

TDK-dolgozat

Gárdi Réka Lorin

Budapesti Corvinus Egyetem
Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinika
ELKH-TTK Humán Pluripotens Össejt Laboratórium

Szkizofrén betegből származó idegsejtek kalcium-jeleinek elemzése Gaussian Mixture és véletlen erdő modellekkel

Gárdi Réka Lorin

Budapesti Corvinus Egyetem | Adatelemzés és Informatika Intézet | Statisztika Tanszék

Dolgozat készítésekor: Alkalmazott közgazdaságtan III. évfolyam

2023 szeptembere óta: Biztosítási és pénzügyi matematika mesterszak

Témavezetők:

Dr. Kovács László BCE: Adjunktus

Dr. Tordai Csongor SOTE: PhD-hallgató

Lezárás dátuma: 2023. 04. 16.

Absztrakt

A Semmelweis Egyetem Pszichiátriai Klinika és az ELKH-TTK Humán Pluripotens Óssejt Laboratóriumának együttműködésében a szkizofrénia biológiai alapjait vizsgáljuk. A laboratóriumban egy sejtes modellrendszer beállítása folyik, melyben a betegséget kísérő mutációk hatása vizsgálható. A modell potenciális gyógyszercélpontok jellemzését teheti lehetővé.

Kontroll és mutáns sejtvonalak összevetésével az egészséges és kóros idegsejtműködés közötti eltéréseket elemezzük. Ezen eltérések többek közt a sejt kalcium (Ca^{2+}) szintjén keresztül monitorozhatók.

A sejt kultúra Ca^{2+} szintjét először beavatkozás nélkül mértük, megkeresve a nyugalmi alapvonal szintet. Ezután ingerlő anyaggal váltottunk ki neuronokra jellemző tipikus Ca^{2+} jeleket.

Kutatási célom, hogy automatizált módszert adjak az alapvonal meghatározására. Ennek pontos megadása a sejt nyugalmi Ca^{2+} szintjének spontán emelkedései (spontán aktivitás) miatt összetett adatelemzési feladat. Ugyanis emiatt a sejt képes nyugalmi állapotban is akkora Ca^{2+} szintet mutatni, mint ingerlés után várható, noha rövidebb ideig. Tehát, az alapvonal meghatározásához ezen spontán aktív hatásokat azonosítani és eliminálni szükséges a Ca^{2+} mérésekből.

Módszertan

A mérés során keletkezett adattömeg az egyes sejtek kalcium-szint idősorai. A sejteket a mikroszkópos felvételen manuálisan jelöltük ki, mint ún. ROI-kat (Region of Interest).

Gaussian Mixture Model (GMM) segítségével adott ROI idősorában az ingerlés előtt szétválasztom az adatpontokat alpopulációkra. Az empirikus sűrűségfüggvényeken megfigyelhető, hogy a pontok túlnyomó többsége a tényleges nyugalmi Ca^{2+} szint körül helyezkedik el, a többi pont magasabb értékek körül tömörült a spontán aktív tüzelések hatására. Alapvonalnak annak az alpopulációnak a várható értékét vettem, amely legnagyobb súllyal szerepelt a teljes populációban.

Biológiai szempontból fontos adat, hogy adott mérésen belül mekkora a spontán aktív sejtek aránya. Így a labortagok segítségével tanuló adatbázisokat készítettünk. Erre egy véletlen erdő modellt hoztam létre a spontán aktív sejtek detektálására.

Eredmények

Létrehoztam egy grafikus felhasználói felületet, ahová a mérések feltölthetők. A kimenet egy táblázat, mely tartalmazza a ROI-khoz tartozó alapvonalakat és reakciókat.

A GMM-mel pontosabban megállapítható a sejtek alapvonala, nem kell külön metodikát használni a spontán aktív és nyugalomban lévő sejteken.

A véletlen erdő AUC értéke külön tesztalmanon 0,96. 105 ROI-ból összesen 8-at azonosított be tévesen a modell.

Tartalom

Absztrakt	1
Módszertan	1
Eredmények.....	2
Ábrajegyzék.....	4
Táblázatjegyzék.....	4
I. Bevezetés	5
II. Szkizofrénia	7
III. A kutatócsoport munkája.....	7
IV. Kiértékelés.....	12
IV. 1. Adatvagyonkezelés	12
IV. 2. Nyugalmi állapot és spontán aktivitás jelenségének bemutatása	13
IV. 3. Alapvonal meghatározása a kutatócsoportunkban és más hasonló tudományos munkákban .	14
IV. 3. 1. Alapvonal mint ingerlés előtti rövid időszak átlaga	16
IV. 3. 2. Alapvonal mint ingerlés előtti (rövid időszak) minimum értékek átlaga	17
IV. 3. 3. Alapvonal mint nyugalmi állapot mediánja	18
IV. 3. 4. Simítófüggvény és dekompozíciós eljárások alkalmazása az alapvonal meghatározására	18
V. Gaussian Mixture modell.....	21
VI. Véletlen erdő modell	25
VII. Eredmények	27
VII. 1. Gaussian Mixture modell eredmények	27
VII. 2. Véletlen erdő eredmények	32
VII. 3. Grafikus felhasználói felület.....	35
VII. 3. 1. GMM alkalmazás.....	35
VII. 3. 2. Véletlen erdő alkalmazás	36
VIII. Összefoglalás.....	38
Hivatkozások	40

Ábrajegyzék

1. ábra: A neuron általános sémája (Forrás: Szentágothai, Réthelyi, 2006, 155. oldal 2/53. ábra)	10
2. ábra: Neuronhálózat mikroszkópos képe (3PC1, 6 hét differenciáció után)	10
3. ábra: XCL1 sejtvonalakról fluoreszcencia mikroszkóppal készített mérések az 5. és 304. időpontban (glutamát beadásának időpontja: 299).....	11
4. ábra: ImageJ szoftverrel kiértékelt adott ROI fluoreszcencia intenzitása (XCL1 sejtvonal)	12
5. ábra: Nyugalmi állapot a spontán aktivitás két típusának jelenlétében és anélkül	14
6. ábra: Alapvonal meghatározása spontán aktivitást nem mutató ROI esetén átlag módszerrel	16
7. ábra: Alapvonal meghatározása spontán aktív ROI esetén átlag módszerrel	17
8. ábra: Alapvonal meghatározása spontán aktív ROI esetén medián módszerrel	18
9. ábra: Periodogram függvényvel megállapított periódushosszak spontán aktív ROI-k esetében.....	19
10. ábra: ROI30 (XCL1 sejtvonal) idősor dekompozíciós felbontása glutamát beadása előtt (1:299 időpontok)	20
11. ábra: H-P szűrő alkalmazása spontán aktív ROI30-on	21
12. ábra: Spontán aktív ROI30 idősora (beadás előtt) és hisztogramja	25
13. ábra: Titanic utasainak túlélés szerinti osztályozása véletlen erdő modellel (forrás: Kovács László órai példa).....	26
14. ábra: XCL1 sejtvonalból származó ROI16 és ROI56 spontán aktív sejtek idősorai és hisztogramjai nyugalmi állapotban, és az alpopulációk normális eloszlásai	28
15. ábra: XCL1 sejtvonalból származó ROI156 nem spontán aktív sejt idősora és hisztogramja nyugalmi és teljes időszakban, és az alpopuláció normális eloszlásai	29
16. ábra: Spontán aktív és spontán aktivitást nem mutató ROI-k alapvonalai 4 különböző módszerrel30	
17. ábra: GMM alkalmazás (saját fejlesztés)	35

Táblázatjegyzék

1. táblázat: Sejtvonalak elnevezése	9
2. táblázat: Alapvonal meghatározására használt módszerek.....	15
3. táblázat: Spontán aktív és spontán aktivitást nem mutató sejtek keresztvalidációval megállapított optimális normális eloszlás komponensei GMM-ben (M), keresztvalidált likelihood érték (neg_log_likelihood), válasz és sp. aktivitás.....	31
4. táblázat: Spontán aktív és spontán aktivitást nem mutató sejtek alapvonalai és reakciói különböző módszerekkel.....	31
5. táblázat: Véletlen erdő modell konfúziós mátrixai 50%-os és 29,75%-os cut-off mellett	34
6. táblázat: Véletlen erdő teljesítménymutatói 50%-os és 29,75%-os cut-off mellett. Piros = csökkenés, zöld = növekedés az eredeti mutatóhoz képest.....	34

I. Bevezetés

Dolgozatomban kísérleti adatvagyonkezeléssel foglalkoztam, továbbá egy olyan grafikus felhasználói felület fejlesztésével, amely elősegíti a kutatók munkáját kísérleti adataik kiértékelésében.

Munkám során a Humán Pluripotens Össejt Laboratóriumhoz csatlakoztam, akik a Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinikával együttműködve a szkizofrénia biológiai alapjait vizsgálják. A szkizofrénia egy súlyos mentális betegség változatos tünetekkel, nehezen diagnosztizálható és kezelhető. Ennek egyik oka, hogy a betegséggel kapcsolatos pontos biológiai folyamatok nem ismertek. A laboratóriumban egy olyan sejtes modellrendszert állítottunk be, melyben a szkizofréniát kísérő mutációk hatását lehet megfigyelni. A modell a betegség hatására lejátszódó biológiai folyamatok pontosabb beazonosítását és potenciális gyógyszercélpontok hatásának jellemzését teheti lehetővé.

Az egészséges és kóros idegsejtműködés közötti eltéréseket kontroll és mutáns sejtvonalak összevetésével elemeztük. Többféle szempontból lehet eltéréseket keresni, többek között a sejteken belüli kalcium (Ca^{2+}) szint vizsgálatán keresztül. Én az utóbbi mérésnek az adataival foglalkoztam. A mérés során keletkezett adathalmaz a mérésben résztvevő egyes sejtek kalcium-szint idősorait jelenti. A mérést követően a mikroszkóp felvételén manuálisan kijelöltük az egyes sejteket, ún. ROI-kat (Region of Interest, „releváns tartomány”) létrehozva. Egy ilyen ROI idősoron az x tengelyen az idő, az y-tengelyen közvetetten a kalcium-szint szerepel, amit az átlagos fluoreszcencia-jelintenzitás segítségével mérünk közvetlenül. A kalcium-szint mérését részletesen a III. fejezetben ismertetem.

A sejtkultúra Ca^{2+} szintjét először beavatkozás nélkül mértük meg, megkeresve a nyugalmi alapvonal szintet. Ezután ingerület átvivő anyagot, glutamát oldatot cseppentettünk a sejtkultúra kamrájába, amivel neuronokra jellemző tipikus Ca^{2+} jeleket váltottunk ki. A Ca^{2+} koncentráció beadás után jelentősen megnőtt, és magasabb szinten maradt hosszabb ideig.

Első kutatási célom arra vonatkozott, hogy automatizált módszert hozzak létre az alapvonal meghatározására. Egyes sejteket a beavatkozás előtti nyugalmi állapotban jellemezheti spontán aktív viselkedés. Ez azt jelenti, hogy a sejt Ca^{2+} szintje nyugalmi állapotban is képes lehet akkora szintet mutatni rövid ideig, esetleg több alkalommal is, mint az ingerlés kiváltása után lenne várható. Az alapvonal meghatározásához ezeket a spontán aktív hatásokat azonosítani és eliminálni szükséges a Ca^{2+} mérésekből.

Gaussian Mixture Model (GMM) segítségével adott ROI idősorában az ingerlő anyag beadása előtt szétválasztottam az adatpontokat alpopulációkra. Az empirikus sűrűségfüggvényeken megfigyelhető volt, hogy a pontok túlnyomó többsége a tényleges nyugalmi Ca^{2+} szint körül helyezkedett el, a többi pont magasabb értékek körül tömörült a spontán aktív tüzelések hatására. Alapvonalnak annak az alpopulációnak a várható értékét vettem, amely legnagyobb súllyal szerepelt a teljes populációban. A modellt hiperparamétereit 10-rétegű keresztvalidációval optimalizáltam.

Létrehoztam egy grafikus felhasználói felületet (GUI), ahová a mérések adatai feltölthetők. A GUI automatikusan szétválasztja két részre a beadás időpontjánál az adatbázist, majd keresztvalidáció segítségével megállapítja, hogy a GMM modell hány normális eloszlást illesszen adott ROI adatpontjaira. Ezután minden ROI-ra felépíti a GMM modellt az optimális alpopuláció számmal, majd a legnagyobb súlyú alpopulációnak a várható értékét, illetve az alapvonal segítségével számolt reakció értékét jeleníti meg kimenetként egy táblázat formájában.

Második kutatási célom a spontán aktivitás jelenlétének detektálása volt. Adott mérésen belül a spontán aktív sejtek aránya fontos információ a biológus kutatók számára. Labortagok segítségével tanuló adatbázisokat készítettünk a spontán aktív sejtekről. Az adatbázis segítségével egy véletlen erdő modellt hoztam létre a spontán aktív sejtek detektálására.

A véletlen erdő AUC értéke külön teszthalmazon 0,96 volt. 105 db ROI-ból összesen 8 db-ot azonosított be tévesen a modell. Ha minden sejtet spontán aktívnak tekintünk, ahol a véletlen erdő modell alapján becsült spontán aktivitási valószínűség legalább 57%, akkor a vizsgált teszthalmazon nem sorol hibásan ROI-kat spontán aktívnak a modell, és megtalálja az összes spontán aktív ROI több, mint felét.

Dolgozatom további felépítése a következő. A II. fejezetben a szkizofrénia betegséget ismertetem röviden, illetve áttekintem a gazdaságra kifejtett hatásait. A III. fejezetben a Humán Pluripotens Őssejt Laboratórium munkáját fogom bemutatni, és bevezetem a biológiai fogalmakat. A IV. fejezetben a kalcium-jel elemzéssel foglalkozó, hasonló tudományos munkák módszertanát vázolom fel az alapvonal meghatározásának szempontjából. Az V. fejezetben a Gaussian Mixture modellt mutatom be. A VI. fejezetben a véletlen erdő modell bemutatása következik, szót ejtek a tanuló adatbázisok létrehozásának nehézségéről is. A VII. fejezetben mutatom be a modelljeim eredményeit, illetve a modelljeimhez fejlesztett grafikus, felhasználói felületeket (GUI-kat). Az utolsó, VIII. fejezetben összegezem az eredményeket.

II. Szkizofrénia

A szkizofrénia egy súlyos mentális betegség, ami a felnőtt populáció kb. 1%-át érinti. (Kahn et al., 2015) Diagnosztizálni nagy kihívás, mert nincs olyan tünet, amely egyértelműen igazolná vagy kizárná a betegség fennállását. Tünetek együtteséből már könnyebben lehet következtetni, ám ezek rendkívül változatosak lehetnek. Tipikus tünetcsoportok a gondolkodászavar, érzécsalódás, a téveszmék, érzelmi közöny. (Tringer, 2010)

A tüneteknek 3 csoportját különböztetik meg: negatív, pozitív és kognitív. Negatív tünetről beszélünk, hogyha bizonyos viselkedésminták elvesztése, tompulása figyelhető meg. Ide sorolható például az érzelmi kifejezőképesség csökkenése, a beszédhang monotonná válása. Pozitív tünetekről van szó, hogyha a beteg új viselkedésmintákat vesz fel, például hallucinálni kezd, tévképzetei lesznek. Kognitív tünetek esetében a beteg észlelése érintett, nehezebben koncentrálni, dolgozza fel a bejövő információt, memóriazavar is előfordulhat.

A szkizofréniát 2013-ban világszerte az első 25 fogyatékosághoz vezető okok között tartották számon. (Vos et al., 2013) A betegek legalább felének ismételt kórházi ellátásra van szüksége. (Tringer, 2010) Az ebből fakadó közvetlen költségek, például a gyógyszerek, a kúra ára, komoly mértékben megterhelik az egészségügyi ellátórendszert. A közvetett, nehezen feltérképezhető költségek ennek többszörösét is kitehetik. Ide sorolható például a kiesett jövedelem, és így az elmaradt adó- és járulékbetételek, a szociális segélyek költségei.

Az Egészségügyi Világszervezet (WHO) 1998-as becslése szerint a szkizofrénia közvetlen költségei a nyugati országokban az egészségügyi kiadások 1,6%-2,6%-át teszik ki, amely a bruttó nemzeti termék (GNP) 7%-12%-át jelenti. (Barbato, 1998, 15. oldal) Úgy vélik, az összköltség 50%-85%-áért nem a közvetlen, hanem a közvetett költség felelős. (Chong et al., 2016). A szkizofréniában szenvedő betegek nem tudnak egészséges mindennapi életet élni, várhatóan 12-15 évvel hamarabb meghalnak az átlagos korhoz képest. (van Os, 2009)

Az előbb leírtak alapján tehát érthető, miért olyan fontos, hogy a szkizofrénia hátterében zajló biológiai folyamatokat megértsük, így diagnosztizálása és kezelése egyszerűbbé és hatékonyabbá váljék. A hatékonyabb kezelés nem csak a betegek életminőségének javulásához járhatna hozzá, hanem a közvetett költségek csökkenéséhez is.

III. A kutatócsoport munkája

A kutatócsoport tagjai azonosítottak egy szkizofrén férfi beteget, akinek a szülei egészségesek. Genetikai vizsgálatok során megállapítottuk, hogy a férfi egy olyan mutációt hordoz génjeiben, amely szülei génjeiben nem található meg. Ezt a jelenséget de novo mutációnak (DNM) nevezik.

Ebben az esetben a mutáció nem szülőktől örökölt, hanem spontán mutáció történt sejtosztódáskor a magzatban a DNS-kettőződés folyamata során. A kutatócsoport egy sejtes modellrendszer felállításán és vizsgálatán keresztül arra a kérdésre keresi a választ, hogy ez a bizonyos de novo mutáció hozzájárulhat-e a szkizofrénia kialakulásához ebben a konkrét esetben.

Ahhoz, hogy a mutációnak a biológiai jelentőségét és hatását vizsgálni tudjuk, a betegtől és egy tőle teljesen független személytől kontrollként vérmintát vettek. A vérminta felhasználásával indukált pluripotens őssejt (iPSC) vonalat készítettek, majd ezekből neurális progenitor sejteket (NPC-k) differenciáltak.

Egyszerűbben fogalmazva, a vérsejteket géntechnológia segítségével vissza lehet programozni pluripotens őssejt állapotba. Az őssejt egy nem specializálódott sejt, azaz a sejtnek nincsen saját, kitüntetett feladata (Evans, 1981). Egy őssejt korlátlanul tud osztódni, és specializálódott utódsejtek létrehozására képes. Pluripotens őssejt állapotról az őssejt érésének 5-7. napja között lehet beszélni. Az ebben az állapotban lévő őssejt minden testi sejt irányába képes differenciálódni. Hétköznapi nyelven fogalmazva, bármilyen sejt kialakulhat belőle az érés során. Amikor ez a specializálódás megtörténik, azt hívják differenciálódásnak. (Apáti, 2016) Ennek az irányát a kutatók meghatározott molekulák hozzáadásával befolyásolni tudják, esetünkben az iPSC vonalból így jöttek létre idegsejtek. (Takahashi, 2006)

Összefoglalva, neuron sejt kultúrákat lehetett kitenyészteni a szkizofrén betegből és a független kontroll személyből. Ezeket a sejt kultúrákat sejt vonalaknak nevezik, mert genetikailag homogénnek tekinthetők, és megfelelő körülmények között korlátlan ideig fenntartani és tenyészteni lehet őket.

A különféle kísérleteket a kutatók ilyen sejt vonalakon végezték el, összesen 2x2 pár neuronsejt vonalat vizsgáltunk.

Az első párban lévő sejt vonalak a szkizofrén betegből származnak. Az egyik sejt vonal hordozza a mutációt, ezt kódoltuk „3PC1”-nek. A pár másik tagja pedig e sejt vonalnak a genetikailag módosított, kijavított („egészséges”, vad típus) változata, melyet a „#23” kód jelöl. Utóbbit vad típusú sejt vonalnak nevezik, mert a génnek azt a formáját tartalmazzák, amely a természetes populációban (tehát a „vadonban”) a leggyakoribb.

A másik pár tagjai egy független kontroll személy egészséges szervezetéből származnak. Itt az eredeti sejtvonal vad típusú (XCL1), a másik pedig a genetikailag módosított elrontott változat, amely mesterségesen tartalmazza a mutációt (#25).

A sejtvonalakat jelölő rövidítéseket az alábbi táblázat tartalmazza:

	mutáns	vad típus
szkizofrén beteg	3PC1	#23
egészséges kontroll	#25	XCL1

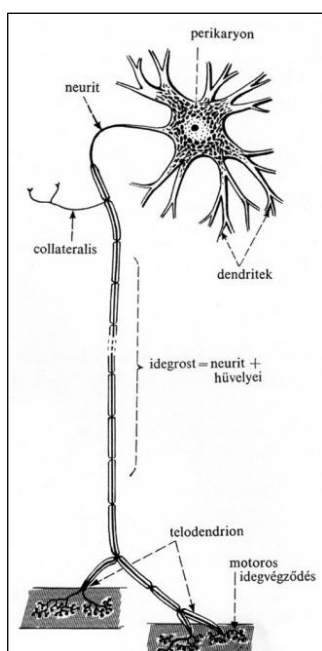
1. táblázat: Sejtvonalak elnevezése

Átfogó kutatási kérdésünk az, hogy található-e a mutáns és vad típusú sejtközösségek tulajdonságai, viselkedése között eltérés. Ha az egyik pár esetében különbséget találunk, a másik párt is meg kell vizsgálni egyfajta kontrollként, hogy ott is fennáll-e ugyanaz az eltérés, így validáljuk az eredményünket. Azaz példának okáért, ha a 3PC1 és #23 sejtvonalak között valamilyen eltérést találunk az egyik kísérlet során, ugyanezt a kísérletet el kell végezni a #25 és XCL1 párra is, és megvizsgálni, ott is fennáll-e az eltérés. Semmi megkötés nincs arra vonatkozóan, hogy hol és milyen különbséget kell keresni. Rengetek különféle módszer áll rendelkezésre (pl.: elektromos aktivitás vizsgálata, különböző fehérjék kifejeződésének vizsgálata, sejtosztódás, sejtmozgás, környezeti hatások tűrőképessége), a dolgozat keretein azonban túlmutat ezek részletezése. Munkám során az egyik jól bevált módszerrel, a sejten belüli kalcium-jel elemzéssel foglalkoztam.

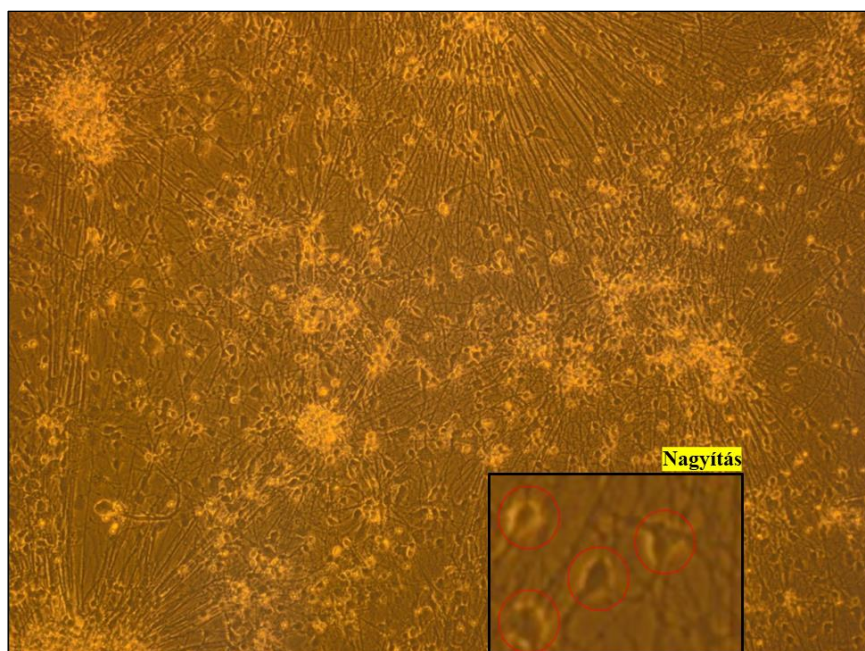
A kalcium ionok (Ca^{2+}) a sejt jelátviteli folyamataiban kulcsfontosságú szerepet töltenek be, a sejtek életének szinte minden aspektusára hatással vannak. A sejtek reagálnak a környezeti változásokra, az ezekhez való alkalmazkodás végett különféle jeleket kell leadniuk. A jelzésekhez olyan hírvivőkre van szükség, amelyek koncentrációja az idő függvényében változik. Ezt a szerepet tölti be a kalcium ion, amely a sejtek jelátvitelét az egyik legnagyobb mértékben szabályozza (a másik fontos másodlagos szabályozó a foszfátion). (Clapham, 2007)

A sejt viselkedését a sejten belüli Ca^{2+} szintjén keresztül vizsgáltuk. A sejtvonal Ca^{2+} szintjét először nyugalmi állapotban mértük, majd bizonyos idő után, miután a mikroszkópot kezelő kutató úgy látta, hogy a neuronsejtek „normálisan” viselkednek (azaz a nyugalomban mért kalcium szint körülbelül állandó: egy vízszintes egyenes, a zajt leszámítva nincsen benne trend), következhetett az ingerlő anyag, a glutamát beadása. Ennek hatására a neuronsejtek tüzelni kezdtek, a kalcium szint jelentősen megugrott, majd fokozatosan ürülni, csökkenni kezdett.

A kalcium szint követéséhez egy ún. kalcium szenzor fehérjét (Fluo-4) használtunk. Ez a kalcium szenzor egy olyan speciális fehérje, aminek a fluoreszcenciája attól függően változik, hogy mekkora a sejten belüli Ca^{2+} koncentráció. A fluoreszcencia a fehérje által kibocsátott fényt jelenti. Fluoreszkálás során az anyag (itt: a fehérje) elnyel egy adott hullámhosszúságú gerjesztő fényt, majd kibocsát egy ettől eltérő hullámhosszúságú fényt. Magasabb intenzitású fluoreszcencia esetében nagyobb Ca^{2+} szintről kapunk információt. Fluoreszcencia mikroszkóppal tudtuk ezt a fluoreszcenciát kiváltani és detektálni, 0,85 frame/secundum időbeli felbontással. Így részletes térbeli és időbeli információt tudtunk gyűjteni a kalcium szint változásairól.



1. ábra: A neuron általános sémája (Forrás: Szentágothai, Réthelyi, 2006, 155. oldal 2/53. ábra)

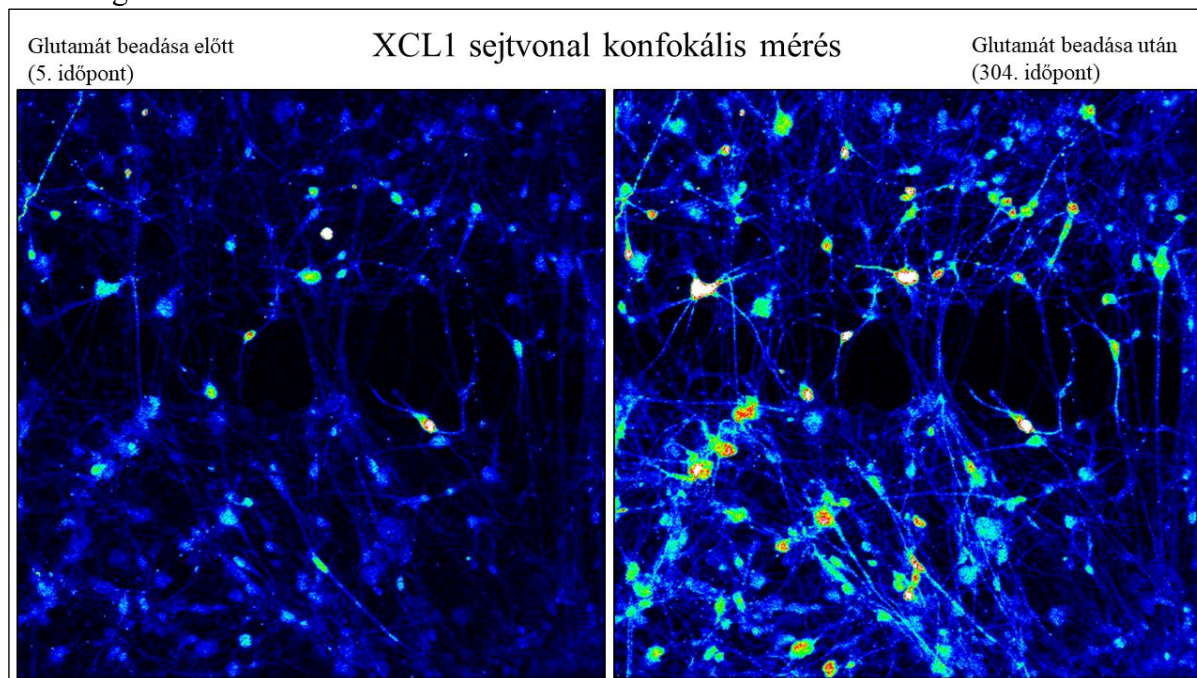


2. ábra: Neuronhálózat mikroszkópos képe (3PCI, 6 hét differenciáció után)

Mielőtt a fluoreszcencia mikroszkópos felvételen lévő sejteket bemutatnám, az 1. ábra segítségével ismertetem, hogy néz ki egy neuron (idegsejt). A perikaryon az idegsejttest, amelyből a kis nyúlványok, a dendritek ágaznak ki. A dendritek fő feladata az információfelvétel. A neurit, másnéven axon pedig az információ továbbítására szolgáló hosszú idegnyúlvány. Dendriteken és neuriten keresztül kapcsolódnak egymáshoz, szerveződnek hálózatba az idegsejtek. A neuron többi alkotóelemének ismertetése ezen dolgozat szempontjából nem releváns.

A következő ábrával szemléltetem, hogy néz ki egy neuronhálózat a valóságban. A 2. ábrán egy neuronhálózatról készített mikroszkópos felvétel látható. A kisebb, barna ovális foltok a

sejttestek. A jobb oldali sarokban egy kis négyzetben a felvétel egy ránagyított része látható, amelyben piros körök jelölik a sejttesteket. Jól látható köztük és az egész felvételen is a sejttestek közötti rengeteg neurit (a hosszúkás vonalak), amelyek hálózatba szervezik a sejtek összességét.



3. ábra: XCL1 sejtvonalaokról fluoreszcencia mikroszkóppal készített mérések az 5. és 304. időpontban (glutamát beadásának időpontja: 299)

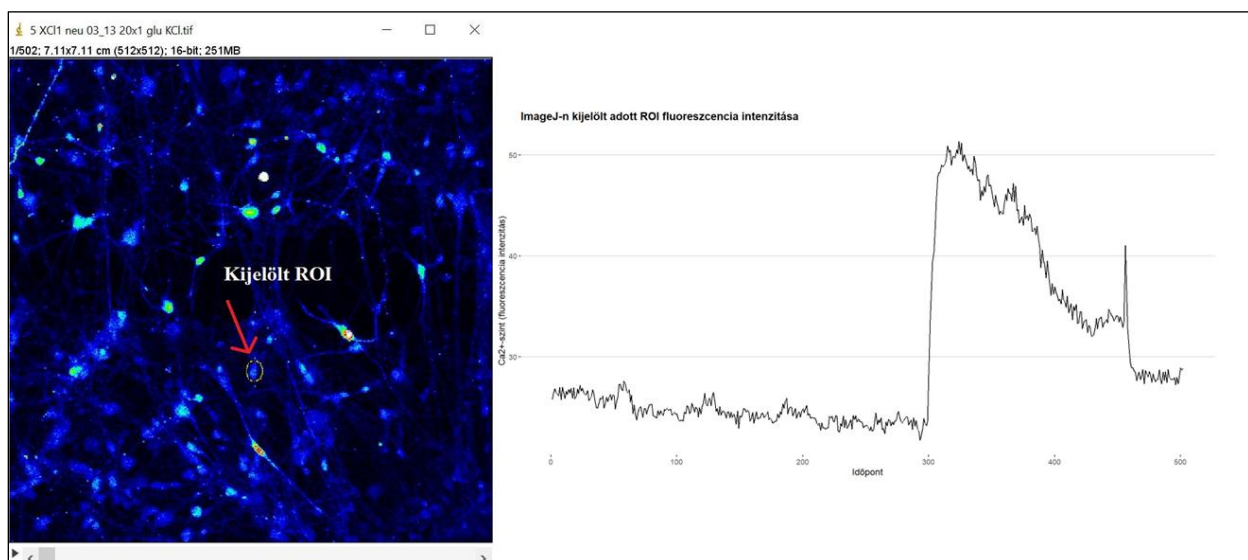
Az 3. ábrán az XCL1 sejtvonalból, azaz a független egészséges kontroll sejtvonalból kitenyésztett sejt kultúra látható. A fény intenzitása jelöli a Ca^{2+} szintet. Ahol fekete a kép, ott nem található sejt. A fluoreszcencia mikroszkópos felvételen (3. ábra) ugyancsak az ovális és kerek foltok jelölik a sejttesteket. A hosszú „csíkok” a nyúlványok, azaz a neuritek. A felvétel a valóságban monokróm, a fluoreszkálás jelenségének láthatóbbá tétele érdekében mesterséges színezetet („16 colors”) raktunk rá.

A sötétkék szín mutatja az átlagos nyugalmi Ca^{2+} szintet, míg az ennél világosabb színek (rendre: világoskék, világoszöld piros, majd fehér) a minél magasabb Ca^{2+} koncentrációt jelölik.

Az 3. ábrán látható, hogy a glutamát beadásának hatására az átlagos jelintenzitás nagy mértékben megnövekedett, magasabb lett a világoskék és világoszöld színek aránya, illetve a pirosé is. Fontos észrevenni azonban, hogy a glutamát beadása előtt is „világítottak” már sejttestek a sötétkéknél világosabb színnel, vagyis a glutamát beadása előtt is láthatóak magas jelintenzitású sejttestek. Ez a spontán aktív viselkedés jelenlétét mutatja, azaz azt, hogy az ingerlő anyag beadása előtt is képesek a sejtek tüzelésre, amikor a Ca^{2+} koncentráció spontán, rövid időre megnövekszik. Látható, hogy a glutamát beadása előtt csak néhány sejt világít

erősen, míg beadás után szinte az összes. A spontán aktivitás jelenségének a megmagyarázására és jelentőségére az IV.2. fejezetben visszatérek.

Egyik kutatási kérdésünk, hogy a glutamát beadása előtt és után hogyan változott a sejteken belül a kalcium szint.



4. ábra: ImageJ szoftverrel kiértékelt adott ROI fluoreszcencia intenzitása (XCL1 sejtvonala)

A 4. ábrán a bal oldalon az ImageJ képfeldolgozó program látható. A szoftverbe a mikroszkóp által készített képsorozatot kell feltölteni, majd az ImageJ-n belül manuálisan lehet kijelölni a sejtttesteket. Ezeket a kijelöléseket (azaz kijelölt sejtttesteket) nevezzük ROI-nak, azaz Region of Interest-nek (magyarul: releváns tartomány). A bal oldali képen a mérés 1. időpontja látható. Az ImageJ minden időpillanatban (itt: 1-502-ig) megméri a fluoreszcencia intenzitást, ennek idősora a 4. ábra jobb oldalán látható. A labor kutatásában nem csak egy ROI-t vizsgálnak a kutatók, hanem az összes sejtttest kijelölésre kerül, mindegyikhez tartozik egy saját idősor.

Egy mikroszkópos mérés során óriási mennyiségű adat keletkezik. A mikroszkóp egyszerre körülbelül 200 sejtről készít felvételt, általában 400 időpontban. Ilyen mérésekből összesen 18-at kaptam kézhez a munkám elején. Összességében több, mint egy millió adatpontról lehet beszélni. A mérések azóta is szüntelenül folytatódnak, az adattömeg csak nő. Az adatok alapos szűrésére és automatizált módszerekre van szükség a minél pontosabb eredmények érdekében.

IV. Kiértékelés

IV. 1. Adatvagyonkezelés

Az első lépés a biológusok által való előzetes szűrés elvégzése. A mikroszkóp által készített felvételen bizonyos sejtek nem kerülnek kijelölésre, mivel ezek a kutatás szempontjából nem

relevánsak. Ilyenek a „glia” típusú támasztó sejtek (pl. asztrociták), ezek szelekcióját morfológia alapú manuális szortírozással végezzük. Egy szakavatott biológus felismeri a jóval nagyobb méretű és lassabban, kevésbé intenzíven reagáló asztrocita sejteket. Ha véletlen mégis bekerülnek a mérésbe ezek a sejtek, az idősoron látszik az eltérő reakció, ilyenkor utólagos szűrésre van szükség.

Ezen felül a halott sejteket is ki kell szűrni. Ezeket az eleve nagyon magas fluoreszcencia intenzitásuk alapján ugyancsak könnyen fel lehet ismerni. A mikroszkóp képén ebben az esetben egy hófehéren világító pontot lehet látni, amely szabadon úszik a kamra felszínén.

A harmadik típusú sejtet, amelyet még szűrni kell, viszont már csak utólagosan lehet felismerni, a ROI-k kijelölésekor. Vannak olyan sejtek, amelyek bizonyos okokból biológiailag eltérően viselkednek a megszokottól, például egyáltalán nem reagáltak a glutamát beadására. Ezek ugyancsak nem képezik a kutatás tárgyát, ezen sejtek sem kerülnek kijelölésre.

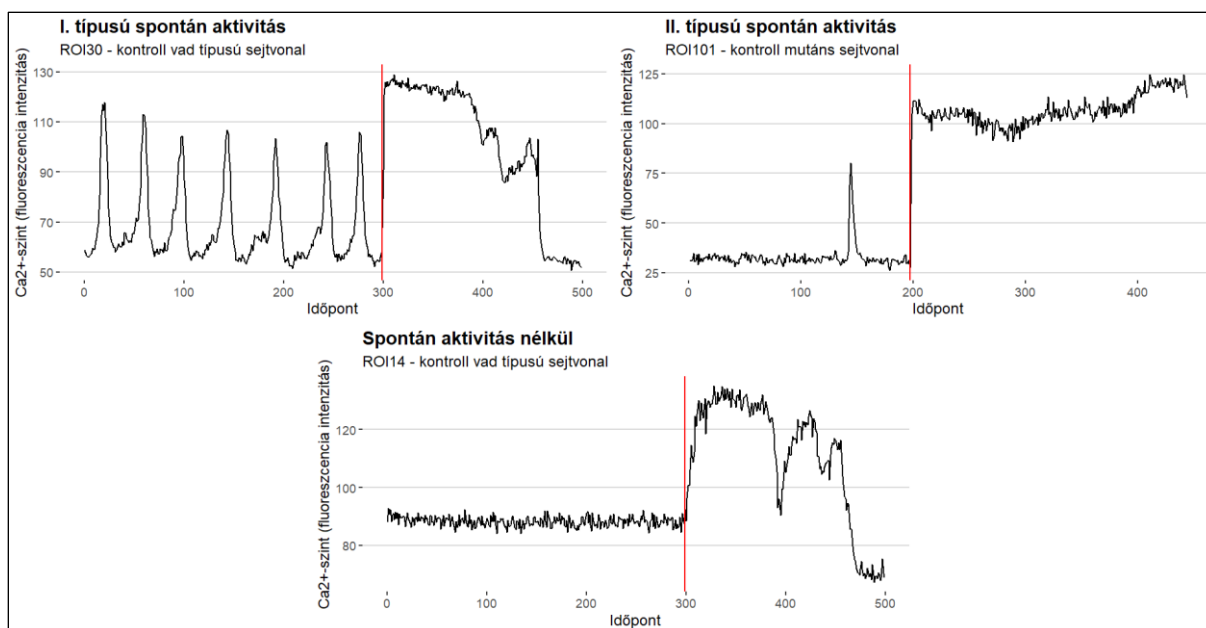
IV. 2. Nyugalmi állapot és spontán aktivitás jelenségének bemutatása

A reakció kiszámításához az első lépés a glutamát beadása előtt egy benchmark-ként szolgáló alapvonal megállapítása. Az alapvonal a nyugalmi állapot átlagos jelintenzitását írja le. Ez nálunk konkrétan a beadás előtti általános, ingerlés nélküli Ca^{2+} -szintet jelöli.

Ennek meghatározása azonban problémákba ütközik a spontán aktivitás jelensége miatt. „A spontán aktivitás a neuronok tüzelését jelenti érzékszervi bemenet hiányában. Ez a nem kiváltott vagy ingerfüggetlen aktivitás az idegrendszerek alapvető tulajdonsága.” (Uddin, 2020) A neuronok ezen tulajdonsága, hogy bármiféle externális hatás, ingerlés nélkül tüzelhetnek, látható az 1. ábra bal oldali felvételén is. A sejtek nyugalmi állapotban voltak, a glutamát ingerlő anyag beadása csak a 299. időpontban következett be. Mégis megfigyelhető a mikroszkópos felvételeken, hogy néhány sejt néha fel-felvillan, esetleg többször is.

Pontos definíció nincs arról, hogy nyugalmi állapotban mekkora Ca^{2+} -szint emelkedés számít spontán aktív viselkedésnek. Általános jellemző azonban, hogy a spontán aktív állapot csupán rövid ideig áll fenn, a tüzelés után hamar visszazuhan a Ca^{2+} -szint a nyugalmi alapvonal szintre. Közös munkánk során a spontán aktivitásnak alapvetően két típusát különböztettük meg. Az I. típusú spontán aktivitás során a Ca^{2+} -szint több alkalommal is megemelkedik. Ekkor a neuronhálózaton egy ingerület jár körbe-körbe, az okozza az emelkedés ismétlődését. Egyfajta szezonális jellege van, ám a periódusok nem feltétlen szabályosak, emiatt a dekompozíciós idősortmodellt nem lehet használni a hatásnak a kiszűrésére. Ennek részletei bővebben az IV.3.4. fejezetben találhatók. A II. típusú spontán aktivitás során a nyugalmi állapotban a Ca^{2+} -szint

csupán egyszer emelkedik meg. Ezt tükének („spike”) szokás nevezni. Ebben az esetben az adott sejt mikrokönyezete úgy alakult, hogy kialakult egy „lokális inger”, amire a sejt tüzelni kezdett, ám az ingerület nem terjedt szét.



5. ábra: Nyugalmi állapot a spontán aktivitás két típusának jelenlétében és anélkül

Az 5. ábrán három ROI Ca^{2+} -szintjének az időszora látható. A függőleges piros vonal jelöli a glutamát beadásának az időpontját. A beadás előtti időszakot nevezzük nyugalmi állapotnak. A spontán aktivitás jelensége ekkor állhat fenn. A ROI14 alsó idősorán egy spontán aktivitás nélküli nyugalmi állapot látható. Az adatpontok egy állandó érték körül mozognak, kisebb-nagyobb zaj kíséretében, az idősorban nincs trend (nyugalmi állapotban). Az idősorban található zaj egyrészt mérés technikai eredetű, másrészt nem idegsejt jellegű szignálfluktuációból tevődik össze. Ezek lassabb, nem olyan nagy változások szoktak lenni. A másik két, a ROI30 és ROI101 idősor esetében azonban nem ez a helyzet, ezek esetében rendre I. és II. típusú spontán aktivitásról beszélhetünk. Kilógó értékek tarkítják a mérést, amelyek megnehezítik a nyugalmi állapot átlagos jelintenzitásának, az általános Ca^{2+} -szintjének, azaz az alapvonal értékének a megállapítását.

IV. 3. Alapvonal meghatározása a kutatócsoportunkban és más hasonló tudományos munkákban

Az alapvonal meghatározására többféle módszer is létezik. A kutatócsoportunk és más, hasonló témával foglalkozó kutatócsoportok által használt főbb módszereket az alábbi, 2. táblázat tartalmazza:

Hivatkozás	Kutatási téma	Alapvonal idézet	Módszer
Hathy et al., 2020, 6. oldal*	De novo mutációk hatásának vizsgálata a szkizofréniában	„(...) és az alapvonal értéke a glutamát beadása előtti 20 időpontnak az átlaga.”	ingerlés előtti rövid időszak átlaga
Gover et al. 2003, 2. oldal	Kalcium-jelátvitel elemzés a környéki idegrendszerben	„(...) az egyes terminálok alapvonal fluoreszcenciáját (F_0) az ingerület kezdetét közvetlenül megelőző három-négy képkockányi fluoreszcencia átlagaként számították ki.”	
Robbins et al., 2021, 5. oldal	Összefoglaló a kalcium-jelátvitel mérésének és elemzésének a folyamatáról és módszereiről	„Az F_0 alapvonal értékek kiszámíthatók az ingerlés kezdete előtti, azonos régióban lévő értékek átlagolásával (Galizia et al., 1999).”	
Newman et al., 2022, 13. oldal	A szinaptikus diverzitás meghatározó tényezőinek feltárása	„(...) az egyes 30 másodperces filmek első 30 képkockájának átlagos minimális intenzitású képét használtuk alapvonalként (F_0). Ez a módszer lehetővé tette a (...) jelek szennyeződésének kiküszöbölését a film során a nyugalmi állapot alatt jelen lévő spontán események által.”	ingerlés előtti rövid időszak minimumainak átlaga
Jennings et al., 2019, 7. oldal	Neuronpopulációk szerepének vizsgálata a viselkedés szempontjából az orbitofrontális agyi kéregben	„Az alapvonal fluoreszcenciát a teljes felvétel során mért fluoreszcencia-értékek mediánjaként számítottuk ki.”	ingerlés előtti teljes (=nyugalmi) időszak mediánja
Hong et al., 2022, 10. oldal	Hormonok (noradrenalin és szerotonin) hatásának vizsgálata a vizuális ingerléssel kiváltott kortikális válaszokra	„a feszültségi jeleket 50 milliszekundumos (ms) ablak mediánjával simítottuk, hogy (...) a szórványos tüskék által okozott szennyeződést kiküszöböljük” & „(...) az alapvonal [az] átlagos feszültség a villogó fény [ingerlés – a szerző] megjelenése előtti 100 ms időszakban (...)”	simítófüggvény használata + ingerlés előtti rövid időszak átlaga

**A kutatócsoportunk tagjának cikke, amely hasonló témában íródott. Kutatócsoportunk korábbi publikációkban az átlag módszert használta az alapvonal megállapítására.*

2. táblázat: Alapvonal meghatározására használt módszerek

Kutatócsoportunkban a glutamát beadására adott válasznak az ingerlés utáni 50 időpont maximum értékét vettük. Azért a maximum értéket választottuk, mivel a mikroszkóp időbeli felbontása nem olyan részletes, 0,85 frame/secundum. Így lehetséges, hogy a valóságban még magasabb választ (azaz Ca^{2+} -szint növekedést) adtak a sejtek, a mikroszkóp azonban túl lassú ahhoz, hogy elkapja ezeket az értékeket, így mire elkészül a kép, már lejjebb esett a Ca^{2+} -szint.

Emiatt vettük el, hogy maximum helyett például a 99. percentilis átlagát vegyük, mivel biológiai szempontból még a maximum érték sem feltétlenül tükrözi teljesen a valós válaszártékot (a nem túl részletes felbontás miatt).

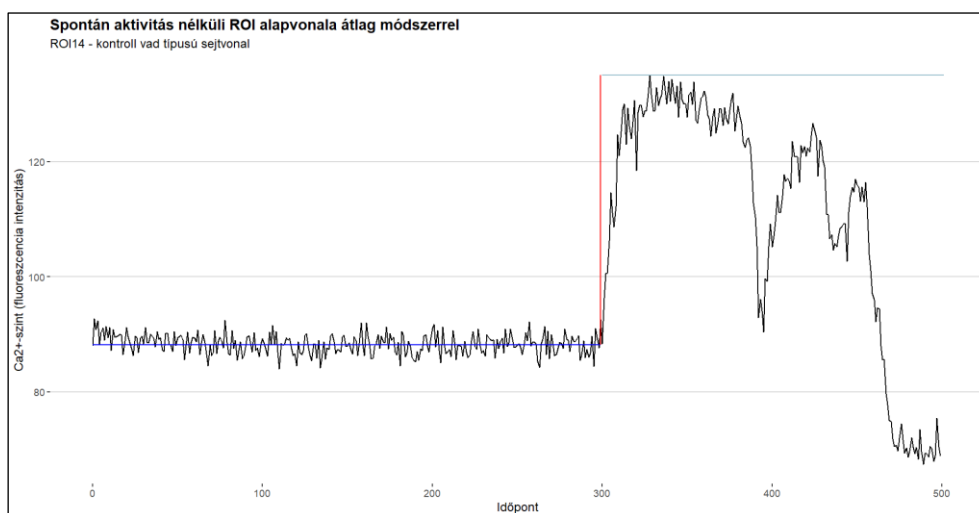
Annak az oka pedig, hogy 50 időponton belül vizsgálódunk és nem vesszük figyelembe a beadás utáni teljes időszakot, az az, hogy bizonyos mérések esetében a glutamát beadása után 1 perccel további oldatokat is hozzáadtunk a sejtekhez. Így a teljes időszak figyelembevételével nem tisztán a glutamát hatását kapnánk meg, hanem glutamát és további oldatok (pl. kálium) együttes ingerlésére adott választ.

A reakció értéke az alapvonal és a válasz értékeinek hányadosa. Szemléltetésül a következő ábrák (6-8. ábrák) tartalmazni fogják a válasz szintjét is.

A következőkben alfejezetekben mutatom be a különböző módszerek eredményeit és azt, hogy spontán aktivitás jelenléte hogyan befolyásolja negatívan ezeket a számításokat.

IV. 3. 1. Alapvonal mint ingerlés előtti rövid időszak átlaga

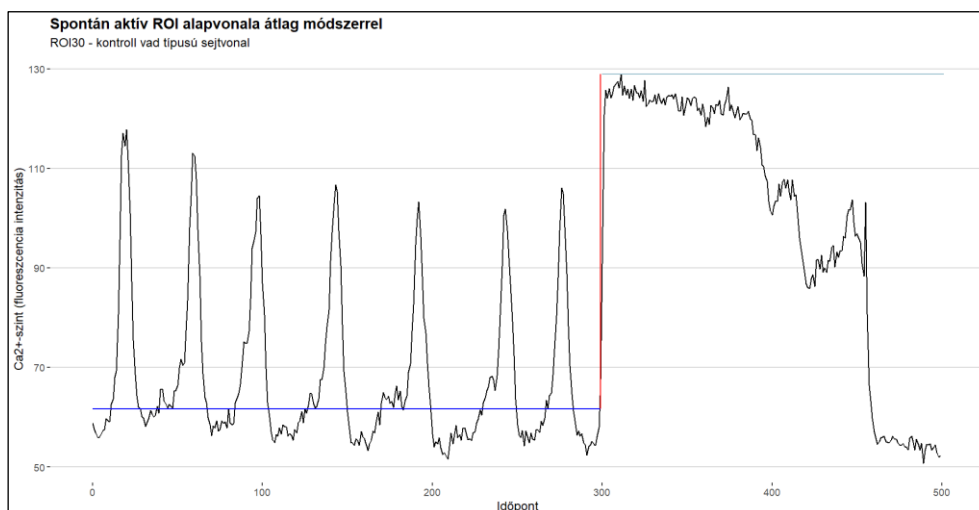
Az alapvonalat meghatározhatjuk az ingerlés előtti néhány időpont (kutatócsoportunk előző cikke alapján 20 (Hathy et al., 2020)) értékének az átlagaként. Egy valóban nyugalmi állapotban lévő sejt esetén a következő ábrát kapjuk ezzel a módszerrel:



6. ábra: Alapvonal meghatározása spontán aktivitást nem mutató ROI esetén átlag módszerrel

A 6. ábrán a sötétkék vonal jelöli az átlaggal számított alapvonalat, a világoskék a válasz értékét, a piros vonal pedig a beadás időpontjában a kettő különbségét mutatja. Spontán aktivitás hiányában valóban jól működhet ez a módszer. A 6. ábrán megfigyelhető, hogy az alapvonal körüli értékek egyszerűen zajnak foghatóak fel.

A II. típusú spontán aktivitás esetében ez a módszer működhet, ha az egyszeri tüzelés nem a beadás előtti „n” időpontban történik. Az I. típusú spontán aktivitás esetében azonban a megfigyelt tendencia az, hogy az ismétlődő emelkedések miatti kilógó értékek nagy eséllyel megjelenhetnek a beadás előtti „n” időpontban, így az átlaggal számított alapvonal értéke torzított lesz a valósághoz képest.



7. ábra: Alapvonal meghatározása spontán aktiv ROI esetén átlag módszerrel

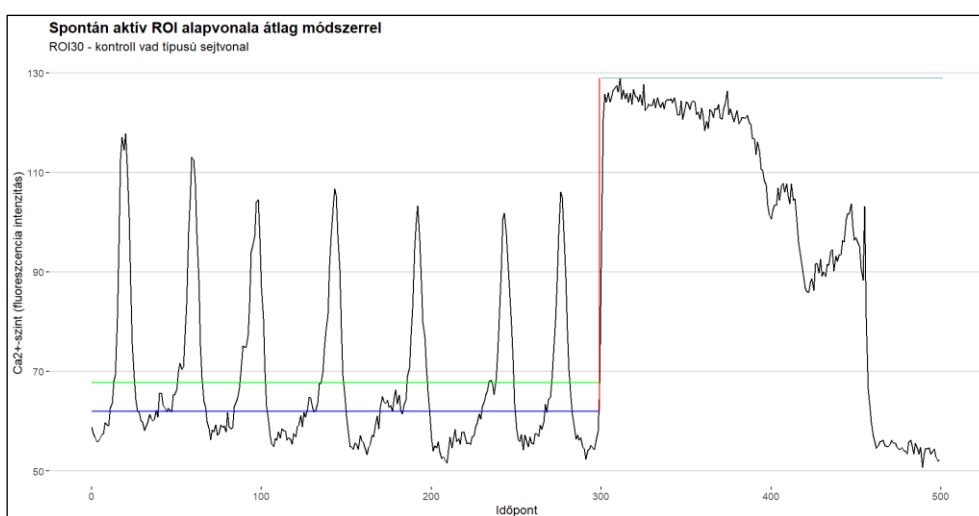
A 7. ábrán jól látszik, hogy az alapvonal mennyivel magasabbra került a spontán aktivitás kilógó értékeinek hatására, hiszen az átlag nem robusztus helyzetmutató. Az alapvonal értékének valahol a tüzelések közötti alacsony szinteken kéne lennie. Tehát az átlag módszer nem alkalmas olyan sejt kultúra alapvonalainak a megállapítására, amelyre jellemző a spontán aktivitás. A neuronsejtekre ez viszont jellemző, úgyhogy itt más módszer szükséges a nyugalmi állapot Ca^{2+} -szintjének meghatározására.

IV. 3. 2. Alapvonal mint ingerlés előtti (rövid időszak) minimum értékek átlaga

A minimum módszer esetében az ingerlő anyag beadása előtti teljes vagy rövid időszakban az „n” legkisebb érték átlagát vesszük, és ez lesz az alapvonal. Ez a módszer természetesen véd az ellen, hogy a spontán aktivitás hatására megjelenő kilógó értékek miatt túl magas legyen az alapvonal értéke. A spontán aktivitás hatásának eliminálására így alkalmas lenne ez a módszer, ám használatát egy általánosabb okból mégis elvetem. Mivel az idősorban található zaj, ha a minimumokat vesszük, akkor azzal a zaj alját kapjuk el, nem a valós nyugalmi állapot alapvonalát. Ekkor a reakció értéke nagyobb lesz a valóságosnál, falsul magas értékeket fogunk kapni akkor is, amikor a valóságban akár a sejt csak egy nagyon kis választ adott az ingerlésre.

IV. 3. 3. Alapvonal mint nyugalmi állapot mediánja

Az alapvonal megállapítására az átlagnál jobb módszer lehet a medián alkalmazása annak robusztussága miatt a kilógó értékekre. Spontán aktivitás hiányában biztonságosabb lehet ezt a módszert használni, hiszen, ha nagyobb zaj van a mérésben (például: fluktuáció), a medián jobban véd ellene, mint az átlag. Spontán aktivitás esetén a következőket lehet elmondani a medián módszerről. A II. típusú spontán aktivitás esetében, mivel ott relatíve az I. típusú spontán aktivitáshoz, nincs sok kilógó érték, a medián jó leírója lehet az alapvonalnak. Az I. típusú spontán aktivitás esetében azonban a kilógó értékek nagyobb aránya miatt a medián értéke ugyancsak eltolódik felfelé, még ha nem is akkora mértékben, mint az átlag. Emiatt a medián módszer sem teljesen alkalmas az alapvonal megállapítására.



8. ábra: Alapvonal meghatározása spontán aktív ROI esetén medián módszerrel

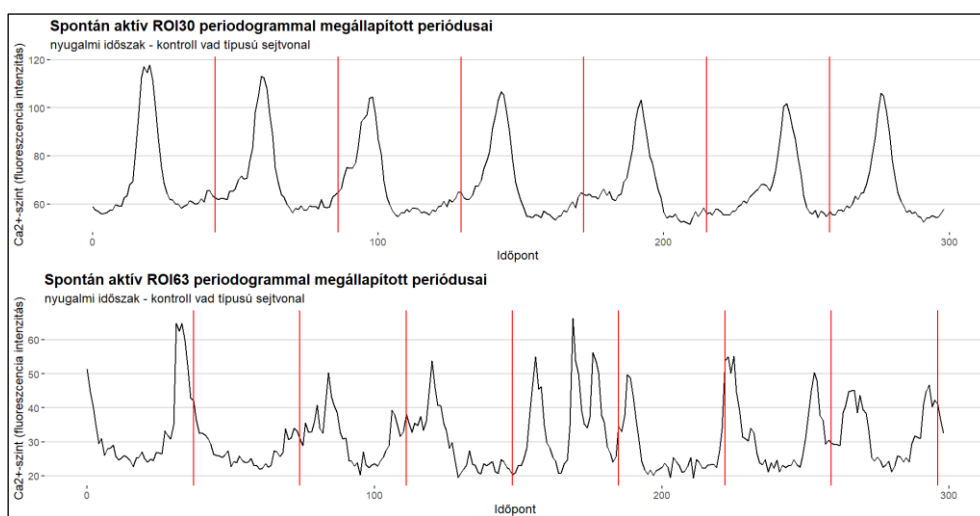
A 8. ábrán jól látszik, hogy a teljes nyugalmi időszakra számított medián (kék egyenes) mennyivel alacsonyabban van, mint az átlag (zöld egyenes), ám mindkét érték továbbra is távol áll a valóságos alapvonaltól, amelyet kutatócsoportunk a tüzelések közötti alsó, adott példában 55-60 közötti szintre szeretne megállapítani, amikor a mikroszkópos felvételeken is látszik, hogy éppen nincs ingerlés alatt a sejt.

IV. 3. 4. Simítófüggvény és dekompozíciós eljárások alkalmazása az alapvonal meghatározására

További módszerként sorolható fel az alapvonal értékének a meghatározására a simítófüggvény használata. A szakirodalomban egy más típusú probléma kapcsán felmerült már biológiai idősorok elemzése során simító függvények használata. (Moore et al., 2022) A tanulmány részletes bemutatása nem releváns, mivel a vizsgált kérdésük és az általuk számított alapvonal típusa eltér a kutatócsoportunk munkájában használt alapvonaltól. Hormonok felszabadításának a folyamatát vizsgálták sejtcsoportok kalcium-jeleinek szinkronizációján keresztül. Stimulálás

nélkül vizsgálták az alapaktivitást. Ötletet adtak arra, hogy simító függvényeket, illetve idősor dekompozíciós modelleket alkalmazzak a Ca^{2+} -szint idősorokon.

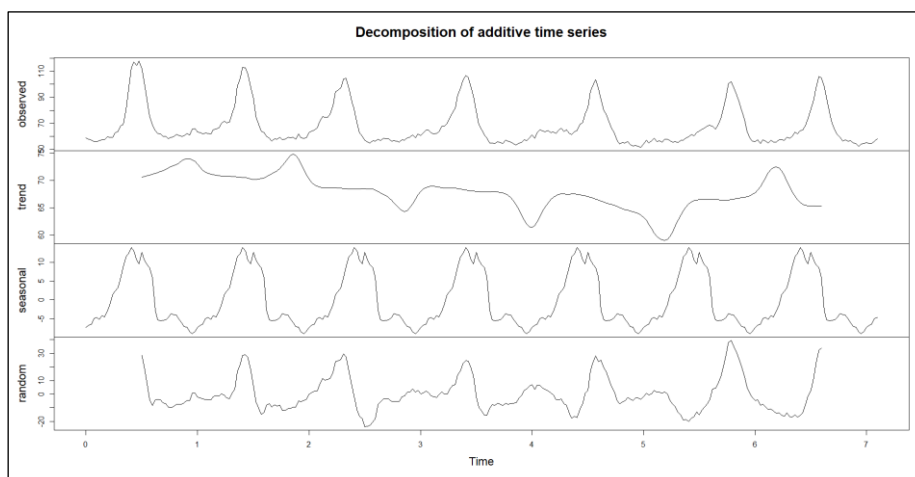
Az első probléma a Ca^{2+} -idősorok hagyományos idősorként való kezelésével az volt, hogy a dekompozíció (ami a simítófüggvények mögötti alapmodell, lásd Hodrick-Prescott szűrő) elvégzéséhez az idősort trend + szezonális + zaj komponensekre kell felbontani. A szezonális összetevőhöz azonban meg kell adni az idősor periódus hosszát. Ránézésre nem lehet megmondani a spontán aktív idősorainkról, hogy mekkora a frekvenciájuk, így a periodogram függvényhez fordultam segítségül. A periodogram függvényt jelfeldolgozás során használják, és alkalmas arra, hogy megállapítsa az idősor domináns frekvenciáját, azaz periódushosszát. (McSweeney, 2006) Hiába végeztem ám el ezeket a számításokat, a spontán aktivitás kiváltotta tüzelések egyáltalán nem ismétlődnek olyan szabályosan, mint például egy makrogazdasági idősor (pl. egy főre jutó GDP) esetében a hónap hatások. Így a dekompozíciós modellek nem működtek, mivel az idősort nem lehetett megtisztítani a „szezonhatástól”, azaz a spontán aktivitástól.



9. ábra: Periodogram függvénnyel megállapított periódushosszak spontán aktív ROI-k esetében

A 9. ábra két spontán aktív ROI a periodogram függvény által megállapított periódusait mutatja be (függőleges piros vonalak). ROI30 esetén a megállapított frekvencia 43, ROI63-nál pedig 37. Jól látható, hogy a szabályosabban viselkedő ROI30 (felső idősor) esetében is periódusonként eltér a viselkedés (hol hamarabb, hol később tör magasra a Ca^{2+} -koncentráció), ám a ROI63 esetében még komolyabb szabálytalanságot láthatunk.

Dekompozíciós modell használata ezeken az idősorokon nem vezet eredményre, a szezonhatást nem lehet eltávolítani az idősorról hagyományos módszerekkel.



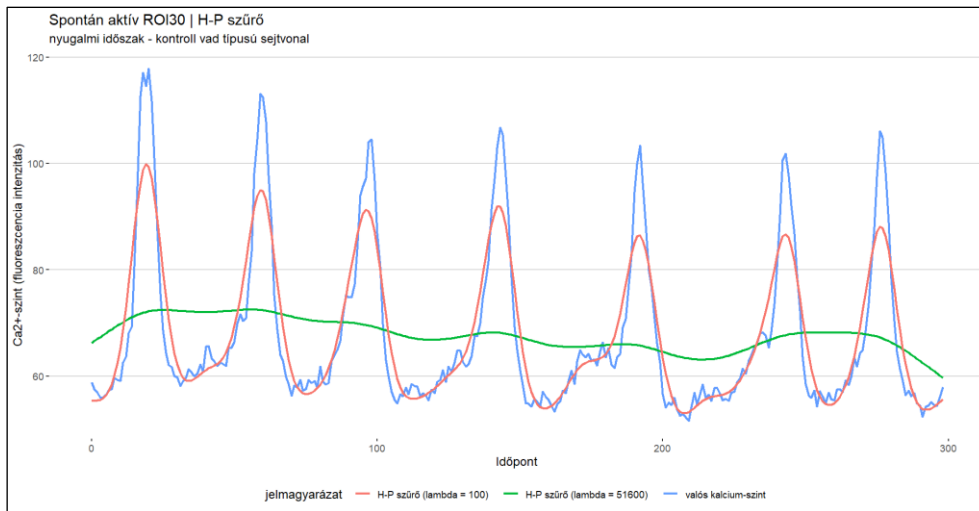
10. ábra: ROI30 (XCL1 sejtvonala) idősor dekompozíciós felbontása glutamát beadása előtt (1:299 időpontok)

A 10. ábra jól illusztrálja, hogy a trendet még a szabályosabban viselkedő ROI30 esetében (így az alapvonalat) sem lehet ezzel a módszerrel megragadni. A trend komponens tele van kisebb nagyobb kilengésekkel, a nem szabályos szezonális túl és alul becslései miatt. Tehát, a szezonális (azaz a spontán aktivitást) nem lehet ilyen módon kiszűrni az idősből. Ugyancsak érdemes megfigyelni, hogy a hibás viselkedése nem írható le stacioner folyamatként.

Simítófüggvények alkalmazását ugyancsak elvetettem. Munkám során a Hodrick-Prescott (H-P) szűrővel próbálkoztam. A H-P szűrő úgy számolja ki a simított idősor értékeit, hogy egy matematikai minimalizálási feladatot old meg. Az egyenlet és tagjainak pontos leírását mellőzném, mivel a dolgozat szempontjából a H-P szűrő alkalmazása nem hangsúlyos, csupán szemléltetés céljából szerepel. A lényeges, hogy egy ún. λ hiperparaméter állítgatásával lehet szabályozni, hogy a H-P szűrő mennyire közelítse a valós idősort. Minél kisebb ennek értéke, annál közelebb kerülnek a simított idősor értékei a valósokhoz. (Hodrick, Prescott, 1997)

A problémát az okozta, hogy két lehetséges kimenetelem volt: a λ értékét állíthatom alacsonyra, és akkor az eredeti idősort kapom vissza, csak zaj nélkül. Ezzel az a gond, hogy a spontán aktív hatástól ez a módszer nem szabadít meg. Ha simítás után veszek átlagot, ahogy Hong és kutatótársai tették, a spontán aktív hatás kilógó értékei továbbra is zavarni fognak, ahogy az előző IV. 3. 1. kifejtettem.

A másik lehetőségem, hogy nagyra állítom λ értékét, ekkor viszont igazából amit teszek, az az, hogy az egész nyugalmi időszakra számítok egy átlagot, amelyet az előző, IV. 3. 1. és IV. 3. 3. fejezetben tárgyaltam, hogy miért nem jó megoldás.



11. ábra: H-P szűrő alkalmazása spontán aktív ROI30-on

A 11. ábrán látható, hogy alacsony λ érték mellett (100) a H-P szűrő körülbelül leköveti az eredeti adatsort zajhatások nélkül. Magas λ érték mellett ($51600 = \frac{42 \cdot 14400}{12}$, a H-P ajánlása 14400 havi periódus esetén) pedig távol vagyunk a valósnak vélt alapvonaltól, az átlaghoz vagyunk közelebb.

Ezek alapján tehát a 2. táblázatban felsorolt Hong és kutatótársai megoldása sem alkalmazható a mi esetünkben. A simítófüggvénnyel eltávolították a zajt, de magát a spontán aktív hatást így nem lehet eltávolítani.

A következő, V. fejezetben azt a módszert mutatom be az alapvonal megállapítására, amely lehetővé teszi, hogy az alapvonal megállapításakor a spontán aktivitás hatása mellőzhetővé váljon.

V. Gaussian Mixture modell

A spontán aktív viselkedést mutató ROI-k esetén a Ca^{2+} alapvonal meghatározása az IV. fejezet 2. és 3. alfejezetei alapján egy zajszűrési feladatként fogható fel. Zajszűrési feladat esetén célunk a vizsgált változó valószínűségi eloszlásának dekompozíciója az eloszlás viselkedését legnagyobb mértékben meghatározó „fő eloszlásra” és zajhatásra. Azonban, a zajhatás több különböző forrásból származhat, eltérő eloszlásokkal. Éppen ezért az alapvonal megállapítására a spontán aktivitás hatásának eliminálása céljából a Gaussian Mixture modellt használtam. Ugyanis, ebben a modellben lehetőség van a vizsgált változónk eloszlását tetszőleges számú különböző hatás lineáris kombinációjaként reprezentálni. (Yu, 2012).

A Gaussian Mixture modell esetén a vizsgált változónk valószínűségi eloszlása felírható egy M komponensből álló Gauss-eloszlások (azaz: normális eloszlások) súlyozott összegeként:

$$p(x|\lambda) = \sum_{i=1}^M w_i g(x|\mu_i, \Sigma_i),$$

ahol x jelöli a D -dimenziós folytonos értékű adatvektort, $p(x|\lambda)$ pedig annak sűrűségfüggvényét. A λ halmaz tartalmazza összefoglalóan a súly (w_i), átlag (μ_i) és kovariancia (Σ_i) paramétereket. A w_i , ($i = 1 \dots M$) vektor, az eloszlások súlyait tartalmazza, míg $g(x|\mu_i, \Sigma_i)$ pedig a normális eloszlások sűrűségfüggvényét reprezentálja:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1}(x-\mu_i)},$$

ahol μ_i az átlagok vektora, Σ_i pedig a kovariancia-mátrix. A w_i súlyok összege 1. Az M egy hiperparaméter, ezzel a változóval azt lehet beállítani, hogy hány normális eloszlás lineáris kombinációjaként reprezentáljuk a vizsgált x adatvektor eloszlását.

A paraméterek becslése Maximum Likelihood (ML) becslésen alapul:

$$p(X|\lambda) = \prod_{t=1}^T p(x_t|\lambda) \rightarrow \max_{\lambda},$$

ahol T a tanuló vektorokat jelöli, $X = \{x_1, \dots, x_T\}$. Az ML becslés során a GMM modellt olyan λ paraméterekkel kívánjuk felépíteni, amelyek maximalizálják annak a likelihood-ját (azaz valószínűségét), hogy véletlen mintavétel esetén az általunk megadott (tanító) adathalmazt kapjuk vissza. (Reynolds, 2009)

Mivel a GMM-likelihood függvény a λ paraméterek nem-lineáris függvénye, közvetlenül nem lehet maximalizálni. Ezt az expectation-maximization (EM) algoritmussal lehet elvégezni. Ennek lényege, hogy felépít egy kezdeti modellt λ paraméterekkel, majd egy újabbat $\bar{\lambda}$ paraméterekkel úgy, hogy $p(X|\bar{\lambda}) \geq p(X|\lambda)$, azaz az újabb modell által becsült értékek bekövetkezési valószínűsége magasabb, mint az előzőé. A következő iterációban az új modell lesz a kezdeti modell. Ez addig ismétlődik, amíg egy konvergencia küszöbértéket el nem ér az algoritmus. A kezdeti modell általában egy bináris vektorkvantálás becslésből származik. (Ballard, 2000)

A modell hiperparaméterének, az eloszlások számának (M) az optimalizálására keresztvalidációt alkalmaztam. A klasszikus tesztek használatát (például likelihood-ratio test) azért vettem el, mert a modellszelekciós tesztek esetében ismerni kéne a teszt próbastatisztikájának határeloszlását igaz nullhipotézis (H_0) esetén. Egy olyan összetett modell

esetében, mint a GMM, ennek ismeretét nem feltételezhetjük. Esetleg bootstrap szimulációk futtatásával lehetne következtetni az eloszlásra, amelyek futásidőben költségesebbek, mint a keresztvalidációs szimulációk alkalmazása. (LeDell et al., 2015)

Az információs kritériumok önálló használatát is elutasítottam, mivel azok a sokaságot (nem megfigyelt adatpontokat) nem veszik figyelembe. A két gyakran használt információs kritérium, az Akaike- (AIC) és a Bayes-Schwarz információs kritérium (BIC) képlete a következőképpen néz ki,

$$AIC = 2k - 2 \ln(\hat{\mathcal{L}})$$

és

$$BIC = k \ln(n) - 2 \ln(\hat{\mathcal{L}}),$$

ahol $\hat{\mathcal{L}}$ a likelihood függvény maximalizált értéke, k a modellben szereplő paraméterek száma és n a mintaelemszám. Látható, hogy egyik egyenletben sem szerepel olyan tag, ami figyelembe venné a megfigyelt mintán kívüli, „out of sample” világot.

Reálisabb képet kaphatunk tehát, hogyha „resampling” módszert használunk, azaz keresztvalidáció segítségével állapítjuk meg, hány normális eloszlást illesszen a GMM. A keresztvalidálás lényege, hogy a bemeneti adathalmaz külön részét használja fel arra, hogy a modellt tanítsa (azaz elvégezze a maximum likelihood paraméterbecslést) és egy másik részét arra, hogy validálja, azaz tesztelje. Ez a folyamat pedig többször ismétlésre kerül (James et al., 2021).

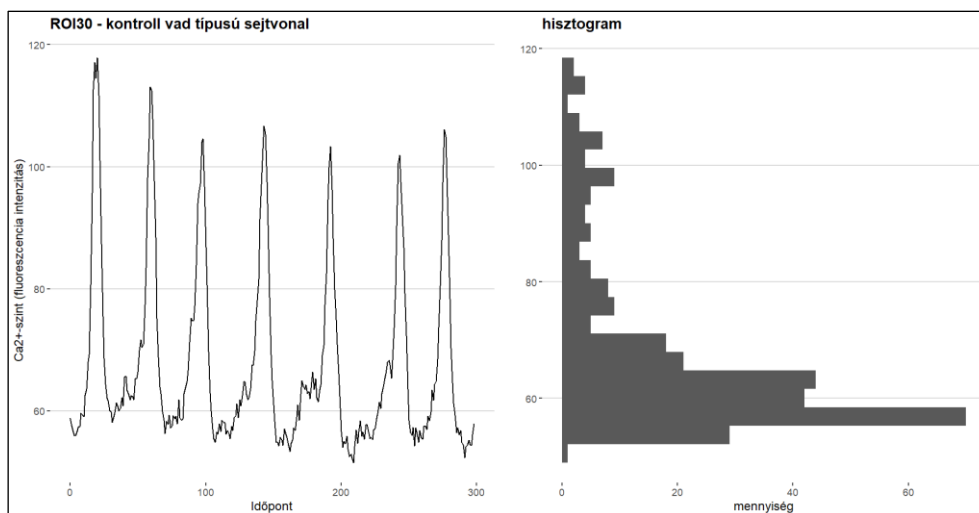
10 rétegű keresztvalidációt alkalmaztam, a hiperparaméter hangolására a Python scikit-learn csomagjának GridSearchCV függvényét használtam. A 10 rétegű azt jelenti, hogy a módszer az adathalmazt összesen 10 rétegre bontja fel, és ezek mindegyike pontosan egyszer lesz tesztelési rész. Azaz 10 iterációban mindig a teljes minta 9/10 része lesz tanuló és 1/10 része tesztelő minta, és minden iterációban az eredeti felosztás egy másik tizede lesz a tesztelő minta. A modellt betanítom a tanuló részen, majd validálom a tesztelőn, és ezt megismétlem összesen 10-szer. Végül átlagolom a validációk során kapott BIC értékeket. Mint láthattuk az információs kritériumok bemutatásánál, egy információs kritérium leírja a modell illeszkedését a maximum likelihood módszerrel megbecsült λ paramétervektor mellett oly módon, hogy a kisebb BIC jelent jobb illeszkedést. A keresztvalidáció során kapott átlagos BIC írja le a GMM teljesítményét egy adott hiperparaméter érték mellett. Amelyik hiperparaméter érték a

legkisebb keresztvalidált BIC értéket szolgáltatja, az tekinthető optimális hiperparaméter választásnak.

A keresztvalidáció kezdeti lépése a lehetséges hiperparaméter értékek megadása egy vektorban. Esetemben csak egy vektort kellett megadnom, amiben a lehetséges normális eloszlások számát 1-6-ig határoztam meg. Ezután a GridSearchCV függvény azt a hiperparaméter értéket adja meg kimenetként, amellyel a felépített modell „értékelő” (scoring) pontszáma a validáló adathalmazon a legmagasabb volt. A GridSearchCV függvény GMM modell hiperparaméterének optimalizálása során automatikusan beépített scoring függvénye a Bayes információs kritérium (BIC). Mivel a scoring függvényt a GridSearchCV maximalizálja, így itt valójában a negatív BIC szerepel, hiszen a maximum negatív BIC = minimum pozitív BIC. Azaz annyi normális eloszlást fog illeszteni, amire a negatív BIC értéke a lehető legmagasabb. A keresztvalidációt minden egyes ROI-ra külön futtattam.

A Gaussian Mixture modell mögötti alapkoncepció tehát az, hogy egy bonyolult (multimodális) eloszlás felírható sok egyszerű, normális eloszlás összegeként is. A modell megközelítése lényegében az approximációs tétel alkalmazása sűrűségfüggvényekre. Azt mondja a tétel, hogy bármilyen bonyolult függvény megközelíthető egyszerűbb elemek lineáris kombinációjaként. (Csáji, 2001) Ez a neurális hálózatok alaptétele is. (Kratsios et al., 2020)

A Ca^{2+} adatok esetében feltételezzük, hogy az adatpontok túlnyomó többsége nyugalmi állapotban az alapvonal szint környékén tömörül, ám az adatpontok egy kisebb, mégis jelentős része egy magasabb értéknél található a spontán aktivitás hatására. Az előzőeket támasztja alá a következő, 12. ábra is, ahol jól látható, hogy a spontán aktivitás hatására felvett magas Ca^{2+} -értékek környékén a jobb oldali hisztogramon a 100-as érték környékén púposodás látható, plurimodális eloszlást olvashatunk le. Az alapvonal környékén (kb. 55-ös érték) található természetesen a legtöbb adat, ám nem elhanyagolható a spontán aktivitás hatására felvett magas, 100 körüli értékek aránya sem, lokális módusról beszélhetünk.



12. ábra: Spontán aktív ROI30 időszora (beadás előtt) és histogramja

VI. Véletlen erdő modell

A véletlen erdő modell döntési fák kombinációjaként állítható elő.

A döntési fa egy olyan osztályozó algoritmus, amely nem rendelkezik matematikai előfeltevésekkel. Az adatok különböző attribútumai, azaz magyarázó változói, prediktorai alapján sorozatosan kettébontja az adathalmazt minél homogénebb csoportokra a kimenet (eredményváltozó) értéke szerint. Más szóval a modell heterogenitást minimalizál.

A döntési fa összesen 3 részből épül fel, ezek a következők: gyökér/törzs (root node), amely a kiindulópontot jelöli, az összes adatot. Ez a kezdeti pont, amikor még nem történt szétbontás. Ágak (branches), ahol már megtörtént az adatok szétválasztása bizonyos attribútumok mentén, de még nem alakultak ki a végleges csoportok, további bontások fognak következni. Levelek (leaf nodes), a végső csoportok, amelyeket az algoritmus már nem bont tovább. A 13. ábra mutatja be ezeket az összetevőket. Az ábrán egy mintaként szolgáló döntési modellt láthatunk, amelyet a Titanic utasainak túlélés szerinti osztályozására építettek. Legfelül látható a gyökér, a teljes minta még megtalálható itt. A következő lépés a bontás, amely itt a nemek szerint történt. A bal oldali ágon találhatóak meg a férfiak, a jobb oldalin a nők. Ezt további bontások követték, amíg el nem jutottunk a levelekig, amelyet a modell már nem bont tovább.

A spontán aktivitás detektálására az osztályozó véletlen erdőt alkalmasnak találtam, mivel pontos predikcióra képes a túlillesztés veszélye és a mélytanuló neurális hálózatok magas számítási kapacitása nélkül. Ez alapvetően egy klasszifikációs probléma, és a véletlen erdő egy népszerű osztályozó modell, bizonyítottan jó eredményekkel. (James et al., 2021) A spontán aktív ROI-k automatikus detektálása a kutatás szempontjából rendkívül fontos információ. Ha a vizsgált sejteket szétbontanánk két csoportra, spontán aktív és spontán aktivitást nem mutató sejtekre, ez további elemzési szempontokat jelentene, és vizsgálhatóvá válna sok új kérdés (például: nagyobb reakciót ad-e egy olyan sejt az ingerlésre, ami spontán aktivitást mutatott; mutáns és vad típus sejtvonalak között eltérő-e a spontán aktív sejtek aránya.)

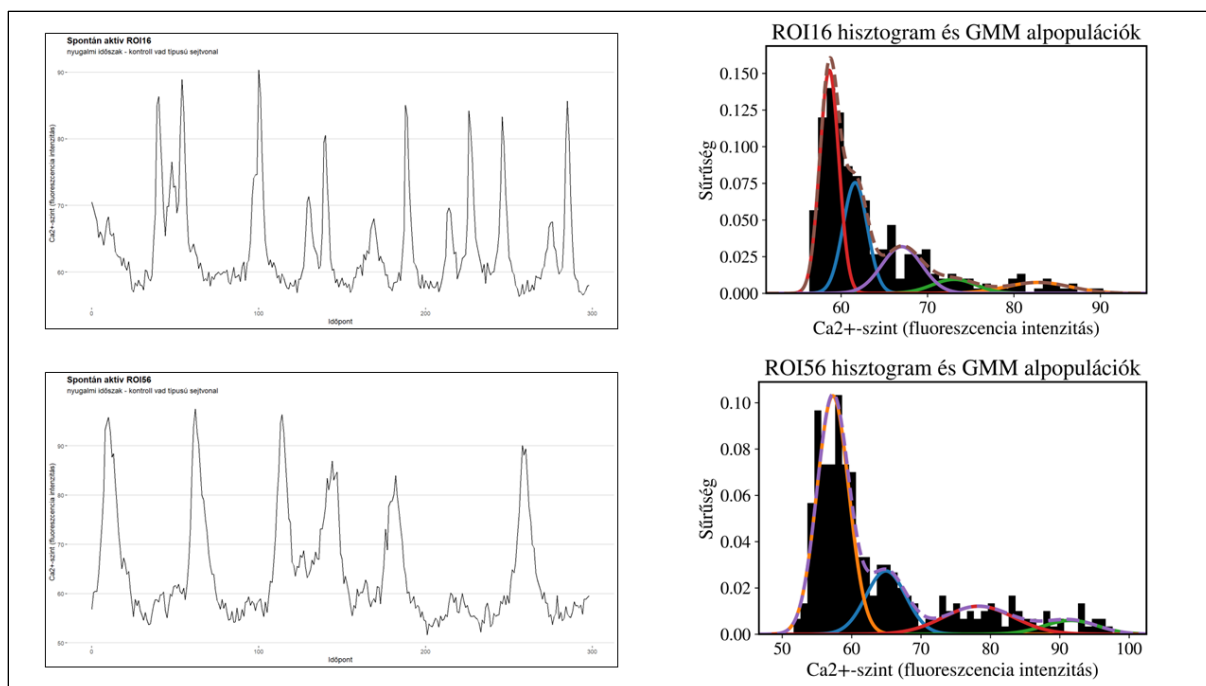
Véletlen erdő modellem felépítéséhez az R nyelvet használtam, a GMM-mel ellentétben, amelyhez pedig a Python nyelvet használtam. A véletlen erdő *mtry* hiperparaméter értékét 10 rétegű keresztvalidáció segítségével hangoltam. Erre a *caret* csomag *train* függvényét használtam, amelynek a GridSearchCV-hez hasonlóan ugyancsak meg kell adni a lehetséges hiperparaméter értékeket. Ezek értékét 1-15-ig adtam meg, mivel összesen 198 változóm volt, és alapesetben a hiperparaméter értékét a változók számának négyzetgyökeként határozzák meg, ami jelen esetben 14,07.

A keresztvalidáció során a tanuló adatokra modellt illeszt az algoritmus, majd a validáló mintán kiszámítja a modell teljesítményét, amelyet itt a pontosság (accuracy) mutató határoz meg. A pontosság mutató a találati arányt jelenti, amit úgy kapunk meg, hogy a helyes találatok számát elosztjuk az összes megfigyelés számával. Végül a legjobb pontossági teljesítményt elérő paraméterek kerülnek a végleges modellbe. Ennek értéke 15 lett.

VII. Eredmények

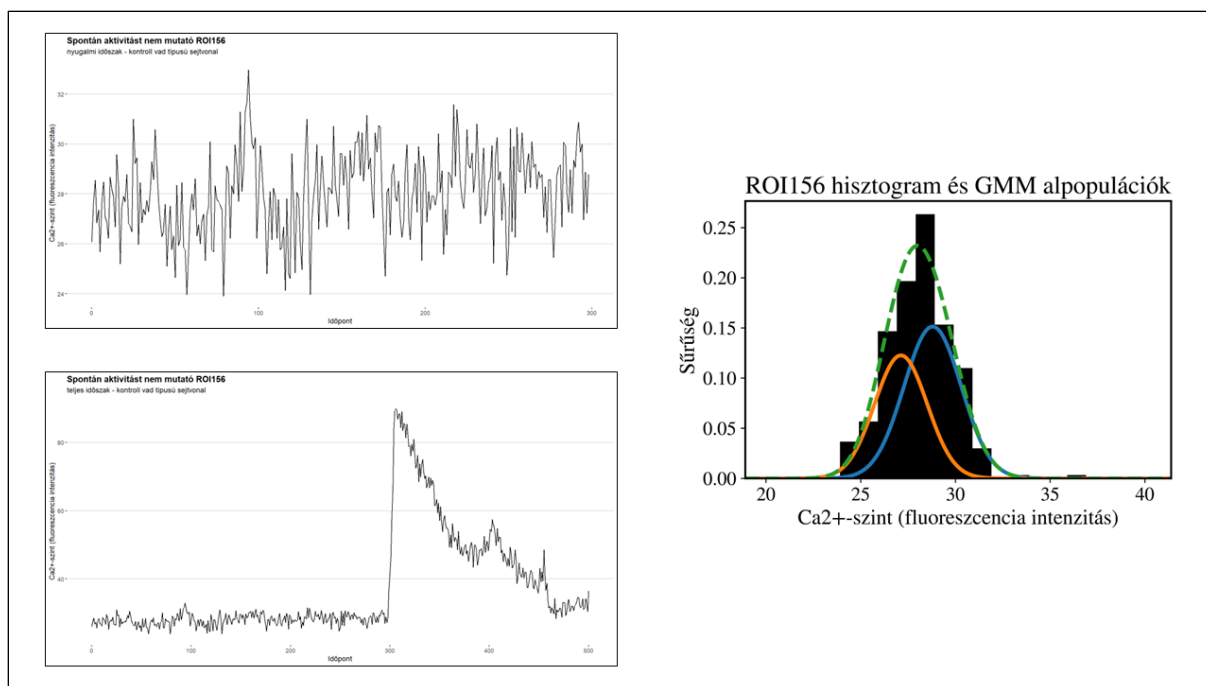
VII. 1. Gaussian Mixture modell eredmények

A Gaussian Mixture Model segítségével klaszterezni tudtam az adatpontokat különböző alpopulációkra. Ahogy az V. fejezet 12. ábráján is szemléltettem, mind a vonaldiagram, mind a hisztogram alapján az a sejtésünk lehet, hogy a spontán aktív ROI-k idősorának az adatpontjai több alpopulációból, azaz a GMM használatából következően több normális eloszlásból épülnek fel. Annak értékét, hogy hány normális eloszlást illesszen a modell adott ROI-ra, keresztvalidált likelihood számításával határoztam meg. A következő, 14. ábrán bemutatom spontán aktív sejtek hisztogram ábráján, hogyan néz ki a gyakorlatban a populációjuk GMM általi alpopulációkra bontása.



14. ábra: XCL1 sejtvonalból származó ROI16 és ROI56 spontán aktív sejtek idősorai és hisztogramjai nyugalmi állapotban, és az alpopulációk normális eloszlásai

Az „XCL1” sejtvonalból származó ROI16 és ROI56 spontán aktív sejtek. A 14. ábra bal oldalán a nyugalmi állapot idősorai láthatóak, jobb oldalon pedig ezen értékek hisztogramjai. A keresztvalidáció elvégzésével a ROI16 esetében 5, a ROI56 esetében 4 normális eloszlást optimális illeszteni a GMM-nek. A különféle színű, folytonos vonalak jelölik az alpopulációkra illesztett normális eloszlásokat, a szaggatott vonal pedig a normális eloszlások súlyozott összege, azaz a változó tapasztalati valószínűségi eloszlása.



15. ábra: XCL1 sejtvonalból származó ROI156 nem spontán aktív sejt időszora és hisztogramja nyugalmi és teljes időszakban, és az alpopuláció normális eloszlásai

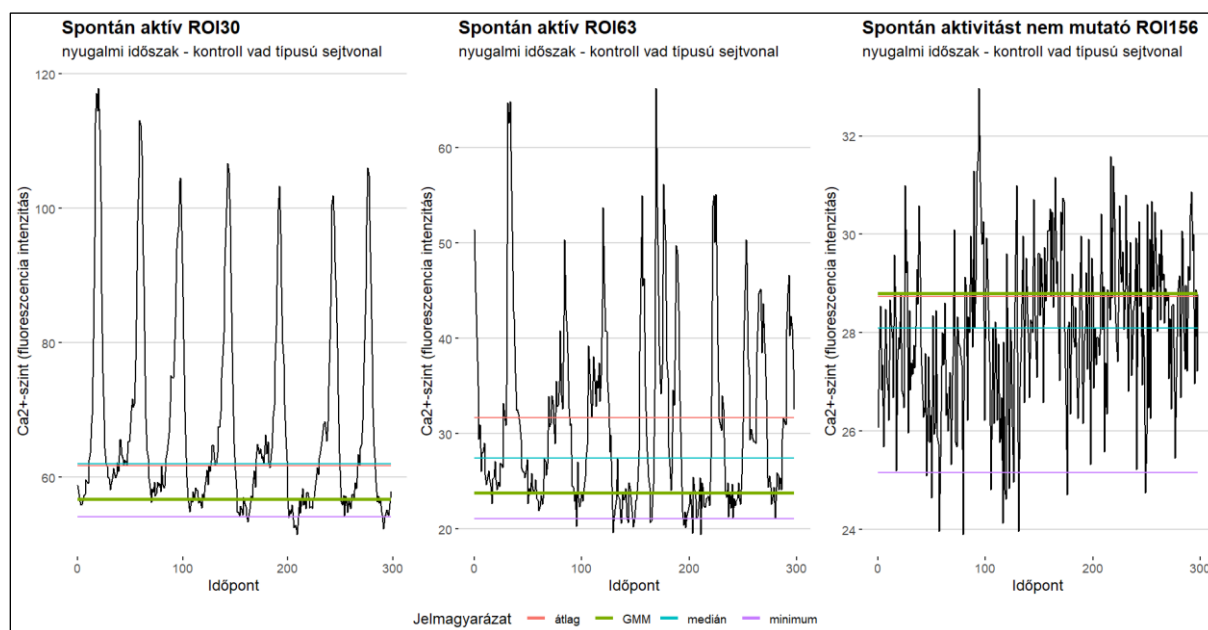
A 15. ábrán a spontán aktív tevékenységet nem mutató ROI időszora és hisztogramja látható, illetve a GMM által megállapított 2 alpopulációk normális eloszlásai. A bal felső ábra a nyugalmi, az alsó a teljes időszak idősorát mutatja.

Az idegsejtekre jellemző spontán aktivitás hiányában is beszélhetünk nem idegsejt jellegű szignálfluktuációról. Így a nem spontán aktív sejtek idősorában található zaj a mérési zaj mellett ezt a kisebb fluktuációs hatást is tartalmazhatja, így indokoltnak találjuk a GMM használatát a nem spontán aktív sejtek esetén is.

Alapvonalként annak a normális eloszlásnak a várható értékét, azaz átlagát vettem, amelyik a legnagyobb súllyal szerepel a modellben, más szóval az a normális eloszlás „klaszter”, ahová az adatpontok a legnagyobb valószínűséggel tartoznak. A következő ábrákon és táblázatokon bemutatom, hogy mekkora lett az alapvonal értéke GMM és egyéb alternatív (átlag, medián, minimum) módszerekkel 2 spontán aktív (ROI30 és ROI63) és egy spontán aktivitást nem mutató (ROI156) ROI esetében.

Az alternatív módszerek a következőket jelentik: átlag módszer = glutamát beadása előtti 20 időpont átlaga; medián módszer = nyugalmi állapot (glutamát beadása előtti teljes időszak) mediánja; minimum módszer = nyugalmi állapot 30 legkisebb értékének átlaga.

A táblázatok emellett a válaszokat (glutamát beadása utáni 50 időpont maximuma) és a reakciókat (válasz és alapvonal hányadosa) is tartalmazzák a különböző alapvonalak esetén.



16. ábra: Spontán aktív és spontán aktivitást nem mutató ROI-k alapvonalai 4 különböző módszerrel

A 16. ábrán az első két idősor a két spontán aktív ROI, ROI30 és ROI63 idősorát mutatja, illetve a különböző módszerekkel meghatározott alapvonalakat. A jobb oldali idősor a ROI156 nem spontán aktív sejt Ca^{2+} -szintjét mutatja be.

A minimum módszerrel meghatározott alapvonal értéke a legalacsonyabb, ha mutat spontán aktivitást a sejt, ha nem. A spontán aktív idősorok esetében megfigyelhető, hogy a GMM-mel meghatározott alapvonal veszi fel a 2. legalacsonyabb értéket. Ezt követi a medián és az átlag, ahol a medián általában az átlaggal meghatározott alapvonal alatt helyezkedik el (mivel az eloszlás jobbra elnyúló), de ahogy a ROI30 ábráján is láthatjuk, a medián és az átlag felvehet közeli értéket is egymáshoz. Azaz láthatjuk, hogy a GMM módszerrel hatékonyan szét tudtuk választani az adatpontokat, és az alapvonalak úgy lettek meghatározva, mintha nem lenne spontán aktivitás jelenség az idősorban. A tüzelések kezdete előtti alacsony alapszintre került az alapvonal, ám mégse olyan alulra, mint a minimum, ami a zaj alsó felére került, így nem ad valós képet a nyugalmi állapot átlagos jelintenzitásáról.

Spontán aktív viselkedést nem mutató sejt esetében, ahogy a 16. ábra jobb oldali idősorán is láthatjuk, a 2. legalacsonyabb értéket a medián veszi fel, és ebben az esetben az átlag és a GMM vesznek fel közeli értékeket egymáshoz. Spontán aktivitás hiányában nem várható olyan drasztikus eltérés az alapvonalak értékei között, mint jelenlétében. A GMM továbbra is annak

az alpopulációnak a várható értékét fogja megadni alapvonalként, amely értékek a legnagyobb súllyal szerepelnek a teljes populációban. Így spontán aktivitás hiányában is használható a GM modell. Méréseink azt mutatják, hogy nem spontán aktív sejt esetében a GMM-mel kapott eredmény nincs messze az átlag módszerrel számított alapvonalától.

ROI	M	neg_log_likelihood	ingerlésre adott válasz	spontán aktivitás
ROI30	5	-3,7331	128,91	1
ROI63	3	-3,4461	74,38	1
ROI156	2	-1,9794	89,97	0

3. táblázat: Spontán aktív és spontán aktivitást nem mutató sejtek keresztvalidációval megállapított optimális normális eloszlás komponensei GMM-ben (M), keresztvalidált likelihood érték (neg_log_likelihood), válasz és sp. aktivitás

ROI/módszer	ROI30		ROI63		ROI156	
	alapvonal	reakció	alapvonal	reakció	alapvonal	reakció
GMM	56,78	2,27	23,75	3,13	28,80	3,12
átlag	61,70	2,09	31,68	2,35	28,73	3,13
medián	61,97	2,08	27,40	2,71	28,10	3,20
minimum	54,16	2,38	21,07	3,53	25,16	3,58

4. táblázat: Spontán aktív és spontán aktivitást nem mutató sejtek alapvonalai és reakciói különböző módszerekkel

A 3. és 4. táblázat jól bemutatja a különböző módszerek előnyeit és hátrányait. Nem meglepő módon a minimum módszerrel meghatározott alapvonal a legalacsonyabb mindhárom ROI esetében, azaz a reakció értéke minimum módszerrel a legmagasabb. Ezzel az a gond, hogy a nagyobb reakció csak annak tudható be, hogy az alapvonal a zaj miatt túl alacsonyra került. A valóságban előfordulhat, hogy a sejt igazából nem is reagált érdemileg az ingerlő anyag beadására, viszont a minimum módszer használata miatt mégis olyan eredményeket kapunk, mintha érdemlegesen reagált volna a sejt. Ha beállítanánk egy abszolút reakció értéket (pl. 2), ami fölött reagálnak soroljuk be a sejtet, alatta pedig nem reagálnak, fals értéket kaphatunk a mérésen belül a reagáló és nem reagáló sejtek arányáról.

Spontán aktív sejt esetében az átlag és medián módszer esetén a reakció alacsonyabb, mint a valóságban. Ez annak tudható be, hogy a kilógó értékek miatt mindkét módszer esetén túl magasra kerül az alapvonal (ahogy a 16. ábrán is látható volt), így a reakció alacsonyabb lesz a valóságosnál. A GMM használatával kiszűrődik a spontán aktív hatás, ám nem kapunk a valóságosnál nagyobb reakció értékeket, mint a minimum módszer esetén. A GMM általi klaszterezés alapján reális értékű reakciót kapunk eredményül.

Spontán aktivitás hiányában a minimum módszert félretéve nagyjából azonos nagyságú reakciókról beszélhetünk. Ez várható volt, hiszen a nem spontán aktív idősorokban nem találhatók a spontán aktivitáshoz hasonló nagy kilógó értékek, így a GMM-mel a

kutatócsoportunk által eddig is használt átlag módszerhez hasonló értékeket kapunk. Keresztvalidáció alapján spontán aktivitás hiányában is akkor kapjuk a maximum likelihood (azaz: negatív logaritmizált likelihood) értéket, ha legalább 2 normális eloszlást illeszt a GMM. 1 normális eloszlás várható értéke, azaz egyszerű átlagszámítás a modell alapján nem javasolt.

Tehát a GMM spontán aktív ROI-k esetében sikeresen eltávolítja a spontán aktivitás hatását az idősből anélkül, hogy túl alacsonyra becsülné az alapvonalat. Spontán aktivitás hiányában pedig az alapvonalat azon normális eloszlás várható értékeként határozza meg, ahova az adatpontok a legnagyobb súllyal (valószínűséggel) kerülnek. Ennélfogva ez esetben a GMM az alpopulációkra bontást nem a spontán aktív kilógó értékei alapján végzi el, hanem az idősorban található zaj alapján.

VII. 2. Véletlen erdő eredmények

A véletlen erdő modell megépítéséhez tanuló adatbázisra volt szükség, amelynek elkészítésében a kutatócsoport tagjai voltak segítségemre. Mivel spontán aktív sejtek ugyanúgy jellemzőek a mutáns, mint a vad típusú sejtvonalakra, a tanuló adatbázisok felépítése során nem tettünk különbséget a sejtvonalak idősorai között, mindkét vonalból használtunk fel adatokat. Összesen 4 mérés idősorait vizsgáltuk meg, 4 különböző sejtvonalon (#25, XCL1, 3PC1, #23), amely 525 sornyi adatot jelent. A beadások időpontjáig vizsgáltuk vonaldiagramon az idősorokat, és 2 csoportba, spontán aktív (1) és spontán aktivitást nem mutató (0) csoportba soroltuk a megfigyeléseket. A ROI-k esetében az egyes időpontok fluoreszcencia intenzitás átlagértékei, azaz a sejtek adott időpontbeli Ca^{2+} -koncentrációi jelentik a magyarázóváltozókat. További két magyarázóváltozóként szerepeltettük emellett a ROI-k nyugalmi időszakában mért átlagos Ca^{2+} -szintet, illetve ennek szórását.

Mivel a spontán aktivitásra nincs egzakt definíció, a kutatók között is eltérések álltak fenn, mi számít spontán aktív viselkedésnek. Munkám során a témavezetőm, Dr. Tordai Csongor csoportosításával dolgoztam, de a kutatócsoport vezetője, Dr. Apáti Ágota is készített besorolást 3 idősorra. Érdekes volt összevetni a csoportosításuk közötti különbségeket. A Cramer-féle V együttható értéke 0,8575 lett, ami erős kapcsolatot mutat a két változó között, erős konzisztenciáról beszélhetünk. Azaz érdemben nem tér el a két besorolás.

További nehézséget okozott, hogy a glutamát beadásának időpontja mérésenként változik. Ennek biológiai okai vannak. Nem lehet automatizálni, hogy a glutamátot mindig ugyanakkor adják be, mivel a mikroszkópos mérés kezdetekor a sejtektől függött, hogy mennyit kell várni a nyugalmi állapot beálltára. Bizonyos esetekben ez hamar megtörtént, ekkor relatíve hamar

megtörténhetett az ingerlés, ám volt, hogy a mérés kezdetekor a Ca^{2+} -szint csökkenő trendet mutatott egy ideig. Ilyenkor a kutatók addig vártak, amíg ez a csökkenés megszűnt, és beállt egy körülbelül egyenes vonallá a Ca^{2+} -szint, és csak ezután következhetett a glutamát beadása.

A 4 mérés különböző beadási időpontjai 196, 254, 298 és 318. Mivel a véletlen erdő nem tud se több, se kevesebb változóval dolgozni, mint az eredetileg megadott változószámmal, ezért a modellt a legrövidebb időszakra, 196-ra készítettem el. Ez úgy történt, hogy a hosszabb méréseket 196 hosszúságúra skáláztam, azaz ezeken az idősorokon bizonyos oszlopokat összevontam a szomszédos oszloppal átlagolás segítségével, bizonyosakat pedig önmagukban szerepeltettem. Az összevonási helyeket külön-külön meghatároztam mindhárom idősorra.

Így a modellben összesen $196+2$ magyarázóváltozó szerepelt. Az első 196 változó a 196 mérési időpont, a két másik változó pedig adott sejt átlaga és szórása az egész nyugalmi állapot alatt. Utóbbi két változó szerepeltetése azért fontos, mivel a Ca^{2+} -szint alapvonalak eltérnek egymástól akkor is, hogyha nincs spontán aktivitás. felépített Az adatbázist 80%-20% arányban osztottam fel, azaz 420 elemű lett a tanító és 105 elemű lett a tesztelő adatbázis. Az adatok normalizálásával is próbálkoztam, ekkor csak a 196 időpont normalizált értékei szerepeltek a modellben, mint magyarázó változók. Ez csökkenést eredményezett a keresztvalidált találati arányban, így ennek használatát elvettem. A normalizált adatokra épített véletlen erdő modell AUC értéke külön tesztalmazon 0,92, 105 db ROI-ból összesen 17-et tévesen azonosít be a modell. A végleges modellben így az eredeti mértékegységükben szerepeltettem a változókat, a továbbiakban e modell bemutatása következik. Ezen modell részletes teljesítményét az 5. és 6. táblázatok tartalmazzál.

Mivel a spontán aktivitás nem olyan gyakori, érdekesnek találtam a modell cut-off értékét átállítani az alapbeállítás 50%-áról alacsonyabb értékre. Azaz azt szeretném, hogy ne 50%-os valószínűség felett becsülje a modell spontán aktívna az adott sejt viselkedését, hanem esetleg már egy kisebb spontán aktivitási valószínűség felett is. Ennek az értéknek a meghatározására először is ROC-görbét számítottam. A pozitív osztály, amit pontosan megbecsülni kívánok, az 1-es osztály, azaz a spontán aktivitás. Olyan cut-off értéket akarok tehát megállapítani, amely esetében a szenzitivitás (True Positive Rate, TPR) a lehető legmagasabb, azaz annak az értéke, hogy a modellünk hány százalékát azonosítja a valóságban az eredményváltozó pozitív osztályába tartozó egyedeknek helyesen, a lehető legnagyobb. Eközben az 1 – specificitás (False Positive Rate, FPR) értékét minimalizálni kívánjuk. A specificitás mutatja, hogy a modellünk hány százalékát azonosítja a valóságban az eredményváltozó negatív osztályába tartozó egyedeknek helyesen. A ROC-görbe az 1 – specificitás értékét használja, azaz azt, hogy

a negatív osztályba tartozó egyedeknek hány százaléka lett tévesen a pozitív osztályba sorolva (így ez a pozitív osztály FPR-ének nevezhető). A ROC-görbe x-tengelyén az 1-specificitás, y-tengelyén pedig a szenzitivitás található, és a görbe pontjai a különböző cut-off értékekhez számolt FPR-TPR értékeket adja meg.

A ROC-görbe alatti terület optimális esetben 1, hiszen ekkor $y = 1$, azaz $TPR = 100\%$ (minden pozitív osztályba tartozó egyed helyesen beazonosít a modell) és $x = 0$, azaz $FPR = 0\%$ (egyik negatív osztályba tartozó egyed sem azonosítja pozitív osztályba tartozóként a modell). A ROC-görbe alatti terület a modellem esetében 0,9598, ami közel van az ideális állapothoz. Akkor vagyunk ehhez a legközelebb, ha az ideális (0,1) pont és a ROC-görbén található összes (x,y) pont euklideszi távolságát megmérjük, és a legkisebb távolsághoz tartozó cut-off értéket vesszük. Ennek értéke esetemben 29,75%. Tehát, ha a véletlen erdő modell által becsült valószínűség legalább 29,75%, akkor már spontán aktívnak számít a ROI.

Tesztelő adatbázison a modellem konfúziós mátrixa 50% és 29,75%-os cut-off mellett a következőképpen néz ki:

		Valós spontán aktivitás			
		cut-off = 50%		cut-off = 29,75%	
		0	1	0	1
Becsült	0	84	7	77	1
	1	0	14	7	20

5. táblázat: Véletlen erdő modell konfúziós mátrixai 50%-os és 29,75%-os cut-off mellett

	cut-off = 50%	cut-off = 29,75%
Accuracy	93,30%	92,40%
Precision(1)	100,00%	74,07%
Precision(0)	92,31%	98,72%
Recall(1)	66,67%	95,24%
Recall(0)	100,00%	91,67%

6. táblázat: Véletlen erdő teljesítménymutatói 50%-os és 29,75%-os cut-off mellett. Piros = csökkenés, zöld = növekedés az eredeti mutatóhoz képest.

A 6. táblázatban látható eredeti, cut-off = 50%-os véletlen erdő teljesítménymutatói a recall(1) mutatót leszámítva magas értékek. Azonban mivel a modellünk célja, hogy a recall(1) mutató minél magasabb legyen (tehát annak értéke, hogy a modell által spontán aktívnak prediktált ROI-k között legyen az összes valójában spontán aktív ROI-k minél nagyobb százaléka), a cut-off értéket optimalizáltam, és már 29,75%-os becsült valószínűség felett spontán aktívnak azonosította a modell a sejtet. Az új modell teljesítménymutatóinak az értelmezése a következő: Accuracy/Pontosság = 92,40%, azaz a modell helyesen osztályozta az összes sejt 92,4%-át;

Precision(1)/Precizitás(1) = 74,07%, azaz a modell által spontán aktívnek becsült sejtek közül 74,07% valóban spontán aktív; Precision(0)/Precizitás(0) = 98,72%, azaz a modell által spontán aktív viselkedést nem mutatónak becsült sejtek közül 98,72% valóban nem spontán aktív; Recall(1) = 95,24%, azaz a spontán aktívnek prediktált sejtek között van az összes valójában spontán aktív sejt 95,24%-a; Recall(0) = 91,67%, azaz a nem spontán aktívnek becsült sejtek között van az összes nem spontán aktív sejt 91,67%-a. Azaz valóban nőtt a pozitív osztályba (spontán aktív) tartozó egyedek becsülésének a helyessége, és a csökkenés ellenére is magasak a teljesítménymutatók értékei. Az új 29,75%-os cut-off használatával a modell, ahogy az absztraktban is szerepelt, csak 8-at hibázik a 105 elemű tesztelő adathalmazon. Habár a hibázások száma 1-gyel több, mint az 50%-os cut-off esetében, megéri a csökkenés a találati arányban a recall(1) mutató nagyfokú javulása okán. A ROC-görbe alatti terület (AUC) értéke 0,96.

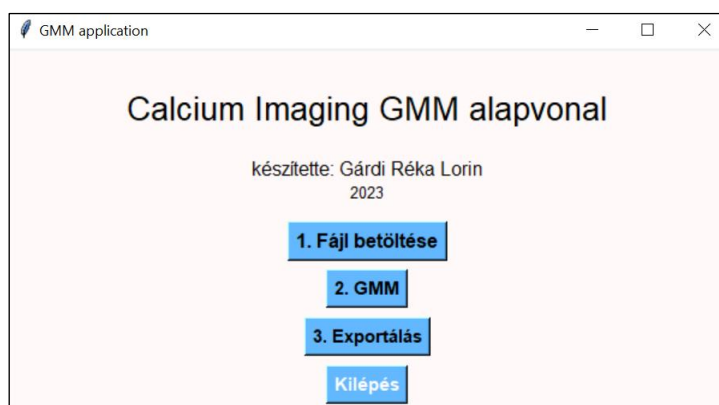
VII. 3. Grafikus felhasználói felület

VII. 3. 1. GMM alkalmazás

A kutatócsoportban dolgozó biológusok és orvosok nem tanultak programozni tanulmányaik során. Emiatt fontosnak tartottam, hogy a Gaussian Mixture és a véletlen erdő modell használatát felhasználóbaráttá tegyem, alkalmazások fejlesztésén keresztül. Célom az egyszerű használat és az automatizáció kialakítása volt.

A Gaussian Mixture modellt Python nyelven készítettem el, így az alkalmazás elkészítéséhez a *tkinter* csomagot használtam. Az alkalmazás python kódja Githubon, a következő linken érhető el: https://github.com/Loreqa/TDK_2023

Az ott található README.md fájlban megtalálhatóak a kód futtatásához szükséges csomagok és verziószámaik, illetve emellett két próba adatsor is található, amellyel a GMM és a véletlen erdő alkalmazás is kipróbálható.



17. ábra: GMM alkalmazás (saját fejlesztés)

A 17. ábrán látható az általam fejlesztett alkalmazás. Az 1. gomb megnyomásával bemenetként egy .csv fájlt kell megadni (első oszlop: időpontok, „Frame” oszlopnévvel; további oszlopok: ROI-k [„ROI0”, „ROI1”, „ROI2”... oszlopnevekkel]; sorok: időpontok; értékek: Ca^{2+} -szint, tehát fluoreszcencia-intenzitás). A fájl elnevezése során a kutatócsoport által követett fájlnev konvenciókat kell követni. A névben az információkat (pl. kísérlet éve, ingerlő anyag típusa) kötőjelekkel kell elválasztani, és az utolsó tagnak kell lennie a számmal megadott beadási időpontnak. Ez azért fontos, mert a program automatikusan a fájlnevből szedi ki a beadás időpontját, és az alapján végzi el a feltöltött mérés idősorának kettéválasztását nyugalmi és ingerlés utáni állapotra. Példának okáért megfelelő név ezek alapján, hogy „xxx-xxx-xxx-300.csv”, ha a 300. időpontban történt az ingerlő anyag beadása. Sikeres betöltés esetén egy felugró ablak értesíti erről a felhasználót.

Ezután a 2. gomb megnyomásával az alkalmazás elvégzi a keresztvalidációt és az alapvonalak kiszámítását GMM módszerrel. Körülbelül 1 perc várakozás után automatikusan megjelenik kimenetként egy adattábla, amely tartalmazza a keresztvalidált normális eloszlások számát (M), a keresztvalidált likelihood értéket (neg_log_likelihood), a GMM által számított alapvonal értéket (gmm_baseline), a válasz értékét (max_value) és a reakció értékét (reaction).

Miután a felhasználó megtekintette a táblát, csukja be, majd a 3. gomb megnyomásával .xlsx formátumban exportálni tudja az előzőleg megtekintett táblát az általa beállított könyvtárba.

VII. 3. 2. Véletlen erdő alkalmazás

A spontán aktivitás detektálására felépített véletlen erdő modell R alapú, így a véletlen erdő alkalmazás megépítéséhez az R shiny csomagját használtam.

Az alkalmazás elérhető a következő linken: https://loreqa.shinyapps.io/sp_akt_veletlen_erdo/

Első lépésként a „.csv fájl tallózás” alatti „Browse” gombbal kell feltölteni egy .csv adatfájlt (első oszlop: Időpont; további oszlopok: ROI; sorok: időpontok és Ca^{2+} -szint (fluoreszcencia-intenzitás)). Az „Eredeti adatfájl” fül alatt megtekinthető a feltöltött adattábla. Az oldalsó panelen lehet megadni, hogy a .csv fájl mivel van elválasztva, tartalmaz-e oszlopneveket, illetve, hogy a fül alatt a teljes adatfájlt, vagy csak az első sorokat szeretné-e megjeleníteni a felhasználó.

Alatta a „Vektor megadása (vesszővel elválasztva)” bemeneti cellába kell beírni vesszővel elválasztva, hogy a program mely oszlopokat vonja össze a tőle jobbra található szomszédjával. Erre a VII. 2. fejezetben is említett ok miatt van szükség a tanuló adatbázisokkal kapcsolatban. A véletlen erdő modellt 196 megfigyeléssel rendelkező tanuló adatbázisra építettem fel (a

további 2 változó már az ebből a 196 változóból számított átlag és szórás). Emiatt tehát a bemenetként megadott adatsor nem lehet rövidebb 196 megfigyelésnél. Ha több nála, akkor a felhasználónak kell megadnia, hogy hol történjen meg az összevonás. Alapértelmezettként azok a helyek vannak megadva, ahol 300 hosszúságú adatsor feltöltése esetén kell összevonni az oszlopokat.

A „Megadott vektor” fül alatt tekinthető meg a felhasználó által megadott vektor, avagy, ha nem ad meg semmit, az alapértelmezettként beállított értékek jelennek meg. Az „Összevont táblázat (modell bemenete)” fül alatt tekinthető meg az adattábla, ami alapján a véletlen erdő modell el fogja végezni a prediktálást, azaz a spontán aktivitás becslését. Az „Eredeti és összevont táblázat vonaldiagram” fül arra szolgál, hogy a felhasználó le tudja ellenőrizni magát és a programot, hogy az ábra alapján sikeres lett-e az összevonás. A fenti ábra tartalmazza az összevont táblázat ROI1 vonaldiagramját, az alsó pedig az eredeti ROI1 vonaldiagramját nyugalmi időszakban. A fenti ábra x-tengelye értelemszerűen 1-196 közötti értékeket vesz fel. Sikeres összevonás esetén látható, hogy alapvetően ugyanaz a két vonaldiagram alakja, csak az alsó ábrán sűrűbbek az értékek.

A „Spontán aktivitás” fül alatt tekinthető meg az adattábla, ami ROI-nként tartalmazza a véletlen erdő modell által becsült spontán aktivitás valószínűségét (`sp_akt_valoszinuseg`) és a becslést, hogy spontán aktív-e a sejt (`becsult_sp_akt`, 1 = spontán aktív, 0 = nem spontán aktív). A véletlen erdő alkalmazásba a VII. 2. fejezetben szereplő 29,75%-os cut-off alapján döntő véletlen erdő modell került a spontán aktivitás becslésére. Alatta egy vektor található, ami a spontán aktív ROI-k mennyiségét és nevét tartalmazza. Az alatt pedig egy interaktív tábla, ahol a felhasználó kiválaszthatja, hogy az eredeti táblából melyik ROI-t szeretné vonaldiagramon ábrázolva látni. A piros függőleges vonal az ingerlő anyag beadásának az időpontját jelzi. Ennek segítségével ellenőrizni tudja a kapott eredményeket, megtekintheti például, hogy néznek ki a spontán aktívnak becsült ROI-k idősorai a teljes időszakon.

Végül a „Spontán aktív táblázat letöltése” gombbal a felhasználó exportálni tudja .csv formátumban a „Spontán aktív” fül alatt található táblázatot. Fontos, hogy a gombot csak azután nyomja meg, hogy a „Spontán aktív” fület egyszer már megnyitotta, hogy legenerálódjon az exportálni kívánt adattábla.

VIII. Összefoglalás

A szkizofrénia súlyos pszichiátriai megbetegedés, egyelőre nem gyógyítható. A kutatócsoporttal mutáns és vad típusú sejtvonalak Ca^{2+} -jeleinek adatgyűjtése és előszűrése után megkezdtem az adatok feldolgozását.

Az ábrák rávilágítottak arra, hogy a spontán aktivitás jelenléte miatt az alapvonal meghatározására az irodalomban eddig használt módszerek nem megfelelőek, torzított eredményeket adnak. A Ca^{2+} -idősorok elemzése hagyományos módszerekkel (idősor-dekompozíció, simítófüggvények) nem vezettek eredményre, szükség volt a spontán aktív hatások elkülönítésére.

Az alapvonal meghatározását GMM módszerrel végeztem el, a spontán aktivitás hatásokat ezáltal klasztereztem. 10-rétegű keresztvalidációval kiszámítottam az egyes sejtekhez tartozó optimális normális eloszlás komponensek számát, majd felépítettem a modellt minden ROI-ra külön-külön. A modellben legnagyobb súllyal szereplő normális eloszlás várható értékét vettem alapvonalnak ROI-nként. Spontán aktív sejtek esetében az átlag és medián módszerek alatti, ám a minimum módszer feletti alapvonal értéket kaptam, ahogy vártam, ott, ahova a kutatótársaim manuálisan is helyezték volna az alapvonalat az ábrák alapján. Spontán aktivitás hiányában pedig az eredeti átlag és medián módszerhez hasonló értékeket kaptam, komoly eltérésről nem lehet beszélni a módszerek között.

Fontos információ a kutatás szempontjából, melyik ROI spontán aktív, így a spontán aktivitás automatizált detektálására véletlen erdő modellt építettem. 1000 döntési fát tartalmaz, az mtry paraméter értéke keresztvalidáció alapján 15. A magyarázó változók a 196 időpont megfigyelései, illetve ezek átlaga és szórása. A cut-off értéke az egységnégyzet és a ROC-görbe közötti legkisebb euklideszi távolságot eredményező 29,75% lett. A modell pontossága tesztelő adathalmazon 92,4%, 21 spontán aktív sejtből 20-at helyesen beazonosít. ROC-görbe alatti területe 0,96.

A modellek használatát alkalmazások fejlesztésével felhasználóbaráttá tettem. A kutatás során használt egyéb programok kimeneti stílusának (pl. ImageJ) és a fájlnevezési konvencióknak megfelelően a mérés során kapott adatokat azonnal fel lehet tölteni az alkalmazásokba, nincs szükség előzetes átalakításra. A GMM alkalmazás használatával nem szükséges a kutatónak egyesével kiszámolnia a reakciókat külön excel fájlokban, hanem a méréseit egymás után feltöltheti és kimenetként egy összefoglaló táblázatban megkapja az alapvonal, a válasz és a reakció értékét is.

A véletlen erdő alkalmazásba hasonlóan feltölthető a mérés adatfájlja, emellett egy vektort kell még megadni az összevonáshoz, hogy a megfigyelések 196 hosszúságúak legyenek. Az alkalmazás ezután megbecsüli a ROI-nként a spontán aktivitás jelenlétét. Interaktív ábra segítségével leellenőrizheti vonaldiagramon a kutató, mely ROI idősort sorolta spontán aktívnak, melyet nem a modell.

Hivatkozások

- Apáti, Á., Berez, T., Sarkadi, B. (2016). *Calcium signaling in human pluripotent stem cells*. Cell Calcium, 59. kötet, 2–3 kiadás, pp. 117–123. <https://doi.org/10.1016/j.ceca.2016.01.005>
- Ballard, D. H. (2000). *An Introduction to Natural Computation*. MIT Press, pp. 189. ISBN 978-0-262-02420-4.
- Barbato, A. (1998). *Schizophrenia and Public Health*. Genf, Svájc: World Health Organization
- Breiman, L. (1984). *Classification and Regression Trees*. Routledge. <https://doi.org/10.1201/9781315139470>
- Breiman, L. (2001). *Random Forests*. Machine Learning 45. kötet, pp. 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chong, H. Y., Teoh, S. L., Wu, D. B., Kotirum, S., Chiou, C. F., Chaiyakunapruk, N. (2016). *Global economic burden of schizophrenia: a systematic review*. Neuropsychiatr Dis Treat, 12. kötet, pp. 357–373. doi: 10.2147/NDT.S96649.
- Clapham D. E. (2007). *Calcium signaling*. Cell, 131(6), pp. 1047–1058. <https://doi.org/10.1016/j.cell.2007.11.028>
- Cryer, J., Chan, K. S. (2008). *Time Series Analysis with Applications in R (second edition)*. Springer.
- Csáji, B. Cs. (2001). *Approximation with Artificial Neural Networks*. Magyarország, Eötvös Loránd Egyetem Természettudományi Kar.
- Evans, M., Kaufman, M. (1981). *Establishment in culture of pluripotential cells from mouse embryos*. Nature, 292. kötet, pp. 154–156. <https://doi.org/10.1038/292154a0>
- Galizia, C., Menzel, R., Hölldobler, B. (1999). *Optical Imaging of Odor-Evoked Glomerular Activity Patterns in the Antennal Lobes of the Ant Camponotus rufipes*. Naturwissenschaften, 86. kötet, pp. 533–537 <https://doi.org/10.1007/s001140050669>
- Gover, T. D., Kao, J. P., Weinreich, D. (2003). *Calcium signaling in single peripheral sensory nerve terminals*. The Journal of neuroscience: the official journal of the Society for Neuroscience, 23(12), pp. 4793–4797. <https://doi.org/10.1523/JNEUROSCI.23-12-04793.2003>

- Hathy, E., Szabó, E., Varga, N. et al. (2020). *Investigation of de novo mutations in a schizophrenia case-parent trio by induced pluripotent stem cell-based in vitro disease modeling: convergence of schizophrenia- and autism-related cellular phenotypes*. Stem Cell Res Ther 11. kötet, 504. számú cikk <https://doi.org/10.1186/s13287-020-01980-5>
- Hodrick, R. J., Prescott, E. C. (1997). *Postwar U.S. Business Cycles: An Empirical Investigation*. Journal of Money, Credit, and Banking, 29(1). pp. 1–16. doi:10.2307/2953682
- Hong, S.Z., Mesik, L., Grossman, C.D. et al. (2022). *Norepinephrine potentiates and serotonin depresses visual cortical responses by transforming eligibility traces*. Nat Commun 13. kötet, 3202. számú cikk. <https://doi.org/10.1038/s41467-022-30827-1>
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An introduction to statistical learning (2nd ed.)*. Springer.
- Jennings, J.H., Kim, C.K., Marshel, J.H. et al. (2019). *Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour*. Nature 565. kötet, pp. 645–649. <https://doi.org/10.1038/s41586-018-0866-8>
- van Os, J., Kapur, S. (2009). *Schizophrenia*. Lancet. 374. kötet, 9690. kiadás, pp. 635–645. [https://doi.org/10.1016/S0140-6736\(09\)60995-8](https://doi.org/10.1016/S0140-6736(09)60995-8)
- Kahn, R. S., Sommer, I. E., Murray, R. M. et al., (2015). *Schizophrenia*. Nature reviews. Disease primers, 1, 15067. <https://doi.org/10.1038/nrdp.2015.67>
- Kratsios, A., Bilokopytov, E. (2020). *Non-Euclidean Universal Approximation*. Advances in Neural Information Processing Systems. Curran Associates, 33. kötet.
- LeDell, E., Petersen, M., van der Laan, M. (2015). *Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates*. Electronic journal of statistics, 9(1), 1583.
- McLachlan, G. J., Rathnayake, S. (2014). *On the number of components in a Gaussian mixture model*. WIREs Data Mining and Knowledge Discovery, 4. kötet. doi: 10.1002/widm.1135
- McSweeney, L. A. (2006). *Comparison of periodogram tests*. Journal of Statistical Computation and Simulation, 76(4), pp. 357–369.

- Moore, A. M., Coolen, L. M., Lehman, M. N., (2022). *In vivo imaging of the GnRH pulse generator reveals a temporal order of neuronal activation and synchronization during each pulse*. PNAS, 119(6) <https://doi.org/10.1073/pnas.2117767119>
- Newman, Z. L., Bakshinskaya, D., Schultz, R. et al. (2022). *Determinants of synapse diversity revealed by super-resolution quantal transmission and active zone imaging*. Nat Commun, 13. kötet, 229. számú cikk. <https://doi.org/10.1038/s41467-021-27815-2>
- Reynolds, D. (2009). *Gaussian Mixture Models*. Encyclopedia of Biometrics, pp. 659-663. https://doi.org/10.1007/978-0-387-73003-5_196
- Robbins, M., Christensen, C. N., Kaminski, C. F., Zlatic, M. (2021). *Calcium imaging analysis - how far have we come?*. F1000Research, 10. kötet, 258. számú cikk. <https://doi.org/10.12688/f1000research.51755.2>
- Szentágothai, J., Réthelyi, M. (2006). *Funkcionális anatómia I*. Budapest, Magyarország: Medicina Kiadó.
- Takahashi, K., & Yamanaka, S. (2006). *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors*. Cell, 126(4), pp. 663–676. <https://doi.org/10.1016/j.cell.2006.07.024>
- Tringer, L. (2010.) *A Pszichiátria tankönyve*. Budapest, Magyarország: Semmelweis Kiadó
- Uddin L. Q. (2020). *Bring the Noise: Reconceptualizing Spontaneous Neural Activity*. Trends in cognitive sciences, 24(9), pp. 734–746. <https://doi.org/10.1016/j.tics.2020.06.003>
- Vos, T., Barber, RM., Bell, B., et al. (Global Burden of Disease Study 2013 Collaborators) (2015). *Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013*. Lancet (London, Anglia), 386(9995), pp. 743–800. [https://doi.org/10.1016/S0140-6736\(15\)60692-4](https://doi.org/10.1016/S0140-6736(15)60692-4)
- Yu, G. (2012). *Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity*. IEEE Transactions on Image Processing, 21(5), pp. 2481–2499. arXiv:1006.3056