неколико ограничења која отежавају примену у већим размерама, као што су очување приватности података, необјашњивост модела (лекар не може да зна зашто је модел дао одређену дијагнозу) и коришћење униформног скупа података (ADReSS) у свим студијама и језицима, да би оне могле лакше да се пореде и понове. Са развојем истраживања и проширењем скупова података на енглеском, а посебно у другим језицима, се очекује да ће доћи до усавршавања вишејезичних модела, развоја корисничких интерактивних алата за дијагностику (нпр. кроз паметне телефоне, паметне куће и паметне градове) и примене мултимодалних приступа (комбинације говора са другим биомаркерима). На пример, данашњи паметни телефони поред говора (тј. дијалога и друштвених интеракција или интеракција са chatbot-ом или AI асистентом као што је Alexa) могу да прате и кретање, сан и брзину куцања срца, које додатно могу указати на поремећаје у свакодневном функционисању. Федеративно учење моделе учи на подацима унутар болнице, без размене приватних података, већ само дељењем параметара, што може да допринесе данашњем проблему приватности и етичности приступа. Проблем необјашњивости ("black box") модела вештачке интелигенције може се превазићи развојем XAI (акроним из explainable AI), који би могао да објасни зашто је предвидео одређену дијагнозу. Такође, вештачка интелигенција омогућава да преко преносивих или кућних уређаја пратимо симптоме особе и тако откривамо мале, персонализоване промене током времена, уместо једнократног тестирања које би могло да те промене занемари, што може бити веома значајно за прогресивне болести попут деменције.

Иако се тренутно користи само као подршка дијагностике Алцхајмерове болести у клиничкој пракси, коришћење AI у овој области показује велики потенцијал за доношење тачних дијагноза које могу бити свакоме приступне, мање инвазивне за пацијенте и јефтиније него друге методе дијагностике.

9 Референце

[1] WHO – Dementia facts sheet https://www.who.int/news-room/fact-sheets/detail/dementiahttps://www.who.int/news-room/fact-sheets/detail/dementia

[2] Alzheimer's Association – What is Alzheimer's? https://www.alz.org/alzheimers-dementia/what-is-alzheimershttps://www.alz.org/alzheimers-dementia/what-is-alzheimers

[3] ScienceDirect – Applications of artificial intelligence to aid early detection of dementia: A scoping review on current capabilities and future directions https://www.sciencedirect.com/science/article/pii/S1532046422000466?via%3Dihubhttps://www.sciencedirect.com/science/article/pii/S1532046422000466?via%3Dihub

[4] Cambridge Core – AI in dementia research https://www.cambridge.org/core/journals/cambridge-prisms-precision-medicine/article/applications-of-artificial-intelligence-in-dementia-research/A62C6F50BFD1E3158578B0D7580E403Ahttps://www.cambridge.org/core/journals/cambridge-prisms-precision-medicine/article/applications-of-artificial-intelligence-in-dementia-research/A62C6F50BFD1E3158578B0D7580E403A

[5] Google Developers – Classification Thresholding https://developers.google.com/machine-learning/crash-course/classification/thresholdinghttps://developers.google.com/machine-learning/crash-course/classification/thresholding

[6] Google Developers – Accuracy, Precision, Recall https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recallhttps://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall **[7] DementiaBank** https://talkbank.org/dementia/https://talkbank.org/dementia/

[8] ADReSS Challenge 2020 https://talkbank.org/dementia/ADReSS-2020/index.htmlhttps://talkbank.org/dementia/ADReSS-2020/index.html

[9] Luz et al. – Alzheimer's Dementia Recognition through Spontaneous Speech https://talkbank.org/dementia/ADReSS-2020/Luz$_e tal.pdf https://talkbank.org/dementia/ADReSS-2020/Luz_e$

[10] Machine Learning University – Logistic Regression Explained https://mlu-explain.github.io/logistic-regression/https://mlu-explain.github.io/logistic-regression/

[11] Google Developers – Logistic Regression https://developers.google.com/machine-learning/crash-course/logistic-regressionhttps://developers.google.com/machine-learning/crash-course/logistic-regression

[12] Machine Learning University – Decision Tree Explained https://mlu-explain.github.io/decision-tree/https://mlu-explain.github.io/decision-tree/

[13] Google Developers – Decision Trees https://developers.google.com/machine-learning/decision-forests/decision-treeshttps://developers.google.com/machine-learning/decision-forests/decision-trees

[14] KNIME https://www.knime.com/https://www.knime.com/

[15] Wikipedia – KNIME https://en.wikipedia.org/wiki/KNIMEhttps://en.wikipedia.org/wiki/KNIME