

---

# Parkinsonin taudin tunnistaminen elektroenkefalogrammista koneoppimisteknologian avulla

---

Pro gradu -tutkielma  
Turun yliopisto  
Tietotekniikan laitos  
Tietojenkäsittelytiede  
2022  
Ilkka Suuronen

Tarkastajat:  
Antti Airola  
Tapio Pahikkala  
Henry Railo

TURUN YLIOPISTO

Tietotekniikan laitos

ILKKA SUURONEN: Parkinsonin taudin tunnistaminen elektroenkefalogrammista koneoppimisteknologian avulla

Pro gradu -tutkielma, 66 s.

Tietojenkäsittelytiede

Kesäkuu 2022

---

Tässä tutkielmassa perehdytään koneoppimisteknologian käyttöön Parkinsonin tautia sairastavien ja terveiden koehenkilöiden EEG-tallenteiden erottamisessa toisistaan. EEG:n käyttö Parkinsonin taudin biomarkerina on herättänyt kiinnostusta, ja aiemmassa koneoppimisteknologiaa käyttävässä tutkimuksessa on saavutettu lupaavia tuloksia. Aiemmassa tutkimuksessa ei ole kuitenkaan tutkittu systemaattisesti EEG:n mittaamiseen käytettävien elektrodien lukumäärän vaikutusta luokittelutarkkuuteen, joka on tämän tutkielman yhteydessä toteutetun koneoppimis-pohjaisen EEG-analyysin ensisijainen tutkimuskysymys. Analyysin aineisto koostuu kolmella eri yliopistolla (Iowan yliopisto, New Mexicon yliopisto ja Turun yliopisto) kerätyistä, yhdistetyistä EEG-aineistoista. Aineistot on esikäsitelty PREP-esiprosessointiputken avulla, ja piirteenekstraktointiin on käytetty tyypillisten EEG-analyysin mukaisten taajuuskaistojen (delta, theta, alpha1, alpha2, beta) näyte-entropia -metriikoita. Aineiston luokitteluun on käytetty logistista regressiomallia. Elektrodien lukumäärän vaikutusta mallin saavuttamaan luokittelutarkkuuteen on tutkittu käyttämällä budjetoitua ja ryhmäperustaista, ahnetta eteenpäinhakualgoritmia piirteenvälintaan. Keskeisenä havaintona huomattiin, että luokittelu onnistuu kymmenellä elektrodilla lähes yhtä hyvällä tarkkuudella (0.72) kuin käyttämällä täyttä elektrodivalikoimaa. Toissijaisesti huomattiin, että luokittelu käyttäen koehenkilöiden silmät auki mitattua EEG:tä onnistuu merkittävästi paremmin kuin käyttäen silmät kiinni mitattua EEG:tä. Lisäksi havaittiin, että elektrodien sijainnilla ei vaikuta olevan erityisen suurta merkitystä. Tämän tutkielman tulosten valossa voi olla aiheellista jatkaa tutkimuksia pienillä elektrodivalikoimilla.

Asiasanat: Parkinsonin tauti, elektroenkefalogrammi, koneoppiminen, näyte-entropia, piirteenvälinta, ahne hakualgoritmi

UNIVERSITY OF TURKU  
Department of Computing

ILKKA SUURONEN: Parkinsonin taudin tunnistaminen elektroenkefalogrammista koneoppimisteknologian avulla

Master of Science Thesis, 66 p.  
Computer Science  
June 2022

---

This thesis addresses the use of machine learning technology for detecting Parkinson's disease from EEG data. The use of EEG as a biomarker for Parkinson's disease has been drawing interest, and previous studies on machine learning based Parkinson's disease detection from EEG have yielded promising results. The effect of the number of electrodes used for recording the EEG on the classification accuracy of the machine learning model has not, however, been systematically studied before, and it is the primary research question of the machine learning based EEG-analysis conducted as an empirical part of this thesis. The material of the analysis consisted of three separate sets of EEG data collected on three sites (University of Iowa, University of New Mexico and University of Turku) that were pooled for the analysis. The data sets were preprocessed using the PREP preprocessing pipeline, and the features extraction was based on sample entropy metrics of the traditional frequency bands (delta, theta, alpha1, alpha2, beta) of EEG analysis. A logistic regression model was employed for classifying the data. The effect of the number of electrodes on the classification accuracy of the model was studied by employing a budget based, grouped, greedy forward search algorithm for feature selection. Remarkably, it was found that the classification accuracy when using only ten electrodes (0.72) is comparable to the classification accuracy when using a full selection of electrodes. As a secondary finding, it was found that classification accuracy on the EEG data collected with the subjects' eyes open is notably higher than the accuracy on the data collected with the subjects' eyes closed. Additionally it was found that the location of the electrodes does not seem to make a big difference on the classification accuracy. The results of this study motivates further study of using machine learning technology for Parkinson's disease detection using small subsets of EEG electrodes.

Keywords: Parkinson's Disease, electroencephalogram, machine learning, sample entropy, feature selection, greedy search algorithm

# Sisällys

<b>1</b>	<b>Johdanto</b>	<b>1</b>
1.1	Lyhyesti Parkinsonin taudista . . . . .	1
1.2	Ihmisaivojen rakenteesta ja toiminnasta . . . . .	3
1.3	Elektroenkefalogrammi . . . . .	5
<b>2</b>	<b>Analyysin metodologia</b>	<b>11</b>
2.1	PREP-prosessointiputki . . . . .	12
2.2	Digitaalisen suodatinteknologian menetelmistä . . . . .	13
2.3	Performanssimetriikat . . . . .	17
2.3.1	Luokittelutarkkuus . . . . .	17
2.3.2	ROC-käyrän alapuolinen pinta-ala (AUC) . . . . .	17
2.4	Epälineaariset systeemit ja entropia . . . . .	20
2.5	Out-of-sample -evaluointi ja ristiinvalidointialgoritmit . . . . .	22
2.5.1	Sisäkkäinen ristiinvalidointi . . . . .	25
2.6	Logistinen regressiomalli . . . . .	26
2.6.1	Regularisointi . . . . .	28
2.7	Ahne joukkoperustainen piirteevalinta-algoritmi . . . . .	30
2.7.1	Piirteevalinnan tavoitteista . . . . .	30
2.7.2	Piirteevalinta joukkopohjaisella hakualgoritmilla . . . . .	31

<b>3 Kirjallisuuskatsaus</b>	<b>36</b>
3.1 Parkinsonin taudin koneoppimispohjainen tunnistaminen . . . . .	36
3.2 Lievä kognitiivinen heikentyminen ja EEG . . . . .	47
3.3 Kirjallisuuden yhteenveto ja keskustelua . . . . .	48
<b>4 Analyysin aineisto, vaiheet ja tulokset</b>	<b>53</b>
4.1 Analyysin tulokset . . . . .	58
4.2 Keskustelua analyysin tuloksista . . . . .	64
<b>Lähdeluettelo</b>	<b>67</b>

# Kuvat

2.1	Erityyppisten alipäästösuodatinten taajuusvasteita. X-akselilla taajuus radiaaneina / sekuntia. Y-akselilla suodattimen vahvistus desibeleinä. . . . .	16
2.2	ROC-käyrä generoidulla datalla sovitetusta logistisesta regressiomallista. Tässä ROC-käyrän alle jäävän alueen suuruus (AUC) on 0.78. .	19
4.1	Vuokaavio, joka esittää piirteengenrointiprosessin vaiheita aineiston mielivaltaiselle tapausesimerkille. . . . .	56
4.2	EO-aineiston luokittelutarkkuuden ja AUC-estimaattien kehittymistä elektrodien lukumäärän funktiona kuvaava graafi. . . . .	59
4.3	EC-aineiston luokittelutarkkuuden ja AUC-estimaattien kehittymistä elektrodien lukumäärän funktiona kuvaava graafi. . . . .	60
4.4	Lämpökartta, joka yhdistää kunkin elektrodin sijainnin ja sen normalisoidun valintakertojen lukumäärän. Valinnat budjeteilla 5, 10, 20 ja 30 elektrodia. . . . .	62

# Taulukot

3.1	Yhteenvedot käsitellyistä EEG-tutkimuksista sisältäen otoskoon ja aineiston jakauman, aineiston tyyppin (EEG/MEG), piirteenekstraktointialgoritmin, luokittelualgoritmin ja luokittelijan suorituskykyestimaatit. . . . .	49
4.1	Turun yliopistolla kerätyn aineiston kuvaus. . . . .	54
4.2	New Mexican yliopistolla kerätyn aineiston kuvaus. . . . .	54
4.3	Iowan yliopistolla kerätyn EO-aineiston kuvaus. . . . .	54
4.4	Iowan yliopistolla kerätyn EC-aineiston kuvaus. . . . .	55
4.5	Keskiarvoistetut testiaineiston luokittelutarkkuusarviot ja AUC-arviot optimoiduille ja satunnaisille yhden, viiden ja kymmenen kanavan valikoimille . . . . .	64

# 1 Johdanto

## 1.1 Lyhyesti Parkinsonin taudista

Parkinsonin tauti on yleinen, etenevä neurodegeneratiivinen sairaus, johon liittyy bradykinesian (hidasliikkeisyyden), lepovapinan ja lihasjäykkyyden kaltaisia klassisia motorisia oireita, sekä vaihtelevia ei-motorisia oireita<sup>1</sup>. Parkinsonin taudin kliininen diagnosointi perustuu ns. parkinsonismin toteamiseen tunnistamalla oireeksi bradykinesian ja vähintään yhden muun mainituista motorisista oireista, jonka jälkeen diagnoosi voidaan vahvistaa Parkinsonin taudiksi riippuen diagnoosia puoltavien tai vastustavien tekijöiden olemassaolosta. Esimerkkinä puoltavasta tekijästä voidaan mainita positiivinen hoitovaste levodopälääkitykselle, siinä missä hoitovasteen puuttuminen tai motoristen oireiden rajoittuminen alavartaloon voivat olla diagnoosia vastustavia tekijöitä. Viime kädessä diagnoosi voidaan varmistaa oikeaksi vain ruumiinavauksen yhteydessä tehdyllä havainnolla keskiaivojen mustatumakkeen hermosolujen tuhoutumisesta, sekä Lewyn kappaleiden<sup>2</sup> muodostumisesta. Vertaamalla kliinisiä diagnooseja ruumiinavauksissa havaittaviin muutoksiin aivois-

---

<sup>1</sup>Parkinsonin taudin ei-motorisiin oireisiin lukeutuu muun muassa uneen, mielialaan, kognitioon sekä autonomiseen hermostoon ja aistien toimintaan liittyviä häiriöitä. Nämä ei-motoriset oireet ovat varsin yleisiä ja voivat heikentää elämänlaatua huomattavasti, mutta jäävät usein vähemmälle huomiolle kuin taudin motoriset oireet.[1]

<sup>2</sup>Lewyn kappaleet ovat tiettyjen neurologisten sairauksien yhteydessä hermosolujen sisälle muodostuvia proteiinikertymiä. ks esim. [2]

sa on arvioitu, että kliinisten diagnoosien tarkkuus on n. 75% – 90%. [3] Kliinisten diagnoosien keskimääräiseksi tarkkuudeksi on arvioitu 81% verrattuna ruumiina-vauspohjaiseen diagnoosiin, ja taudin edistymisen on raportoitu vaikuttavan kliinisen diagnoosin onnistumistodennäköisyyteen. Vähemmän kuin viisi vuotta kestäneen taudin tapauksessa kliinisen diagnoosin onnistumisprosentiksi on arvioitu vain 58%. [4] Etenkin varhaisen diagnoosin tueksi voi olla tarpeellista tunnistaa soveltuvia ns. biomarkereita, mitattavia suureita tai muita vastaavia objektiivisesti havainnoitavia ominaisuuksia jotka indikoivat biologisten prosessien normaalia tai patologista toimintaa. Biomarkerien käyttö Parkinsonin taudin tutkimuksessa voisi paitsi tukea luotettavan kliinisen diagnoosin tekemistä, myös auttaa seuraamaan tai ennustaa taudin etenemistä tai auttaa suunnittelemaan yksilöllistä hoitoa ja seuraamaan hoidon tehoamista. [5] Parkinsonin taudin biomarkereiksi on käytetty tai ehdotettu ainakin puheääntä, askellusta, käsialaa ja aivosähkökäyrää eli elektroencefalogrammia (engl. electroencephalogram, lyh. EEG). Puheäänen käyttö biomarkerina perustuu siihen, että Parkinsonin tautia sairastavan puheääni on normaalia monotonisempi, paikoin epäsujuva ja erityisesti tiettyjen konsonanttien ääntäminen on epätarkempaa. Askelluksen muutokset näkyvät huojuntavana askeleena hitaasti kävellessä, kävelyn jähmettymisenä ja lyhyinä askelina. Käsialanäytteestä puolestaan tarkkaillaan kinemaattisia piirteitä, käsialan epälineaarista dynamiikkaa ja pikselitiheyttä. EEG:ssä on havaittu muutoksia useimpien aivolohkojen alueilla (otsalohko, keskilohko, päälakilohko, takaraivolohko). [6]

Parkinsonin taudin nimi periytyy englantilaiselta monitietelijältä James Parkinsonilta (1755 - 1824), joka vuonna 1817 kirjoitti tautia käsittelevän tutkielman *An Essay on the Shaking Palsy*, kuvaillen kuuden potilaan oireita. Parkinsonin taudin oireisiin sopivia kuvauksia on dokumentoitu eri puolilla maailmaa kautta historian ja jopa ennen ajanlaskun alkua, joten kyse on ollut ainakin jossakin määrin tunnetusta sairaudesta jo ennen Parkinsonin tutkielman julkaisua, mutta Parkinsonin

tutkielmaa voinee pitää ensimmäisenä tunnettuna tautia koskevana kokonaisvaltaisena esityksenä. Parkinson tunnisti muun muassa taudin hitaan etenemisen, sen varhaisiksi oireiksi heikkouden tunteen ja käsien vapinan sekä etenevän vaikeuden ylläpitää ryhtiä. Pidemmälle edenneen sairauden oireina hän kuvaili muun muassa potilaidensa vaikeuksia käyttää haarukkaa syömiseen, suolentoiminnan hidastumisen sekä puhekyvyn asteittaisen menettämisen. 1900-luvun aikana Parkinsonin tautia tutkittiin systemaattisesti, ja mainittavina edistysaskelina vuonna 1958 havaittiin taudin motoristen oireiden johtuvan dopamiinivajauksesta mustatumakkeessa, vuonna 1961 kehitettiin levodopaan perustuva tehokas hoitomuoto taudin oireille, ja vuonna 1987 luotiin UPDRS-arviointikriteeristö taudin etenemiselle, jota vuonna 2007 täydennettiin johtaen MDS-UPDRS-kriteeristön<sup>3</sup> käyttöönottoon.[8]

## 1.2 Ihmisaivojen rakenteesta ja toiminnasta

Ihmisen aivot koostuvat hermokudoksesta joka jaetaan harmaaseen ja valkoiseen aineeseen. Hermokudos muodostuu gliasoluista ja hyvin suuressa määrin yhteenliittyneistä hermosoluista, joista suurin osa on afferenttien (ts. tuovien) ja efferenttien (ts. vievien) hermosolujen välisiä ns. välittäviä hermosoluja, jotka yhdistävät aivojen sensorisia ja motorisia alueita. Aivot rakentuvat isoaivoista (lat. cerebrum), aivorungosta ja pikkuaivoista (lat. cerebellum). Suurin näistä, isoaiivot, rakentuu puolestaan vasemmasta ja oikeasta hemisfääristä eli aivopuoliskosta, sekä väliaivoista (lat. diencephalon). Isoaivojen rakenne koostuu aivopoimuista (lat. gyrus, mon. gyri) sekä niiden välissä olevista uurteista (lat. sulcus, mon. sulci), joiden perusteella aivopuoliskot voidaan jakaa neljään funktionaalisesti erikoistuneeseen aivolohkoon. Keskiurre (engl. central sulcus, myös Rolandic fissure) erottaa etulohkon (engl.

---

<sup>3</sup>MDS-UPDRS täydensi UPDRS-kriteeristöä erityisesti ottamalla huomioon entistä paremmin Parkinsonin taudin muut kuin motoriset oireet, sekä sisältämällä yksityiskohtaiset ohjeet arvioinnin tekemistä varten.[7]

frontal lobe) ja päälakilohkon (engl. parietal lobe) toisistaan, lateraaliuurre (engl. lateral sulcus, myös Sylvian fissure) erottaa etulohkon ja päälakilohkon ohimolohkosta (engl. temporal lobe), ja parieto-okkipitaaliuurre erottaa päälakilohkon ja ohimolohkon takaraivolohkosta (engl. occipital lobe). Etulohkot liittyvät muun muassa motoriseen toimintaan, kieleen ja käytökseen. Ohimolohkot liittyvät ääniärsykkeiden vastaanottamiseen ja käsittelyyn, kielen ymmärrykseen sekä muistin ja tunteiden toimintaan. Päälakilohkot liittyvät tuntoaistimusten käsittelyyn, avaruudellisen hahmotuskyvyn toimintaan sekä lukemisen ja laskemisen kaltaisiin tehtäviin. Takaraivolohkot puolestaan liittyvät ensisijaisesti näköhavaintojen vastaanottamiseen ja käsittelyyn. Aivolohkojen toiminta on osin erilaista aivopuoliskojen välillä: esimerkiksi avaruudellinen hahmotuskyky kuuluu erityisesti oikean aivopuoliskon päälakilohkon tehtäviin, siinä missä vasemman aivopuoliskon tehtäviin kuuluvat lukeminen ja laskeminen. Erikoistuneiden aivoalueiden välillä on suuri määrä valkoisen aineen muodostamia yhteyksiä, jotka mahdollistavat viestien välittymisen alueiden välillä nopeasti ja tehokkaasti.[9] Viestien välittyminen hermosolujen välillä perustuu hermosolujen viejähaarakkeiden eli aksonien johtamiin sähköisiin impulsseihin. Hermostulun pinnalla on kanavia, jotka päästävät läpi tiettyntyyppisiä ioneja<sup>4</sup>, estäen pääsyn toisilta. Näiden kanavien avulla hermostolu ylläpitää epätasapainotilaa solun sisä- ja ulkopuolen ionipitoisuuksien välillä, jolloin solun sisäpuolella on negatiivinen jännite (n. -70 mV) suhteessa sen ulkopuoleen. Kun hermostolu depolarisoituu - siis sen negatiivinen jännite suhteessa sen ulkopuoleen purkautuu - sen pinnan kanavat avautuvat sallien positiivisesti varautuneiden ionien pääsyn soluun, mikä puolestaan aiheuttaa solun toimintapotentiaalin (engl. action potential) laukeamisen. Tämä ilmiö propagoituu hermostulun viejähaaraketta pitkin, saavuttaen lopuksi viereisen hermostulun tuojahaarakkeen, jolloin impulssi siirtyy synaptisen yhteyden

---

<sup>4</sup>Keskeisimmät hermostolujen toimintaan liittyvät ionit ovat positiivisesti varautuneet kalium- ( $K^+$ ) ja natrium-ionit. ( $Na^+$ )

välityksellä seuraavaan hermosoluun.[10]

Aivopuoliskojen lohkot ovat tämän tutkielman kannalta relevanttia taustatietoa sen vuoksi, että elektroenkefalogrammin nauhoittamiseen käytettävät elektrodit yksilöidään niiden sijainnin perusteella suhteessa tähän anatomiseen malliin. Tieto siitä, millä alueella Parkinsonin taudin tunnistamisen kannalta hyödyllisimmät elektrodit sijaitsevat voi näin tarjota näkökulmia paitsi kustannustehokkaista EEG:n nauhoitusmenetelmisiin, myös Parkinsonin taudin aiheuttamiin muutoksiin ihmisaivoissa.

### 1.3 Elektroenkefalogrammi

Elektroenkefalogrammi (jatkossa EEG) on menetelmä aivojen sähköisen toiminnan monitoroimiseen ja tallettamiseen päänahkaan kiinnitettävillä elektrodeilla, joiden tarkoituksena on vastaanottaa hyvin matalan jännitteen (n. 10 - 100  $\mu V$ ) suhteellisen matalan taajuuden oskillaatiota (n. 1-30 Hz). Perinteinen EEG talletettiin piirtämällä kunkin elektrodin signaali paperille reaaliajassa; nykyään elektrodien vastaanottama signaali talletetaan digitaalisilla laitteilla, ja kootaan visuaaliseen muotoon jälkikäteen. EEG tyypillisesti mitataan lepääviltä tai nukkuvilta testihenkilöiltä, mutta tilanne- ja tutkimusasetelmakohtaisesti testihenkilöille voidaan myös tuottaa erilaisia ärsykeitä, joiden vastetta EEG:stä tutkitaan. [11] EEG:n avulla voidaan havainnoida vain osaa kaikesta aivojen sähköisestä aktiivisuudesta, toisaalta elektrodit vastaanottavat myös muista fysiologisista lähteistä<sup>5</sup> sekä ympäristöstä peräisin olevaa signaalia. EEG ei varsinaisesti perustu hermosolujen toimintapotentiaalin, vaan sen yhteydessä esiintyvän postsynaptisen potentiaalin monitoroimiseen. EEG:n merkittävinä etuina voidaan pitää sen kustannustehokkuutta ja ei-invasiivisuutta tarkasteltaessa aivojen toimintaa ja sen häiriöitä. [12]

---

<sup>5</sup>Esim. sydämen toimintaan sekä silmien ja lihasten liikkeisiin liittyvä sähköinen aktiivisuus.

EEG-tallenteen nauhoittamista varten voidaan sijoittaa vaihteleva määrä elektrodia erilaisten järjestelmien mukanaan. 1958 esitellyssä kansainvälisessä 10-20 -järjestelmässä esimerkiksi elektrodit sijoitetaan kallon anatomian määrittämien merkkipaalojen mukaan siten, että niiden etäisyydet toisistaan ovat 10% tai 20% koko pääläen pitkittäisestä tai poikittaisesta pituudesta. 10-20 -järjestelmän kaltaisen standardin käyttäminen on mahdollistanut tallennusmenettelyn toistamisen eri koehenkilöillä ja laboratorioden välillä. Elektrodien sijaintien määrittäminen suhteellisina sijainteina absoluuttisten sijaan mahdollistaa muun muassa johdonmukaisen sijoittelun esimerkiksi pitkittäistutkimuksissa, joissa nuoren koehenkilön pää saattaa vuosien aikana kasvaa. 10-20 -järjestelmä käyttää 21 elektrodia, joista parittomalla luvulla numeroidut sijaitsevat pään vasemmalla puolella, ja parillisella luvulla numeroidut pään oikealla puolella. Elektrodin sijaintia aivokuoren anatomian suhteen ilmaistaan yleensä kirjainyhdistelmällä, joka ilmoittaa minkä anatomisen alueen kohdalla elektrodi sijaitsee.<sup>6</sup> Käytännössä monissa tilanteissa käytetään suurempaa määrää elektrodia kuin 10-20 -järjestelmässä, sekä toisinaan myös pienempää määrää, kuten esimerkiksi pienten lasten EEG-tallenteen nauhoittamiseen. Aivojen neurofysiologinen aktiivisuus tuottaa signaalia jonka taajuus voi olla jopa 2000 Hz, mutta EEG-tutkimuksessa ollaan harvoin kiinnostuneita yli 50 Hz:n taajuuksista. Usein EEG-tallenteen nauhoitustilanteessa käytetään jonkin rajataajuuden (esim. 35 Hz tai 70 Hz) mukaan määritettyä alipäästösuodatinta muun muassa suhteelli-

---

<sup>6</sup>Aivopuoliskojen anatomisten alueiden tunnistena käytetään yleensä alueen englannin- tai latinankielisen nimen ensimmäistä kirjainta: etulohkoja merkitään "F", ohimolohkoja "T", päälakilohkoja "P" ja takaraivolohkoja "O". Näiden lisäksi pääläen keskellä sijaitsevia elektrodia voidaan merkitä kirjaimella "C". Kahden eri alueen välillä sijaitsevat elektrodit yksilöidään usein molempien alueiden tunnisteilla. Esimerkiksi pään vasemmalla puolella päälakilohkon ja takaraivolohkon välillä sijaitseva elektrodi voisi olla tunnisteeltaan "PO7". Tunnisteiden käyttö näyttää vaihtelevan jonkin verran tapauskohtaisesti, eikä pelkän tunnisteiden perusteella voi välttämättä tietää tarkasti elektrodin sijaintia.

sen korkeataajuuksisten häiriösignaalien suodattamiseksi (esim. lihasten aktiivisuuden tai sähkölinjan aiheuttamat artefaktit). Vastaavasti hyvin matalan taajuuden komponenttien (esim. hikoilun aiheuttamat matalan taajuuden artefaktit) vaimentamiseen voidaan käyttää ylipäästösuodatinta (rajataajuuksina esim. 0.1 Hz tai 1 Hz).[13] Mainitun 10-20 -järjestelmän lisäksi voidaan käyttää huomattavasti lukuisempia elektrodeja sisältäviä järjestelmiä: esimerkiksi tämän tutkielman kirjoittamisen yhteydessä toteutettu analyysi hyödynsi 59:ää elektrodia vastaavaa kanavaa.<sup>7</sup>

EEG-analyysissä tarkastellaan usein tiettyjen taajuuskaistojen rytmistä aktiivisuutta, josta etsitään poikkeavuuksia. Taajuuskaistalla tarkoitetaan tietyn parametrein rajattua osuutta koko EEG:n taajuusspektristä. EEG-analyysin kannalta keskeisimpiä taajuuskaistoja ovat delta-, theta-, alfa- ja betakaistat joita vastaavat taajuusalueet ovat (edellä annetussa järjestyksessä) 0-4 Hz, 4-8 Hz, 8-13 Hz ja 13-40 Hz. [14] Käytännössä taajuuskaistojen määritelmien on huomattu vaihtelevan tutkimuskirjallisuudessa huomattavasti lähteestä riippuen. Siinä missä theta- ja alfakaistojen määritelmistä on selkeämpi konsensus, vaihtelee deltakaistan alaraja kirjallisuudessa välillä 0 - 2 Hz ja yläraja välillä 3.5 - 6 Hz. Betakaistan alaraja puolestaan vaihtelee kirjallisuudessa välillä 12 - 15 Hz ja yläraja välillä 20 - 50 Hz.[15] Taajuuskaistojen määritelmien epäjohdonmukaisuus voi ilmeisellä tavalla vaikeuttaa tutkimusraporttien tulosten vertailua. Yksiselitteisyyden vuoksi tässä tutkielmassa käytän oletusarvoisesti edellä annettuja määritelmiä EEG:n spektrianalyysin kannalta keskeisistä taajuuskaistoista, ja sikäli kun esimerkiksi referoitavassa tutkimuskirjallisuudessa poiketaan merkittävästi näistä määritelmistä, esittelen eksplisiittisesti käytetyt vaihtoehtoiset määritelmät. Alfakaista toisinaan jaetaan ala- ja ylätaajuuksiin joiden välisenä rajataajuutena toimii usein 10 Hz.

EEG:n taajuuskaistojen rytmisen aktiivisuuden tulkinta on kontekstisidonnaista:

---

<sup>7</sup>Luvussa 4 kerrotaan tarkemmin elektrodien lukumäärästä, sekä eri laboratorioissa kerättyjen aineistojen yhdistämisen yhteydessä tehdystä ylimääräisten elektrodien karsimisesta.

tietyä taajuuskaistan aktiivisuus voi olla normaalia tai patologista riippuen (mm.) tutkimushenkilön vireystilasta ja iästä. Valvetilassa nauhoitetussa, normaalina pidetyssä EEG:ssä hallitsee niin kutsuttu alfarytmi, joka sijoittuu symmetrisesti takaraivolohkoille, ja vaimenee silmien ollessa auki. Betakaitan aktiivisuutta esiintyy niin ikään normaalissa EEG:ssä, ja se lisääntyy uneliaassa tilassa, kevyessä unessa tai mentaalisen aktiivisuuden yhteydessä. Thetakaistan aktiivisuutta esiintyy etulohkoissa ja etummaisissa päälaen lohkoissa normaalissa EEG:ssä osalla ihmisistä, ja se voi liittyä keskittymiseen tai mentaaliseen tai emotionaaliseen aktiivisuuteen. Lisäksi esimerkiksi hyperventilointi, uneliaisuus ja unitila lisäävät thetakaistan aktiivisuutta. Thetakaistan normaali aktiivisuus ohimolohkojen alueilla voi myös lisääntyä ikääntymisen seurauksena. Deltakaistan aktiivisuus valvetilassa voi olla normaalia lapsilla ja yli 60-vuotiailla, joilla delta-aktiivisuutta esiintyy uneliaisuuden, hyperventiloinnin ja syvän unen yhteydessä.[16]

Parkinsonin taudin tiedetään aiheuttavan muutoksia EEG:n rytmisessä aktiivisuudessa - erityisesti alfarytmin hidastuminen ja matalampien delta- ja thetataajuuksien aktiivisuuden lisääntymisen tiedetään olevan tyypillistä Parkinsonin taudille. EEG:n käyttöä Parkinsonin taudin tunnistamisessa voi vaikeuttaa se, että Parkinsonin taudista kärsivien testihenkilöiden taajuuspektrit eroavat merkittävästi riippuen siitä, kärsiikö testihenkilö myös dementiaasta.[17] Magnetoencefalogrammi-datan<sup>8</sup> tilastollisessa spektrianalyysissä on vahvistettu, että edellä kuvattu hidastuminen on tilastollisesti merkittävää sellaisilla potilailla, jotka eivät kärsi dementiaasta. Samassa analyysissä on havaittu hidastumista tapahtuvan Parkinsonin taudin varhaisvaiheessa jo matalilla alfaaajuuksilla (8-10 Hz). [19] Tämä havainto voi antaa teoreettista tukea menettelylle tarkastella alfaaajuuksia jaettuna matalaan ja korke-

---

<sup>8</sup>Magnetoencefalografia (lyh. MEG) mittaa aivojen magneettikenttiä, jotka aiheutuvat samasta aivojen sähköisestä toiminnasta jota EEG:ssä seurataan päänahan jännitteen vaihteluna. Signaalia välittävän fyysikaalisen ilmiön lisäksi MEG:n ja EEG:n sisältämän informaation laatu, määrä ja tarkkuus vaihtelevat jonkin verran ja niitä voikin pitää toisiaan täydentävinä teknologioina.[18]

aan komponenttiin. Saman tutkimusryhmän myöhemmässä tutkimusraportissa on raportoitu, että Parkinsonin taudille on ominaista MEG-kanavien synkronisoituminen<sup>9</sup>, jolla on tilastollinen korrelaatio taudin vaiheen ja vakavuuden kanssa. Taudin varhaisessa vaiheessa havaittiin synkronisaatiota tapahtuvan interhemisfäärisesti matalilla alfaaajuuksilla (8-10 Hz), siinä missä pidemmälle edistyneessä vaiheessa havaittiin synkronisaatiota tapahtuvan paikallisesti theta-, alfa- ja betataajuuksilla (tässä tutkimuksessa 13-30 Hz). [21] Nämä tulokset antavat osaltaan aihetta etsiä Parkinsonin taudin biomarkereita laajalta taajuuspektriltä.

Parkinsonin taudin EEG-aineistoille perustuvaa koneoppimispohjaista analyysiä koskeva kirjallisuus on pääosin keskittynyt korkean luokittelutarkkuuden tavoittelemiseen Parkinsonin tautia sairastavien ja terveiden koehenkilöiden välillä, missä onkin saavutettu hyviä tuloksia. Toissijaisemmin on tarkasteltu Parkinsonin tautia sairastavien koehenkilöiden EEG-tallenteiden poikkeavuuksia esimerkiksi aivolohkoittain. Tällaiseen tutkimukseen on tyypillisesti käytetty suurta määrää elektrodia, esittämättä kysymystä täyden elektrodivalikoiman tarpeellisuudesta luokittelun onnistumisen kannalta. Sen sijaan systemaattista tutkimusta optimaalisesta elektrodien määrästä tai sijainnista ei ilmeisesti ole aikaisemmin tehty. Suuren elektrodimäärän käyttäminen voi kuitenkin olla epäkäytännöllistä ja potilaan kannalta epämiellyttävää, joten kysymys siitä, voiko pienemmällä elektrodijoukolla mitatun EEG-tallenteen pohjalta luokitella koehenkilöt Parkinsonin tautia sairastaviin ja terveisiin korkealla tarkkuudella on perusteltu. Tämän tutkielman kirjoittamisen yhteydessä on suunniteltu ja toteutettu koneoppimispohjainen EEG-analyysi, jonka ensisijaisena tutkimuskysymyksenä on Parkinsonin tautia sairastavien (jatkossa “PD-positiivisten”) koehenkilöiden erottaminen terveistä (“negatiivisista”) kontrollikoehenkilöistä mahdollisimman suurella tarkkuudella käyttäen mahdollisimman

---

<sup>9</sup>Synkronisoitumista kvantifioidaan tässä yhteydessä synkronisaatiokovavuuden avulla (engl. synchronization likelihood). ks. esim. [20]

suppeaa elektrodien joukkoa. Elektrodien valintamenetelmäksi valikoitui määrätyn suuruisen, lokaalisti optimaalisen elektrodien osajoukon löytävä budjettiperustainen ahne hakualgoritmi. Toisena, merkittävänä uutuusarvoa lisäävänä ominaisuutena voidaan pitää sitä, että analyysin aineisto on suurempi kuin aikaisemmissa koneoppimispohjaisissa EEG-analyyseissä. Tämän tutkielman keskeinen sisältö on tämän analyysin vaiheiden ja tulosten kuvailussa, sekä sen taustoittamisessa ja kontekstoinnissa suhteessa aiempaan tutkimukseen.

Tutkielman toisessa luvussa esitellään analyysissä käytetyt menetelmät datan prosessointiin, piirteiden ekstraktoimiseen ja valintaan, aineiston luokitteluun sekä luokittelumallin suorituskyvyn arviointiin. Kolmannessa luvussa tehdään katsaus EEG- ja osin MEG-aineistoja käyttävään Parkinsonin taudin koneoppimispohjaiseen tutkimukseen. Neljännessä luvussa kuvaillaan tämän tutkielman kirjoittamisen yhteydessä toteutettua Parkinsonin taudin tunnistamiseen liittyvää koneoppimispohjaista analyysiä aineiston, analyysin vaiheiden ja tulosten osalta.

## 2 Analyysin metodologia

Tämän tutkielman kirjoittamisen yhteydessä on toteutettu koneoppimispohjainen EEG-analyysi, jonka tavoitteena on tutkia yksinkertaisen logistisen regressiomallin kykyä erottaa PD-positiiviset tapaukset negatiivisista kontrollitapauksista. Analyysin lähtökohtana oli ajatus siitä, että monikanavaisen EEG-tallenteen eri elektrodikanavien välillä on huomattavaa korrelaatiota. Toisaalta EEG-tallenteen eri elektrodikanavien sisältämä informaatio on luokittelun - siis käytännössä taudin tunnistamisen - näkökulmasta vaihtelevan arvoista. Olisi toivottavaa, että esimerkiksi Parkinsonin taudin tunnistamisessa käytettävä EEG-tallenne voitaisiin kerätä käyttäen mahdollisimman suppeaa elektrodien osajoukkoa, olettaen että tunnistustarkkuus pysyisi lähes yhtä korkealla tasolla - tai parhaassa tapauksessa jopa kohen-tuisi - verrattuna johonkin usean elektrodikanavan standardiin. Kun huomioidaan EEG:n elektrodikanavien oletettu keskinäinen korrelaatio ja niiden vaihteleva arvo luokittelun näkökulmasta, sekä edellä kuvattu motivaatio etsiä minimaalista elektrodien osajoukkoa, on hyvin perusteltua tutkia koneoppimismallin opetusaineistona käytettävän EEG-datan kokoonpanon (elektrodikanavien valinnan näkökulmas-ta) vaikutusta mallin suorituskykyyn. Tässä analyysissä on valittu lähestymistavaksi etsiä testiperformanssiestimaatin lokaalia optimia ahneen hakualgoritmin avulla. Tämän tutkielman kirjoittamisen yhteydessä toteutettu analyysi on tehty käyttäen Python3-ohjelmointikieltä[22], sekä sen kirjastoja Numpy[23], Scipy[24], Scikit-learn[25], Neurokit2[26] ja AntroPy[27].

Tässä luvussa esitellään tutkielman yhteydessä toteutetun koneoppimispohjaisen analyysin metodologista käsitteistöä, sisältäen esiprosessointiin, piirteenekstraktointiin, luokitteluun sekä luokitteluperformanssin estimointiin käytettyjen menetelmien käytännöllistä ja teoreettista pohjustusta.

## 2.1 PREP-prosessointiputki

EEG-aineistojen käyttökelpoisuus (esim. koneoppimispohjaisessa) analyysissä usein edellyttää laajamittaista esiprosessointia, johon kuuluu tavallisesti häiriöisten kanalien tunnistamista ja korjaamista tai poistamista, silmien ja lihasten liikkeistä aiheutuvien häiriöartefaktien poistamista sekä signaalin häiriökomponentin poistamista. Näiden esiprosessointivaiheiden toteuttamiseen on olemassa suuri valikoima erilaisia algoritmisia menetelmiä, joista osa saattaa estää tai tehdä vaikeaksi prosessoidun datan käyttämisen jatkossa tiettyihin tarkoituksiin, mistä syystä näitä menetelmiä ei ole omaksuttu EEG-aineistojen esiprosessoinnin standardiksi. Standardimenettelyn olemassaolo olisi kuitenkin toivottavaa muun muassa aineistojen jakamisen ja yhdistettyjen aineistojen analysoinnin näkökulmasta. PREP-prosessointiputki on suunniteltu tällaiseksi EEG-aineistojen esiprosessoinnin minimaaliseksi standardimenettelyksi, joka mahdollistaa aineistojen jakamisen ja laajamittaisten koottujen analyysien toteuttamisen rajoittamatta aineistojen mahdollisia käyttötarkoituksia.

PREP-prosessointiputki koostuu kolmesta päävaiheesta: verkkovirran aiheuttaman häiriön poistaminen signaalista, “robustista” referentiointimenettelystä, sekä huonojen kanalien tunnistamisesta ja niiden interpoloinnista. Verkkovirtahäiriön poistaminen tapahtuu käyttäen numeeriseen laskentaan tarkoitettua MATLAB-ohjelmointikielen ja -ympäristön EEG-nauhoitteiden prosessointiin tarkoitettua EEGLAB-laajennuksen[28] *cleanline*-liitännäistä. Referentiointimenettely perustuu iteratiiviseen prosessiin, jossa kanalien “todellinen” keskiarvo lasketaan niin, että huonoiksi määritellyt kanavat ensin interpoloidaan keskiarvon laskemis-

ta varten. Huonojen kanavien tunnistamiseen käytetään (muiden kriteerien ohella) laskettua keskiarvoa, joten todellinen keskiarvo tarkentuu iteraatioiden myötä. Näin laskettua kanavien todellista keskiarvoa käytetään huonojen kanavien tunnistamiseen ja interpolointiin käyttäen seitsemännen asteen legendren polynomiaaleja.<sup>1</sup>[30]

## 2.2 Digitaalisen suodatinteknologian menetelmistä

Signaalilla tarkoitetaan matemaattisesti tulkittuna funktiota, joka välittää informaatiota jostakin fyysikaalisesta ilmiöstä. Tämän tutkielman kannalta keskeisiä signaaleja ovat EEG-nauhoitteen kanavat, jotka välittävät informaatiota aivojen toiminnasta. Mielivaltaisen EEG-nauhoitteen kanavan voi esittää ajan funktiona  $x(t)$ , jolloin se ilmaisee elektrodin mittaamaa jännitettä ajanhetkellä  $t$ . Koska tämän tutkielman yhteydessä toteutetun analyysin aineistona on säännöllisin aikaintervallein näytteistetty digitaalinen EEG-nauhoite, on aikasarja diskreetti, jolloin eroa voidaan ilmentää merkinnällä  $x(n)$ , missä  $n$  merkitsee ajanhetken sijaan aikasarjan indeksii. Signaalin lisäksi oleellinen käsite tässä alaluvussa on järjestelmä (engl. system), jolla tarkoitetaan sellaista prosessia, joka ottaa vastaan syötesignaalin ja tuottaa tulostesignaalin. Tässä alaluvussa käsitellyt suodattimet ovat eräänlaisia lineaarisia järjestelmiä, eli ne ovat homogeenisia ja additiivisia.<sup>2</sup>

Suodattaminen (engl. filtering) on hyvin yleinen toimenpide monenlaisten signaalien esiprosessointivaiheena. Tarkoituksena on parantaa signaalin laatua sille aiotun käyttötarkoituksen näkökulmasta vaimentamalla määrätty ei-toivottu osa signaalin taajuusspektristä, ja päästämällä muu osa läpi mahdollisimman muuttumattomana.

---

<sup>1</sup>Legendren polynomiaalien käytöstä EEG-kanavien interpolointiin ks. esim. [29]

<sup>2</sup>Homogeenisuudella tarkoitetaan ominaisuutta, että muutos syötesignaalin amplitudissa aiheuttaa samansuuruisen muutoksen tulostesignaalin amplitudissa. Additiivisuus taas tarkoittaa sitä, että kahden tai useamman mielivaltaisen yhteenlasketun syötesignaalin tulostesignaali on sama kuin syötesignaalien yhteenlasketut tulostesignaalit.

Signaalin spektrillä tarkoitetaan sen taajuuskomponenttien järjestettyä listausta, ja sen voi muodostaa fourier-muunnoksen avulla. Rekursiivisilla suodattimilla tarkoitetaan sellaisia suodattimia, joiden toteutettava algoritmi käyttää tulostesignaalin arvon laskemiseen syötesignaalin lisäksi myös tulostesignaalin aiempia arvoja. Rekursiivisiin suodattimiin kuuluvat tyyppien I- ja II- Chebyshev-suodattimet, elliptiset suodattimet sekä Butterworth-suodattimet.

Jaksollisten signaalien, kuten siniaaltofunktioiden, amplitudilla tarkoitetaan yksinkertaisesti sen jakson ääripisteiden etäisyyttä jaettuna kahdella. Käytännössä usein käytetään amplitudin neliöllistä keskiarvoa (engl. root mean square, lyh. RMS). EEG-sovelluksissa amplitudin suuruutta ilmaistaan yleensä mikrovolteissa. Järjestelmän vahvistuksella (engl. gain) tarkoitetaan muutosta signaalin amplitudissa, kun se kulkee järjestelmän läpi. Merkitään syötesignaalin amplitudia  $V_{in}$  ja tulostesignaalin amplitudia  $V_{out}$ . Tällöin järjestelmän vahvistus  $gain$  on:

$$gain = 10 \log_{10} \left( \frac{V_{out}}{V_{in}} \right) dB \quad (2.1)$$

Järjestelmän päästökaistaksi (engl. passband) kutsutaan järjestelmän frekvenssivasteen sitä aluetta, johon ei kohdistu (suurta) negatiivista vahvistusta, ts. vaimennusta (engl. attenuation). Estokaistaksi (engl. stopband) puolestaan kutsutaan järjestelmän frekvenssivasteen sitä aluetta, johon kohdistuu merkittävää vaimennusta. Rajataajuudeksi (engl. cut-off frequency) kutsutaan sellaista taajuutta, jonka kohdalla järjestelmän vahvistus on  $-3dB$ , mikä vastaa signaalin amplitudin puolittumista. Rajataajuus merkitsee päästö- ja pysäytyskaistan välistä rajaa. Frekvenssivasteen alueesta, jolla vahvistus muuttuu merkittävästi kutsutaan siirtymäkaistaksi (engl. transition band).

Suodattimen suorituskykyä voidaan arvioida muun muassa sen frekvenssivasteen siirtymäkaistan kulmakertoimen (engl. roll-off) jyrkkyyden tai sen päästö- tai pysäytyskaistan amplitudin aaltoilun (engl. ripple) perusteella. Siirtymäkaistan mahdolli-

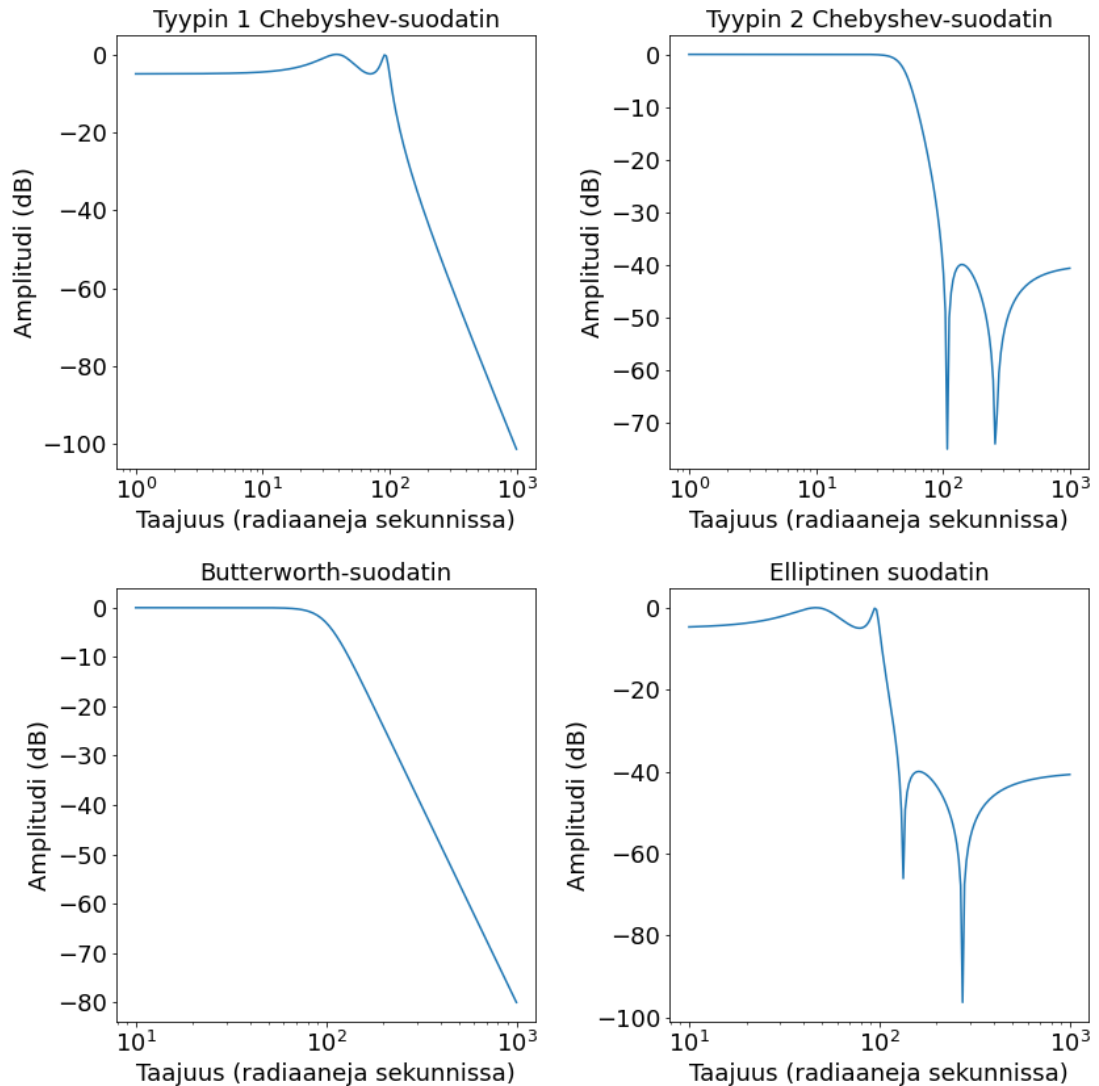
simman jyrkkää kulmakerrointa pidetään toivottavana, sillä se merkitsee tarkempaa erottelua päästö- ja pysäytyskaistojen välillä. Aaltoilu päästö- ja pysäytyskaistoilla puolestaan vääristää suodatettua signaalia, joten sen minimoiminen on niin ikään toivottavaa.

Idealisesti suodattimen siirtymäkaistan kulmakertoimen tulisi olla mahdollisimman jyrkkä, ja aaltoilu päästö- ja pysäytyskaistoilla minimaalista. Käytännössä yleisten rekursiivisten suodatinmallien yhteydessä on tehtävä valinta jyrkän kulmakertoimen ja päästö- tai pysäytyskaistan aaltoilun väliltä. Sallimalla enemmän aaltoilua, voidaan saavuttaa jyrkempi kulmakerroin.

Chebyshev-suodatin on rekursiivisen äärettömän impulssivasteen suodattimen malli, jossa on suhteellisen jyrkkä siirtymäkaistan kulmakerroin, mutta joka aiheuttaa aaltoilua päästö- tai pysäytyskaistassa. Tyypin 1 Chebyshev-suodatin aiheuttaa aaltoilua päästökaistassa, ja tyypin 2 Chebyshev-suodatin puolestaan pysäytyskaistassa. Butterworth-suodattimen päästökaistan frekvenssivaste on täysin tasainen, mutta sen siirtymäkaistan kulmakerroin on loivempi kuin Chebyshev-suodattimilla. Elliptinen suodatin yhdistää Chebyshev-suodattimen tyyppien 1 ja 2 ominaisuuksia aiheuttaen aaltoilua sekä päästö-, että pysäytyskaistaan, mutta mahdollistaen hyvin jyrkän siirtymäkaistan kulmakertoimen. Siirtymäkaistan kulmakertoimen jyrkkyyttä voi lisätä käyttämällä korkeampaa suodattimen kertalukua, mikä teoriassa merkitsee usampien napapisteiden käyttämisestä suodattimen suunnitteluun. Käytännössä korkean ( $> 2$ ) kertaluvun suodattimet toteutetaan yleensä porrastamalla (engl. cascade) useita toisen kertaluvun suodattimia. Alipäästösuodattimesta voi ns. spektraalinversion avulla tehdä ylipäästösuodattimen, ja yhdistämällä ali- ja ylipäästösuodattimia voi tehdä kaistanpäästö- tai kaistanestosuodattimen.[31] Kuvassa 2.1 on esitettyinä erityyppisten rekursiivisten äärettömän impulssivasteen suodatinmallien alipäästösuodatinten tyypillisiä taajuusvasteita.

Butterworth-suodattimen suunnittelu perustuu Laplace-muunnoksen diskreet-

Kuva 2.1: Erityyppisten alipäästösuodatinten taajuusvasteita. X-akselilla taajuus radiaaneina / sekuntia. Y-akselilla suodattimen vahvistus desibeleinä.



tiin muunnelmaan, niin kutsuttuun z-muunnokseen. Laplace-muunnokset ovat menetelmä differentiaaliyhtälöiden ratkaisemiseen, ja soveltuvat sellaisten lineaaristen järjestelmien analysointiin, joiden matemaattinen mallintaminen edellyttää differentiaaliyhtälöiden käyttämistä. Siinä missä Laplace-muunnos on kuvaus jatkuvasta kompleksiselta aika-alueelta kompleksiselle s-alueelle (engl. s-domain), on z-muunnos kuvaus diskreetiltä aika-alueelta kompleksiselle z-tasolle (engl. z-domain). Sekä Laplace- että z-muunnosten käyttäminen suodatinsuunnitteluun perustuu niin

kutsuttujen nolla- ja napapisteiden etsimiseen, jotka esiintyvät muunnoksen laskukaavan integraalilausekkeen konvergenssialueella, ja sen rajapinnoilla. Butterworth-suodattimen suunnitteluprosessin teknisiin yksityiskohtiin voi perehtyä tarkemmin esim.[32]

## 2.3 Performanssimetriikat

### 2.3.1 Luokittelutarkkuus

Ensisijainen koneoppimismallin suorituskyvyn laskemiseen käytetty metriikka oli luokittelutarkkuus (engl. accuracy, lyh. ACC), joka ilmoittaa oikein luokiteltujen tapauksien lukumäärän suhteessa sekä oikein että väärin luokiteltujen tapauksien lukumäärään. Luokittelutarkkuusmetriikan täsmällisempi määrittely perustuu binäärisen luokittelutehtävän tapauksessa oikein luokiteltujen positiivisten tapauksien (engl. true positive, lyh. TP), väärin luokiteltujen positiivisten tapauksien (engl. false positive, lyh. FP), oikein luokiteltujen negatiivisten tapauksien (engl. true negative, lyh. TN) ja väärin luokiteltujen negatiivisten tapauksien (engl. false negative, lyh. FN) käsitteisiin, jotka ilmoittavat tarkastelussa olevan observaation ennustetun ja todellisen luokan. Näiden käsitteiden avulla luokittelutarkkuus voidaan binäärisen luokittelutehtävän tapauksessa määritellä seuraavasti:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.2)$$

### 2.3.2 ROC-käyrän alapuolinen pinta-ala (AUC)

Luokittelutarkkuuden lisäksi toissijaisena kriteerinä hakualgoritmissa käytettiin AUC-estimaattia (lyhenne sanoista "area under the curve") tilanteissa, joissa kahden vaihtoehdoisen mallin luokittelutarkkuus olisi ollut samansuuruinen. AUC on binäärisen luokittelutehtävän performanssimetriikka, jonka laskeminen perustuu sen

todennäköisyyden estimointiin, että luokittelija arvottaa mielivaltaisesti valitun positiivisen observaation mielivaltaisesti valitun negatiivisen observaation edelle, kun oletetaan että positiivisen tapauksen tulisi olla ensisijainen negatiiviseen nähden. AUC-metriikan laskeminen perustuu niin kutsutun ROC-käyrän (lyhenne sanoista "receiving operating characteristic") alle jäävän alueen pinta-alan laskeminen. ROC-käyrän piirtäminen puolestaan perustuu sensitiivisyyden (engl. sensitivity, myös true positive rate, lyh. TPR) ja spesifisyyden (engl. specificity, myös true negative rate, lyh. TNR) käsitteisiin, jotka määritellään seuraavasti:

$$TPR = TP / (TP + FN) \quad (2.3)$$

$$TNR = TN / (TN + FP) \quad (2.4)$$

Spesifisyyden avulla voidaan määritellä väärin negatiivisten esiintyvyys (engl. false positive rate, lyh. FPR) esimerkiksi seuraavasti:

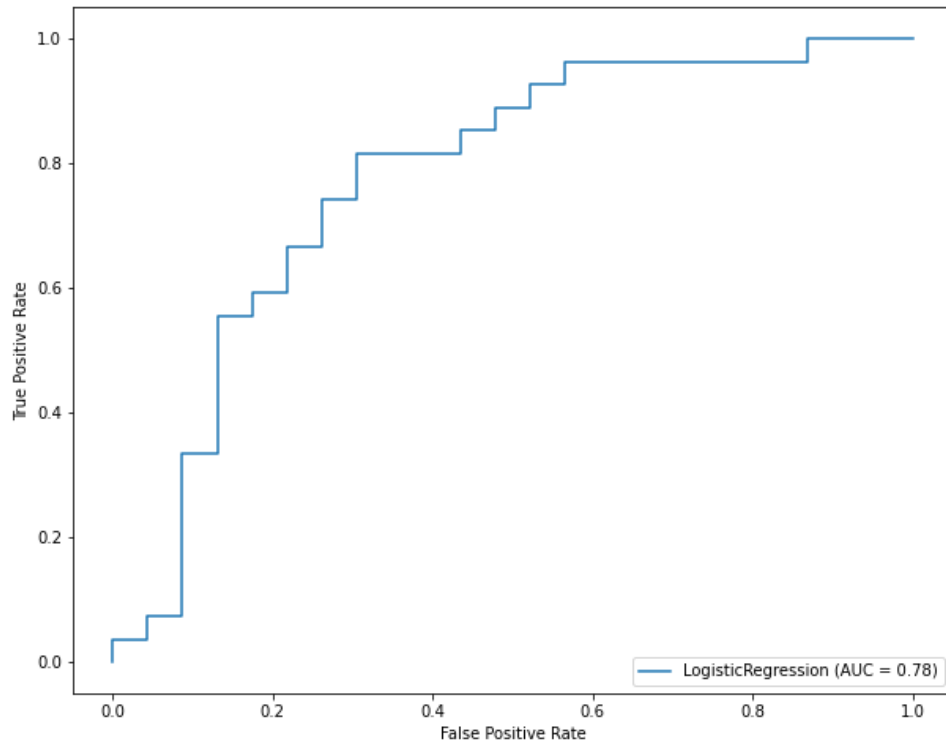
$$FPR = 1 - TNR \quad (2.5)$$

ROC-käyrä piirretään koordinaatistoon, jonka X-akseli vastaa yleensä FPR-arvoa ja Y-akseli TPR-arvoa, madaltamalla vähitellen luokittelijan päätöskynnystä (engl. decision threshold), jolloin luokittelijan todennäköisyys luokitella mielivaltainen observaatio positiiviseen luokkaan kasvaa. Päätöskynnyksen madaltaminen kasvattaa näin molempia arvoja, ja ROC-käyrän päätepisteessä (1, 1) jokainen tapaus luokitellaan positiiviseksi. Näin käyrän voidaan tulkita ilmentävän sitä, miten nopeasti TPR-arvo kasvaa suhteessa FPR-arvoon, kun päätöskynnystä madalletaan. AUC-arvon voi laskea tällaisesta kuvaajasta laskemalla käyrän alle jäävän alueen pinta-alan, joka on rajoittunut välille [0, 1], ja sitä suurempi mitä todennäköisemmin luokittelija arvottaa mielivaltaisesti valitun positiivisen tapauksen satunnaisen negatiivisen tapauksen edelle. [33] Kuvassa 2.2 esimerkin vuoksi generoitua dataa käyttäen

tuotettu ROC-käyrä.

Kuva 2.2: ROC-käyrä generoidulla datalla sovitetusta logistisesta regressiomallista.

Tässä ROC-käyrän alle jäävän alueen suuruus (AUC) on 0.78.



Tässä analyysissä ei ole tehty varsinaista ROC-analyysiä siinä mielessä, että luokittelijan tulosten pohjalta olisi laadittu ROC-käyrää, ja onkin varsin yleistä että AUCia käytetään puhtaasti performanssiestimaatin metriikkana<sup>3</sup> luokittelutarkkuuden rinnalla.

---

<sup>3</sup>AUC-estimaatin laskemiseen on käytetty Scikit-Learn -kirjaston `roc_auc_score`-funktiota.

## 2.4 Epälineaariset systeemit ja entropia

Muun muassa fysiikassa sekä matematiikan sovellusaloilla systeemillä tarkoitetaan jonkinlaista tarkasteltavaa kohdetta, joka voi olla konkreettinen tai abstrakti. Systeemiä luonnehtii yksi tai useampi muuttuja, joiden arvot annetulla ajanhetkellä määrittävät systeemin tilan. Systeemin tilaa voidaan kuvata pisteenä tila-avaruudeksi (engl. phase/state space) kutsutussa koordinaatistossa, jossa jokaista muuttujaa vastaa koordinaattiakseli. Systeemin tilaa ajan funktiona esittävää polkua tila-avaruudessa kutsutaan systeemin liikeradaksi (engl. trajectory).

Dynaamiseksi systeemiksi (engl. Dynamical system) kutsutaan sellaista systeemiä, jonka tila kehittyy ajan kanssa jonkin tietyn säännön mukaisesti. Dynaamisen systeemin kehitykseen vaikuttavat mainittu sääntö, sekä systeemin alkutila. Mikäli systeemin liikerata voidaan yksikäsitteisesti määrittää alkutilan perusteella, on kyseessä deterministinen systeemi. (engl. Deterministic system/process). Jos sitä vastoin systeemin tilan kehitykseen sisältyy satunnaisuutta, puhutaan epädeterministisestä tai stokastisesta systeemistä (engl. Stochastic system/process).

Lineaariset systeemit (engl. Linear system) noudattavat superpositioperiaatetta, mikä tarkoittaa että ne täyttävät additiivisuuden ja homogeenisyyden ehdot. Epälineaariset systeemit (engl. Nonlinear system) sen sijaan eivät täytä näitä ehtoja, mikä käytännössä merkitsee sitä, että systeemin syötearvojen muutokset (nk. interventiot) voivat aiheuttaa muutoksia systeemissä, jotka ovat mittakaavaltaan suhteettomia syötearvojen muutoksiin nähden.

Edellä mainittujen kategorioiden lisäksi on syytä mainita kompleksiset (engl. complex) ja kaoottiset (engl. chaotic) systeemit. Kompleksiset systeemit koostuvat useista alisysteemeistä, joiden keskinäinen vuorovaikutus saa aikaan monimutkaista yhteistoimintaa, siinä missä kaoottiset systeemit voivat koostua suhteellisen harvoista alisysteemeistä, joiden yhteistoiminta on kuitenkin huomattavan monimutkaista ja mahdollisesti näennäisen satunnaista. Kaoottisten systeemien tuottamat aikasar-

jat ovat tyypillisesti fraktaalisia, millä tarkoitetaan sitä, että ne ilmentävät samanlaista rakennetta monessa eri mittakaavan tasossa. Sekä kompleksiset että kaoottiset systeemit ovat epälineaarisia dynaamisia systeemejä, mistä syystä ne molemmat ovat herkkiä systeemin muuttujien alkuarvoille.[34]

Näyte-entropia on epälineaaristen dynaamisten systeemien analyysiin soveltuva menetelmä, jota voidaan käyttää esimerkiksi fysiologisten aikasarjojen kompleksisuuden kvantifointiin.[35] Sen laskeminen perustuu todennäköisyyden arvioimiseen, että kaksi tarkasteltavan aikasarjan  $m$ :n pisteen pituista alisekvenssiä ovat edelleen samanlaisia  $m + 1$ :n pituudessa. Se liittyy läheisesti likiarvoisentropiana (engl. approximate entropy) tunnettuun kompleksisuusmittaan, mutta eliminoi likiarvoisentropiassa esiintyvän vinouman, joka aiheutuu siitä, että likiarvoisentropiassa jokainen alisekvenssi lasketaan täsmäämään myös itsensä kanssa. Likiarvoisentropia perustuu puolestaan Kolmogorov-Sinain entropiana tunnettuun eksaktiin säännöllisyysmittaan, joka ei sellaisenaan yleensä sovellu häiriötä sisältävän datan analysointiin.[36]

Olkoon  $X = x_1, x_2, \dots, x_N$   $N$ :n pisteen aikasarja,  $m$  upotusdimensio (tai kertaluku),  $r$  kahden alisekvenssin samankaltaisuuden toleranssiarvo ja  $d(u, v)$  Chebyshevin<sup>4</sup> etäisyysmetriikka. Upotusdimension avulla konstruoidaan templaattivektoreiden joukko  $u_m(i) = [x_i, x_{i+1}, \dots, x_{i+m-1}]$  kullekin  $i$ :n arvolle siten, että  $1 \leq i \leq N - m$  upotusavaruudessa  $\mathbf{R}^m$ . Toinen templaattivektoreiden joukko  $u_{m+1}(i)$  konstruoidaan upotusavaruudessa  $\mathbf{R}^{m+1}$ . Näyte-entropia määritellään tällöin seuraavasti:

$$SE(x(n), d(i, j), m, r) = -\log \frac{A}{B} \quad (2.7)$$

missä

---

<sup>4</sup>Chebyshevin etäisyys kahden vektorin välillä määritellään vektorien suurimpana etäisyytenä minkä tahansa koordinaattidimension suhteen, ts.

$$d(u, v) = \max_i (|u_i - v_i|) \quad (2.6)$$

Toisinaan käytetään myös muita etäisyysmetriikoita, kuten euklidista etäisyysmetriikkaa.

$$A = \frac{1}{N-m} \sum_{i=1}^{N-m} \frac{1}{N-m-1} \sum_{j=1, j \neq i}^{N-m} \theta(r - d(u_{m+1}(i), u_{m+1}(j))) \quad (2.8)$$

ja

$$B = \frac{1}{N-m} \sum_{i=1}^{N-m} \frac{1}{N-m-1} \sum_{j=1, j \neq i}^{N-m} \theta(r - d(u_m(i), u_m(j))) \quad (2.9)$$

ja missä  $\theta$  merkitsee Heavisiden askelfunktiota.<sup>5</sup>

## 2.5 Out-of-sample -evaluointi ja ristiinvalidointialgoritmit

Ohjatussa koneoppimisessa koneoppimismallin suorituskykyä arvioidaan tavallisesti jonkin soveltuvan metriikan avulla mallin kykynä ennustaa (engl. predict) mallin sovittamiseen käytetyn opetusjoukon ulkopuolisia observaatioita vastaavien ennustettavien muuttujien arvoja. Koneoppimismallin sovittamisella tarkoitetaan koneoppimismallin parametrien optimoimista niin, että mallin (mahdollisesti regularisoitu) virhetermi opetusaineistolla on mahdollisimman pieni. Sovitetun koneoppimismallin yleistykyvyn arvioimista opetusaineiston ulkopuolisella aineistolla kutsutaan out-of-sample -testaamiseksi, ja siihen käytettävää aineistoa testiaineistoksi. Menettely on perusteltu, sillä mallin testaaminen sen sovittamiseen käytetyllä aineistolla johtaa huomattavaan positiiviseen vinoumaan maliin suorituskykyä ilmaisevissa tuloksissa, eikä anna realistista kuvaa mallin kyvystä tehdä ennusteita aineistolle, jota ei ole käytetty sen sovittamiseen.

---

<sup>5</sup>Heavisiden askelfunktio yhdistää jokaiseen syötearvoon  $n$  tulosteen 0 jos  $n < 0$  ja 1 if  $n \geq 0$ ;  
ts.

$$\theta(n) = \begin{cases} 0 & \text{if } n < 0, \\ 1 & \text{if } n \geq 0 \end{cases}$$

Yksinkertaisin tapa koneoppimismallin yleistyskyvyn arvioimiseen out-of-sample -aineistoa käyttäen lienee niin sanotun holdout-menettelyn käyttäminen, mikä merkitsee koko aineiston jakamista opetusaineistoon, jota käytetään vain mallin sovittamiseen, ja testiaineistoon, jota käytetään vain mallin yleistyskyvyn arvioimiseen. Kun mallin yleistyskyvyn arviointi tehdään aineistolla, jota ei ole käytetty mallin sovittamiseen, saadaan realistinen käsitys mallin yleistyskyvystä olettaen, että testiaineisto on riittävän suuri ja että se on edustava näyte koko aineistosta. Holdout-menettelyn käyttämistä rajoittaa kuitenkin menettelyn toteuttamiseen tarvittavan aineiston suhteellisen suuri koko, mikä vähentää puolestaan mallin sovittamiseen käytettävissä olevan aineiston kokoa, johtaen vaatimattoman kokoisen aineiston tapauksessa negatiivisesti vinoutuneeseen arvioon yleistyskyvystä mallin alisovittuessa<sup>6</sup> liian pieneen opetusaineistoon. Monilla sovellusaloilla on käytettävissä hyvin rajoitettuja määriä koneoppimiseen soveltuvaa korkealaatuista annotoitua dataa, joten riittävän suuren testiaineiston asettaminen syrjään ei ole usein käytännössä mahdollista.

Suhteellisen pienen aineiston tapauksessa on usein mielekkäämpää käyttää holdout-menettelyn sijaan jotakin ristiinvalidointialgoritmia koneoppimismallin yleistyskyvyn arvioimiseen. Ristiinvalidoinniksi (engl. cross-validation, lyh. CV) kutsutaan menettelyä, jossa aineisto partitioidaan (likimain) yhtä suuriin osajoukkoihin. Osajoukkoja kutsutaan "laskoksiksi" (engl. fold) ja niitä on yleensä kaikkiaan joko 5, 10 tai erityistapauksena aineiston observaatioiden lukumäärä, jolloin kyse on ns. "leave-one-out" -ristiinvalidoinnista. Käytettävien laskosten määrä on tapana ilmoittaa algoritmin nimen yhteydessä. Jokainen laskoksista toimii vuorollaan testiaineistona lopun aineiston muodostaessa mallin opetusaineiston.  $K$ :n laskoksen ristiinvalidoin-

---

<sup>6</sup>Alisovittumisella tarkoitetaan tilannetta, jossa koneoppimismalli ei onnistu mallintamaan opetusaineiston observaatioiden ja kohdemuuttujan välistä suhdetta, jolloin mallin suorituskyky sekä opetus- että testiaineistoilla jää heikoksi.

tialgoritmissa siis sovitetaan ja testataan kaikkiaan  $k$  mallia. Kokonaissuorituskykyarvio voidaan tuottaa kahdella eri tavalla: arvio yleistyskyvystä voidaan laatia jokaiselle testiaineistolle erikseen ja sitten laskea näiden testiaineistokohtaisten arvioiden keskiarvo, tai kaikkien mallien tuottamat ennusteet voidaan yhdistää (ns. “pooling”-menettely) ja laskea mallin suorituskyvyn arviointiin käytettävä metriikka vertaamalla niitä samoin yhdistettyihin testiaineistoihin. Näiden kahden laskentatavan välillä ei ole yleensä merkittävää eroa, mutta tiettyjä suorituskykymetriikoita käytettäessä valinnalla voi olla merkitystä. Erityisesti käytettäessä F-arvoa<sup>7</sup> suorituskykymetriikkana usean luokan luokittelutehtävässä voi keskiarvoistusmenettely johtaa merkittävään vinoumaan.[37] Tämän tutkielman näkökulmasta oleellisemmin AUC-estimaatin laskemisen pooling-menettelyllä pienillä datajoukoilla<sup>8</sup> on havaittu aiheuttavan huomattavaa negatiivista vinoumaa, siinä missä sen laskeminen keskiarvoistusmenettelyllä on havaittu aiheuttavan huomattavaa varianssia.[38] Ohjelmalistaus 1 esittää K:n laskoksen ristiinvalidointialgoritmin keskiarvoistuksella.

Koska tämän tutkielman kirjoittamisen yhteydessä toteutetussa analyysissä ei ole käytetty F-arvoa suorituskykymetriikkana, eikä käytettävissä oleva datajoukko ole kohtuuttoman pieni, oletetaan jatkossa että ristiinvalidoinnin kokonaissuorituskykyarvio lasketaan keskiarvoistusmenettelyä käyttäen. Ristiinvalidoinnin merkittävä etu on siinä, että sen käyttäminen mahdollistaa jokaisen observaation esiintymisen

---

<sup>7</sup>F-arvo lasketaan seuraavalla kaavalla:

$$precision = TP / (TP + FP) \quad (2.10)$$

$$recall = TP / (TP + FN) \quad (2.11)$$

$$F = 2 * \frac{precision * recall}{precision + recall} \quad (2.12)$$

<sup>8</sup>Referoidun artikkelin tutkimusasetelmassa pienen aineiston koko on  $N = 30$ .

---

**Ohjelmalistaus 1** K:n laskoksen ristiinvalidointi

---

**Input:** Luokittelija  $Clf$ , Observaatiot  $X$ ,Kohdemuuttujat  $Y$ , Laskosten lukumäärä  $K$ **Output:** Luokittelijan  $Clf$  yleistyskyvyn estimaatti  $P$  $P \leftarrow 0$  $\{F_1, \dots, F_K\} \leftarrow part(X, K)$  ▷ Laskosten indeksien alustaminen**for**  $F_i$  **in**  $\{F_1, \dots, F_K\}$  **do** ▷ Iteroidaan laskosten yli $X_{train} \leftarrow X - X[F_i]$  ▷ Opetusaineisto $Y_{train} \leftarrow Y - Y[F_i]$  $X_{test} \leftarrow X[F_i]$  ▷ Testiaineisto $Y_{test} \leftarrow Y[F_i]$  $Clf \leftarrow fit(Clif, X_{train}, Y_{train})$  ▷ Mallin sovittaminen $P \leftarrow P + score(Clif, X_{test}, Y_{test})$  ▷ Mallin testaaminen**end for** $P \leftarrow \frac{P}{K}$  ▷ Keskiarvo laskosten testiperformanssien summasta**Return**  $P$ 

---

mallin opetus- ja testijoukoissa, lisäksi näin aineistosta saatavaa hyötyä mallin sovittamisen ja testaamisen kannalta, mikä on mallin suorituskykyarvion vinouman ja varianssin vähentämisen näkökulmasta toivottavaa erityisesti pienten aineistojen tapauksessa. Ristiinvalidointialgoritmit ovat kuitenkin itsestäänselvästi laskenta-ajan näkökulmasta kalliita, sillä koneoppimismalleja on sovitettava kaikkiaan  $k$  yhden sijaan, ja mallin sovittaminen on useimmiten koneoppimis pohjaisen analyysin kallein operaatio.

### 2.5.1 Sisäkkäinen ristiinvalidointi

Sisäkkäistä ristiinvalidointia (engl. nested) käytetään silloin, kun halutaan sisällyttää koneoppimismallin valinta osaksi ristiinvalidointiprosessia. Käytännössä tämä

tarkoittaa sitä, että kun aineisto on partitionoitu laskoksiin joista jokainen toimii vuorollaan testiaineistona, jäljelle jäävä opetusaineisto partitionoidaan vielä kerran. Näistä “sisemmän” partition laskoksista jokainen toimii vuorollaan validaatioaineistona, jota käytetään arvioimaan sisemmän partition varsinaisella opetusaineistolla sovitettujen vaihtoehtoisten mallien alustavaa yleistyskykyä, jota voidaan kutsua vaikakapa validaatiosuorituskyvyksi. Vaihtoehtoisilla koneoppimismalleilla tarkoitetaan tässä yhteydessä esimerkiksi erilaisilla hyperparametrien<sup>9</sup> arvoilla alustettuja malleja. Vaihtoehtoisista malleista valitaan testattavaksi se, jonka kokonaisvalidaatio-suorituskyky on paras, malli sovitetaan käyttäen koko opetusaineistoa ja sen yleistyskykyä arvioidaan testiaineistona toimivan laskoksen avulla. Kokonaistestisuorituskyky lasketaan kuten tavallisessa “litteässä” ristiinvalidointimenettelyssä, eli mallin yleistyskyvyn arvio voidaan laskea joko testiaineistokohtaisten suorituskykyarvioiden keskiarvona tai pooling-menettelyn avulla. Sisäkkäisen ristiinvalidoinnin käyttäminen mallinvalinnan yhteydessä on perusteltua siksi, että yleistyskyvyn arvioiminen samalla aineistolla jolle mallin hyperparametrit on optimoitu johtaa tyypillisesti ylioptimistiseen arvioon mallin yleistyskyvystä. Edellä kuvattu sisäkkäinen menettely jossa mallinvalinta tehdään erikseen jokaisella ristiinvalidoinnin (tai muun uudelleenotannan) laskoksella noudattaa puolestaan mallinvalinnan parhaita käytänteitä. [39]

## 2.6 Logistinen regressiomalli

Lineaarinen regressio on tilastollinen mallinnusmenetelmä, jota käytetään selittävän muuttujan tai -muuttujien (observaatiot) ja vastemuuttujan (ennustettava muuttu-

---

<sup>9</sup>Hyperparametreja voivat olla esimerkiksi koneoppimismallin regularisointiparametrin arvo, tai esimerkiksi tarkasteltavien naapuriobservaatioiden määrä  $K$ :n lähimmän naapurin menetelmässä, joka ennustaa annetulle testiobservaatiolle enemmistöluokkaa  $k$ :n samanlaisimman opetusobservaation muodostamassa naapuristossa.

ja) välisen lineaarisen suhteen mallintamiseen. Oletetaan, että  $K$ :n selittävän muuttujan ja vastemuuttujan suhde voidaan ilmaista yhtälöllä

$$Y = \alpha + \beta_1 * X_1 + \dots + \beta_K * X_K + \epsilon \quad (2.13)$$

missä  $X_1, \dots, X_K$  ovat selittävät muuttujat,  $Y$  vastemuuttuja,  $\alpha$  intercept-vakiotermi, joka merkitsee regressiosuoran ja  $Y$ -akselin leikkauspistettä,  $\beta_1, \dots, \beta_K$  mallin parametreja, jotka yhdessä määräävät regressiosuoran kulmakertoimen ja  $\epsilon$  virhetermi, joka huomioi sen mahdollisuuden että muuttujien suhdetta ei ole mahdollista ilmaista virheettömästi yllä esitetyn yhtälön muodossa. Tällöin lineaarisen regressioanalyysin päämääränä on estimoida muuttujien välistä lineaarista suhdetta määräämällä intercept-vakiotermin ja parametrien arvot pienimmän neliösumman menetelmää<sup>10</sup> (engl. ordinary least squares, lyh. OLS) käyttämällä. Päädytään yhtälöön

$$\hat{Y} = a + B_1 * X_1 + \dots + B_K * X_K \quad (2.14)$$

Logistinen regressio on luokitteluun soveltuva malli, joka perustuu lineaariseen regressioon, mutta mukautettu soveltumaan tilanteeseen, jossa vastemuuttuja on jatkuvan sijaan binäärinen. Oletetaan, että vastemuuttuja  $Y \in [0, 1]$ , missä 0 ja 1 edustavat luokittelutehtävän kategorioita (ts. luokkia), joihin mielivaltainen tapaus  $x \in X$  voidaan yhdistää. Merkitään  $P(Y = 1)$  todennäköisyyttä, että mielivaltainen tapaus kuuluu luokkaan "1", jolloin vastaavasti todennäköisyys, että mielivaltainen tapaus kuuluu luokkaan "0" on  $1 - P(Y = 1)$ .

Todennäköisyyksien avulla voidaan tilanteessa määritellä niin kutsuttu odds-suure. Olkoon  $p$  jonkin tapahtuman toteutumisen todennäköisyys, jolloin  $1 - p$  on sen vastatapahtuman todennäköisyys. Tällöin odds-suure lasketaan yksinkertaisesti:

---

<sup>10</sup>OLS valitsee sellaiset lineaarisen regressiomallin parametrit jotka minimoivat vastemuuttujan todellisten ja ennustettujen arvojen neliön.

$$\Omega(p) = \frac{p}{1-p} \quad (2.15)$$

Odds-suureen avulla voidaan puolestaan määritellä logit-suure, joka on yksinkertaisesti odds-suureen logaritmi.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (2.16)$$

Arvioimalla  $Y$ :n logit-suuretta lineaarisen regression avulla, ja manipuloimalla saatavaa yhtälöä algebrallisesti voidaan johtaa logistisen regression yhtälö:

$$\text{logit}(Y) = \alpha + \beta_1 * X_1 + \dots + \beta_K * X_K \quad (2.17)$$

Eksponentioimalla puolittain yhtälö saadaan muotoon:

$$\Omega(Y) = e^{\alpha + \beta_1 * X_1 + \dots + \beta_K * X_K} \quad (2.18)$$

Odds-suureen voi muuntaa yhtälön algebrallisella manipuloinnilla takasin todennäköisyyksiksi, jolloin päädytään logistisen regression yhtälöön:

$$P(Y) = \frac{e^{\alpha + \beta_1 * X_1 + \dots + \beta_K * X_K}}{1 + e^{\alpha + \beta_1 * X_1 + \dots + \beta_K * X_K}} \quad (2.19)$$

Yllä esitetty yhtälö rajaa asianmukaisesti vastemuuttujan ennustetut arvot välille  $[0, 1]$ , joten niitä on mielekästä käsitellä todennäköisyyksinä observaation kuulumiselle luokkaan "1".[40]

### 2.6.1 Regularisointi

Lineaaristen koneoppimismallien parametrien optimoinnin yhteydessä malliin liitetään usein jonkinlainen regularisointitermi rajoittamaan optimoidun mallin kompleksisuutta. Mallin kompleksisuuteen vaikuttavat mallin parametrien lukumäärä sekä

niiden mittakaava, joita molempia voidaan rajoittaa regularisoinnin yhteydessä. Regularisointi ehkäisee mallin ylisovittumista, jolla tarkoitetaan ilmiötä jossa malli oppii opetusaineistoon sisältyvän satunnaisuuden yksityiskohtaisesti ja saavuttaa näin korkean suorituskyvyn opetusaineistolla, mutta sen yleisyky jää samalla tästä syystä heikoksi, mikä on havaittavissa matalana testiaineiston suorituskvyraviona.

Regularisointimenetelmää, jossa rajoitetaan vain parametrien yhteenlaskettua kokonaisarvoa lisäämällä skalaarilla kerrottu parametrien neliöiden summa minimoitavaan virhetermiin kutsutaan Tikhonovin regularisointimenetelmäksi, tai harjanne-regressioksi (engl. ridge regression). Tikhonovin regularisointimenetelmän regularisointitermiä kutsutaan l2-rangaistustermiksi<sup>11</sup> (engl. l2-penalty). l2-regularisoinnin virhetermi on muotoa<sup>12</sup>:

$$\sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2 \quad (2.21)$$

missä  $y_i$  on i:nnen observaation vastemuuttujan todellinen arvo,  $\hat{y}_i$  sitä vastaava ennustettu arvo,  $\lambda$  regularisoinnin vahvuutta määrittävä skalaari ja  $\beta_j$  j:nnen selittävän muuttujan parametri.

Tässä tutkielmassa esiteltävän analyysin yhteydessä on käytetty vakiovahvuista ( $C = 1.0$ ) l2-regularisointia ehkäisemään mallin ylisovittumista eräänlaisena kom-

---

<sup>11</sup>Nimitys johtuu siitä melko ilmeisestä syystä, että rangaistustermi perustuu l2-normiin, joka tunnetaan myös eukleideen normina.

<sup>12</sup>Scikit-Learn kirjaston implementaatio logistisesta regressiosta käyttää hieman eri tavoin määriteltyä l2-regularisointia:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_i \log(\exp(-y_i(X_i^T w + c)) + 1) \quad (2.20)$$

missä keskeisenä erona on regularisointikertoimen  $C$  esiintyminen regularisointitermin sijaan varsinaisen virhetermin yhteydessä. Tässä yhtälössä  $C$  toimii pikemminkin regularisointikertoimen käänteislukuna.

promissina laskenta-ajan ja ylisovittumisen ehkäisyn välillä. l2-regularisoinnin suosiminen l1-regularisoinnin sijaan johtuu siitä, että l1-regularisointi soveltuu huonosti käytettäväksi tilanteissa, joissa observaatioiden lukumäärä on pienempi kuin selittävien muuttujien lukumäärä. l1- ja l2-regularisointia yhdistävä ElasticNet-regularisointi sen sijaan edellyttää laskennallisesti kallista hyperparametrien säätämistä, eikä tästä syystä soveltunut analyysissä käytettäväksi regularisointimenetelmäksi. [41]

Analyysissä on käytetty logistisen regressiomallin optimointitehtävän ratkaisijana liblinear-kirjaston algoritmia.[42]

## 2.7 Ahne joukkoperustainen piirteevalinta-algoritmi

### 2.7.1 Piirteevalinnan tavoitteista

Piirteevalinta (engl. feature selection) on eräs piirreavaruuden dimensionaalisuuden pienentämiseen tarkoitettuista menettelyistä. “Piirteet” voidaan koneoppimisen kontekstissa ymmärtää aineistoon kuuluvia observaatioita luonnehtiviksi ominaisuuksiksi. Piirteet ovat usein numeerisia ominaisuuksia, mutta ne voivat olla myös esimerkiksi merkkijonoja, aikaleimoja tai muita kategorisia ominaisuuksia. Piirreavaruudella puolestaan tarkoitetaan matemaattista koordinaatistoa, jossa aineiston koostavat piirrevektorit voivat sijaita, ja sen dimensionaalisuuden pienentäminen tarkoittaa jonkinlaista aineiston transformaatiota pienempiulotteiseen koordinaatistoon. Piirteevalinta-algoritmit muodostavat aineiston  $X$  piirteiden joukosta  $F$  jonkin (aidon) osajoukon  $S \in F$ , mieluiten siten, että  $|S| \ll |F|$ . Piirteevalinta-algoritmit jaetaan usein kolmeen kategoriaan sen perusteella käyttäkö annettu piirteevalinta-algoritmi aineiston kohdemuuttujia piirteiden osajoukon valitsemiseen. Algoritmit, jotka käyttävät kohdemuuttujia, kuuluvat ohjatun (engl. supervised) piirteevalinnan kategoriaan, siinä missä kohdemuuttujia käyttämättömät algoritmit kuuluvat ohjaamattoman (engl. unsupervised) piirteevalinnan kategoriaan. Algoritmit, jot-

ka on suunniteltu käytettäväksi jos vain osa aineiston kohdemuuttujista on tiedossa, kuuluvat puoliohjatun (engl. semi-supervised) piirteenvallinnan kategoriaan.

Piirreavaruuden dimensionaalisuuden pienentäminen piirteenvallinnan keinoin on usein perusteltua muun muassa mallin suorituskyvyn kohentamisen ja mallin sovitamiseen kuluvan laskenta-ajan tehostamisen näkökulmasta, sekä mahdollisesti mallin tulkittavuuden parantamisen näkökulmasta. Piirteenvallinnan tavoitteeksi usein ymmärretään “optimaalisen” tai “parhaan” piirteiden osajoukon selvittämisen koneoppimismallin sovittamista varten, mutta yleisesti ei välttämättä vallitse kiistatonta yhteisymmärrystä siitä, mitä tarkalleen näillä käsitteillä tarkoitetaan; usein kuitenkin mallin luokittelusuorituskyky on keskeinen kriteeri.

Toisinaan ohjatun piirteenvallinnan kontekstissa erotetaan piirteiden malliin sisällyttämisen aiheuttama luokitteluperformanssin paraneminen ja piirteiden relevanssi, jonka määritelmästä ei myöskään ole täydellistä yhteisymmärrystä. Eräs määritelmä perustuu “vahvan” ja “heikon” relevanssin käsitteisiin, ja esittää, että vahvasti relevanttia piirrettä ei voi poistaa piirrejoukosta ilman mallin luokittelusuorituskyvyn heikkenemistä, siinä missä heikosti relevantilla piirteellä voi joissakin tapauksissa olla vaikutusta mallin luokittelusuorituskykyyn. Käytännön sovelluksissa piirteiden relevanssia voi estimoida (Pearsonin) korrelaatiokertointen ja ristientropian kaltaisten metriikoiden avulla.[43][44]

### 2.7.2 Piirteenvallinta joukkopohjaisella hakualgoritmeilla

Vastemuuttujien käytön lisäksi piirteenvallinta-algoritmien joukosta voidaan erottaa kolme perustyyppiä: suodatintyyppiset menetelmät (engl. filter), wrapper-menetelmät sekä sulautetut menetelmät (engl. embedded). Suodatintyyppisiin menetelmiin luokituvat algoritmit kuuluvat ohjaamattomaan piirteenvallintaan, ja perustuvat piirteistä laskettaviin tunnuslukuihin (esim. edellämainitut Pearsonin korrelaatio ja ristientropia) joiden arvojen perusteella jotkin piirteistä voidaan sisällyttää malliin ja

toiset hylätä. Wrapper-tyyppiset algoritmit kuuluvat ohjatun piirteevalinnan kategoriaan, ja perustuvat koneoppimismallin luokittelusuorituskyvyn evaluointiin erilaisilla piirteiden osajoukoilla, joista parhaaksi arvioitu valikoituu lopulliseen malliin. Sulautetut menetelmät kuuluvat niin ikään ohjatun piirteevalinnan kategoriaan, ja niiden toiminta ajoittuu koneoppimismallin sovitusvaiheeseen. Wrapper-tyyppisten piirteevalintamenetelmien ongelma on siinä, että optimaalisen piirrejoukon hakuvaruuden koko on eksponentiaalinen suhteessa piirteiden lukumäärään, joten koko hakuvaruuden kattava haku ei usein käytännössä ole käyttökelpoinen.[45] Lisäksi optimaalisen piirrejoukon löytymisen varmistaminen edellyttää tyhjentävää hakuvaruudessa, mikä on harvoin realistinen lähtökohta piirteevalinnalle. Heuristista strategiaa käyttävien wrapper-menetelmien on osoitettu suoriutuvan hyvin kohtuullisella laskenta-ajalla.[43]

Hakualgoritmillä tarkoitetaan sellaista proseduuria, jossa tavoitteena on löytää ratkaisu annettuun ongelmaan etsimällä hakuvaruudesta ratkaisuksi soveltuvaa tietuetta jonkin hakustrategian mukaisesti. Ahneella hakualgoritmillä (engl. greedy search algorithm) tarkoitetaan heuristista menettelyä, jossa annettu ongelma ratkaistaan osittamalla se sarjaksi osaongelmia, joista kukin ratkaistaan vuorollaan valitsemalla tarjolla olevista ratkaisuvaihtoehdoista paras mahdollinen. Ongelman kokonaisratkaisu muodostuu osaongelmien ”paikallisten” ratkaisujen muodostamasta kokonaisuudesta. Mikäli ongelman optimaalisen ratkaisun tiedetään koostuvan sen osaongelmien optimaalisista ratkaisuista, johtaa ahne hakualgoritmi ongelman optimaaliseen ratkaisuun, mutta tämä ominaisuus ei ole yleisesti pätevä.[46] Wrapper-tyyppisen piirteevalintamenettelyn hakustrategiana ahne hakualgoritmi ei näin ollen välttämättä löydä optimaalista kokonaisratkaisua, vaan jonkin lokaalisesti optimaalisen ratkaisun.

Piirteenekstraktointimenettely, jossa monikanavaisen EEG-nauhoitteen jokaisesta kanavasta generoidaan erikseen ennaltamäärätyn pituinen piirrevektori johtaa

tyypillisesti suureen joukkoon voimakkaasti keskenään korreloivia - ja näin potentiaalisesti redundantteja - piirteitä, mistä syystä voi olla aiheellista käyttää jotakin piirteevalintamenettelyä suppeamman, parhaassa tapauksessa optimaalisen piirrejoukon löytämiseksi. Koneoppimispohjaisessa EEG-tutkimuksessa kiinnostuksen kohteena voi olla luokitteluperformanssin lisäksi esimerkiksi niiden elektrodien lukumäärä ja sijainti, joiden mittaamista signaaleista valitun piirteiden osajoukon sisältämät piirteet on ekstraktoitu. Sellaisessa tapauksessa, jossa kustakin kanavasta ekstraktoidaan yhtä useampia piirteitä, edelämainitunkaltaisten toissijaisten tutkimuskysymysten tarkastelu voi vaikeutua huomattavasti jos piirteevalinta-algoritmi ei jollakin tavoin säilytä tietoa samasta kanavasta ekstraktoitujen piirteiden yhteisestä alkuperästä, vaan kohtelee kaikkia piirteitä täysin itsenäisinä. Muun muassa kiinnostus EEG-nauhoitteen sisältämästä tilallisesta ulottuvuudesta voi toimia perusteluna sille, että piirteitä voidaan haluta käsitellä ryhmittelemällä samasta kanavasta ekstraktoidut piirteet yhteen, ja joko valitsemalla ne yhdessä malliin, tai jättämällä ne sen ulkopuolelle. Tähän tarpeeseen voi soveltaa jonkinlaista ryhmittäistä piirteevalinta-algoritmia.

Ryhmittäinen piirteevalinta perustuu usein sulautetun piirteevalinnan kategoriaan kuuluvan,  $l_1$ -regularisointiin perustuvan lasso-algoritmin muunnelmiin. Vaihtoehtona lasso-algoritmille ryhmittäisessä piirteevalinnassa on esitetty kustannuspohjaista, wrapper-tyyppistä ahnetta eteenpäinhakualgoritmia (engl. greedy forward search) ryhmittäistä piirteevalintaa varten[47], johon tämän tutkielman yhteydessä tehdyssä analyysissä käytetty piirteevalinta-algoritmi perustuu. Algoritmin lähtökohtana on tyhjä valittujen piirteiden joukko, jota täydennetään iteratiivisessa silmukkarakenteensa piirreryhmä kerrallaan eniten koneoppimismallin suorituskykyä parantavilla piirreryhmillä laskevassa järjestyksessä, kunnes ennaltamäärätty valittavien piirreryhmien lukumäärää säätelevä budjetti on ehdytetty. Valittu piirreryhmä poistetaan valinnan jälkeen potentiaalisten piirreryhmien joukosta. Lopuksi malli

sovitetaan valittujen piirreryhmien sisältämistä piirteistä muodostetulla piirteiden osajoukolla. Menetelmä on laskennallisesti suhteellisen kallis, sillä prosessi, jossa jokaista potentiaalista malliin sisällytettävää piirreryhmää testataan evaluoimalla sillä täydennetyyn koneoppimismallin luokittelusuorituskykyä toistetaan iteratiivisessa silmukkarakenteessa budjetin määrittämä määrä kertoja. Menettely ei myöskään takaa optimaalisen ratkaisun löytymistä, vaan on heuristinen menetelmä ns. lokaalin optimin löytämiseksi. Budjettiperustainen, ryhmitelty, ahne eteenpäinhakualgoritmi piirteenvalintaan on esitetty ohjelmalistauksessa 2.

Tämän tutkielman kirjoittamisen yhteydessä toteutetussa analyysissä on sovellettu edellä kuvailtua budjetoitua ahnetta eteenpäinhakualgoritmia ryhmittäistä piirteenvalintaa varten. Piirteenvalintamenettely on toteutettu sisäkkäisellä 10:n laskoksen ristiinvalidointialgoritmillä siten, että sisempää ristiinvalidointiprosessia on käytetty mallin optimoimiseen parhaan piirreryhmien osajoukon löytämiseksi annetulla budjetilla, ja ulompaa ristiinvalidointiprosessia on käytetty asianmukaisesti optimoidun mallin yleistyskyvyn estimointiin. On huomattava, että  $K$ :n laskoksen ristiinvalidointialgoritmin yhteydessä toteutettu piirteenvalinta ei valitse yksikäsitteistä parasta piirreryhmien osajoukkoa (tai yleisemmin yksikäsitteistä mallia), sillä piirteenvalinta-algoritmi voi valita erilaisia piirreryhmien osajoukkoja eri validaatioaineistoilla optimoituna.

---

**Ohjelmalistaus 2** Budjettiperustainen, ryhmittäinen ahne eteenpäinhakualgoritmi

---

**Input:** Luokittelija  $Clf$ , Piirreindeksit  $F$ ,Kohdemuuttujat  $Y$ , Budjetti  $B$ , Piirteiden lukumäärä ryhmää kohden  $I$ ,Piirreryhmien lukumäärä  $J$ **Output:** Lokaalisti optimaalinen malli  $M_{opt}$  budjetilla  $B$ Piirreindeksien partitioiminen ryhmittäin:  $G_i \subset F; i = 1, \dots, J; |G_i| = I$ Alustetaan valittujen piirteiden joukko:  $S \leftarrow \emptyset$ Alustetaan koneoppimismallin hintaa merkitsevä muuttuja:  $c \leftarrow 0$ **while**  $c < B$  **do** $s \leftarrow \operatorname{argmax}_{i \in 1, \dots, J, G_i \not\subset S} (\operatorname{score}(C(S \cup G_i, Y)))$  $S \leftarrow S \cup G_s$  $c \leftarrow c + 1$  $M_{opt} \leftarrow Clf(S, Y)$ **end while****Return**  $M_{opt}$ 

---

## 3 Kirjallisuuskatsaus

### 3.1 Parkinsonin taudin koneoppimispohjainen tunnistaminen

Tässä luvussa tutustutaan kronologisessa järjestyksessä aiempaan Parkinsonin taudin koneoppimispohjaista tunnistamista käsittelevään EEG-, sekä menetelmien samankaltaisuuden vuoksi soveltuvilta osin myös MEG-tutkimukseen. Aihetta koskevat analyysit eroavat oleellisesti useiden tekijöiden, kuten analyysiin käytetyn aineiston määrän ja laadun, analyysin metodologian (esim. esikäsittelymenetelmät, piirteengenerointimenetelmät, luokittelualgoritmit sekä validointimenettely) ja osin analyysin tavoitteidenkin osalta, eivätkä eri analyysien raportoidut tulokset ole välttämättä täysin vertailukelpoisia. Voi esimerkiksi olla syytä epäillä, että hyvin pienellä otoskoolla saavutettu korkea testijoukon luokittelutarkkuus ei täysin yleisty suuremmalle otoskoolle. Olettaen, että raportoidut tulokset on tuotettu asianmukaisesti out-of-sample -evaluointimenetelmää käyttäen, voidaan lähtökohtaisesti ajatella, että mitä suurempi analyysissä käytetty otoskoko on, sitä luotettavampana tuloksia voi pitää. Sitä, miten suuri vaikutus eri laboratorioissa kerättyjen EEG-aineistojen yhdistämisellä on testijoukon luokittelutarkkuuteen (tai -AUC-estimaattiin), ei kuitenkaan vaikuta olevan systemaattisesti tutkittu. Huomattavasti analyyseissä ei ole systemaattisesti tutkittu kysymystä siitä, kuinka pienellä elektrodivalikoimalla luokittelu onnistuu korkealla luokittelutarkkuudella. Koska analyysit eroavat niin mo-

nessa suhteessa toisistaan, on tässä luvussa pyritty referoimaan mahdollisimman tiiviisti kunkin analyysin oleelliset seikat. Luvun lopussa esitetään yhteenveto edeltävästä tutkimuksesta. Käsittelyn ulkopuolelle on jätetty erilaisia hermoverkkoarkkitehtuureja käyttävät viimeaikaiset Parkinsonin taudin tunnistamiseen liittyvät EEG-analyysit (esim. [48], [49]) metodologian eroavuuden, sekä tulosten arvioinnin vaikeuden vuoksi. Näille analyyseille on ominaista huomattavan korkeat testipermanssitulokset (luokittelutarkkuus testijoukolla jopa  $> 0.99$ ), sekä varsin pienet otoskoot ( $|N| < 50$ ). Eräs mielenkiintoinen tutkimuskysymys voisikin olla hermoverkko-pohjaisten EEG-analyysien toistettavuuden tutkiminen mahdollisesti suuremmalla otoskoolla. Tämän lisäksi katsauksen ulkopuolelle on jätetty jonkinlaista koehenkilön aktiivisuutta edellyttävät EEG-koeasetelmat, sillä on epäselvää miten tulokset vertautuisivat lepotilassa kerätyn EEG-aineiston kanssa.

Stoffers et al. (2008) on tilastollinen MEG-analyysi, jonka aineisto koostuu 70:n PD-potilaan ja 21:n terveen kontrollihenkilön silmät kiinni lepotilassa kerätystä 151-kanavaisesta MEG-tallenteesta. Jokaisesta tallenteesta valittiin kolme noin 13:n sekunnin mittaista mahdollisimman häiriötöntä epookkia, jotka joista erotettiin digitaalilla suodattimella delta- (0.5–4.0 Hz), theta- (4.0–8.0 Hz), alpha1- (8.0–10.0 Hz), alpha2- (10.0–13.0 Hz), beta- (13.0–30.0 Hz) ja gammakaistat (30.0–48.0 Hz). Tallenteiden kanavat ryhmiteltiin aivoalueittain (kummankin hemisfääriin otsalohko, keskialue, ohimolohko, päälaenlohko ja takaraivolohko), missä yhteydessä yhdeksän aivojen keskilinjalle sijoittuvaa kanavaa, ja yksi takaraivolohkolle sijoittuva kanava hylättiin. Tämän esiprosessoinnin jälkeen aineistossa oli jäljellä 141 kanavaa jaetuna kymmenelle aivoalueelle. Kanavapareille laskettiin jokaista epookkia kohden synkronisaatiokovettavuus (jatkoissa SL) aivoalueiden sisällä, samalle hemisfäärille kuuluvien aivoalueiden välillä, sekä eri hemisfääreille kuuluvien aivoalueiden välillä. Kaikkien näin saatujen kanavaparien SL-arvot keskiarvoistettiin aivoalueittain ja epookeittain. Näin ekstraktoidut piirre-arvot pyrittiin normalisoimaan, minkä jäl-

keen kuitenkin havaittiin etteivät deltakaistaa vastaavat SL-arvot ole normaalisti jakautuneet, ja ne hylättiin. Näin tuotetuilla piirteillä suoritettiin neljä tilastollista analyysiä (A-D) tilastollisella merkittävyydellä  $p = 0.05$ . Analyysissä A tutkittiin hoitamattoman varhaisvaiheen Parkinsonin taudin näkymistä SL-arvoissa, ja potilailla havaittiin korkeampia paikallisia, hemisfäärien sisäisiä (jatkossa intrahemisfäärisiä) ja hemisfäärien välisiä (jatkossa interhemisfäärisiä) alpha1-kaistan SL-arvoja kuin kontrollihenkilöillä. Analyysissä B tutkittiin Parkinsonin taudin keston ja motorisen toimintakyvyn yhteyttä, ja havaittiin yhteys taudin keston ja lokaalisen, intrahemisfäärisen ja interhemisfäärisen alpha2-kaistan sekä lokaalisen thetakaistan SL-arvojen välillä. Lisäksi havaittiin yhteys UPDRS-arvon<sup>1</sup> ja lokaalisen, intrahemisfäärisen ja interhemisfäärisen thetakaistan sekä interhemisfäärisen betakaistan SL-arvojen välillä. Analyysissä C tutkittiin keskipitkälle edenneen taudin näkymistä SL-arvoissa, ja potilailla havaittiin korkeampia SL-arvoja paikallisilla, intrahemisfäärisillä ja interhemisfäärisillä thetakaistolla, paikallisilla ja intrahemisfäärisillä alpha1-kaistoilla, paikallisilla ja interhemisfäärisillä alpha2-kaistoilla sekä paikallisilla betakaistoilla. Analyysissä D tutkittiin hoitamattoman varhaisvaiheen Parkinsonin taudin vaikutuksia kognitiiviseen toimintaan ja havaittiin yhteys kognitiivisen perseveranssin<sup>2</sup> ja interhemisfääristen alpha1-kanavan SL-arvojen kanssa. Tämän tutkielman kannalta erityisen oleellisenä tuloksena voi pitää analyysin A tulosta varhaisvaiheen PD-potilaiden SL-arvojen eroamista terveiden kontrollihenkilöiden vastaavista arvoista. [21]

Gomez et al. (2011) on koneoppimispohjainen MEG-analyysi, jossa aineistona on 18 varhaisvaiheen PD-potilaan ja 20 kontrollihenkilön 151-kanavaiset lepotilassa kerätyt MEG-tallenteet. Analyysissä käytetyt MEG-tallenteet kerättiin 312.5 Hz näytteenottotaajuudella käyttäen 0.25-125 Hz laitteistotason kaistapäästösuo-

---

<sup>1</sup>UPDRS-arvolla (Unified Parkinson Disease Rating Scale) arvioidaan Parkinsonin taudin edistymistä ja vakavuutta.

<sup>2</sup>Kognitiivisella perseveranssilla tarkoitetaan taipumusta ajatella jatkuvasti kielteisiä asioita.

datinta. Esiprosessointina jokaista testihenkilöä kohden valittiin 12 neljän sekunnin artefaktitonta epookkia, ja jokainen epookki suodatettiin äärellisen impulssivasteen -kaistanpäästösuodattimella (engl. finite impulse response filter, lyh. FIR) käyttäen Hamming-ikkunaa<sup>3</sup> ja rajataajuuksia 0.5 Hz ja 45 Hz. Piirteenekstraktointivaiheessa datalle laskettiin kanavakohtaiset Lempel-Ziv -kompleksisuusarvot (jatkossa LZC).<sup>4</sup> sekä kahden että kolmen symbolin konversiolla<sup>5</sup>. Kanavakohtaiset LZC-kompleksisuusarvot laskettiin epookkikohtaisten kompleksisuusarvojen keskiarvona. Lopuksi koneoppimismallin yleistyskykyä arvioitiin käyttäen luokittelualgoritmina lineaaridiskriminanttianalyysiä ja out-of-sample -evaluointiin leave-one-out -ristiinvalidointialgoritmia. Analyysin yhteydessä PD-positiivisten koehenkilöiden LZC-kompleksisuusarvojen (sekä kaksi- että kolmesymbolisten) huomattiin olevan alhaisempia kuin terveiden kontrollihenkilöiden. Mallin rakentamiseen käytettiin askelittaista LDA-algoritmia, jossa malliin valittiin inkrementaalisesti aivoalue kerrallaan luokittelutarkkuutta eniten parantavat piirteet.<sup>6</sup> Valmis malli saavutti luokittelutarkkuuden 84.21%. (sensitiivisyys 77.78%, spesifisyys 90.0%). Tutkimuksessa huomattiin lisäksi, että erikseen arvioituina aivokuoren alueista eniten luokittelutarkkuutta parantavaa informaatiota sisältyi oikean ohimolohkon alueeseen (luokittelutarkkuus 81.58%), jota seurasivat oikea päälakilohko (78.95%) ja vasen

---

<sup>3</sup>ks. esim. [50]

<sup>4</sup>Menetelmä perustuu signaalin kvantisointiin hyvin matalaan resoluutioon käyttäen jotakin kynnysarvoa tai arvoja, joiden mukaan aikasarjan arvoille määrätään vastaava symboli (esim. 0, 1) sen mukaan, onko aikasarjan arvo pienempi vai suurempi kuin kynnysarvo. Tästä uudesta kvantisoidusta aikasarjasta lasketaan erilaisten alasekvenssien normalisoitu lukumäärä, joka palautetaan aikasarjan kompleksisuusestimaattina.

<sup>5</sup>Kahden symbolin konversion kynnysarvona oli käytetty aikasarjan mediaania. Kolmen symbolin konversion kynnysarvoina oli puolestaan käytetty aikasarjan mediaanista  $x_m$ , minimistä  $x_{min}$  ja maksimista  $x_{max}$  laskettuja arvoja  $T_{d1} = x_m - |x_{min}|/16$  ja  $T_{d2} = x_m + |x_{max}|/16$

<sup>6</sup>Ilmeisesti tässä yhteydessä LDA-luokittelija optimoitiin käyttäen sisäkkäistä leave-one-out -algoritmia. Artikkelin on tämän tärkeän yksityiskohdan suhteen hieman epäselvä.

keskeinen alue (76.32%). Vähiten luokittelutarkkuuta parantavaa informaatiota sisältyi vasemman takaraivolohkon alueelle (63.16%), vasemman etulohkon alueelle (65.79%) ja oikean etulohkon alueelle (65.79%). Kolmea symbolia käyttävä LZC-konversio suoriutui yleisesti paremmin kuin kahta symbolia käyttävä. [51]

Lainscsek et al. (2013) on tilastollinen EEG-analyysi, jonka piirteenekstraktointimenettely perustuu EEG-signaalin dynamiikan analyysiin viivedifferentiaaliyhtälöiden (engl. delay differential equation, jatkossa DDE. ks. [52]) avulla. Tutkimuksen aineistona on yhdeksältä lievästä tai keskivaikeasta Parkinsonin taudin muodosta kärsivää potilasta ja yhdeksältä kontrollihenkilöltä kerätyt 64-kanavaiset EEG-tallenteet käyttäen 512 Hz:n näytteenottotaajuutta. PD-potilailta kerättiin tutkimusta varten EEG-tallenteet sekä lääkityksen alaisena että lääkitsemättömänä. Analyysissä käytettiin näistä EEG-tallenteista valikoituja yhden sekunnin mittaisia lepotilassa kerättyjä epookkeja<sup>7</sup>, jotka esiprosessoitiin poistamalla matalataajuuksinen ns. drift-siirtymä ja sähkölinjan aiheuttama häiriökaistanpäästösuodattimella rajataajuuksina 1 Hz ja 55 Hz sekä poistamalla silmänliikkeistä (yms. lähteistä) aiheutuvat artefaktit itsenäisten komponenttien analyysillä (engl. independent component analysis, lyh. ICA. ks. esim. [53]). Esiprosessoinnin lopuksi tallenteen 64 kanavaa jaettiin kymmeneen ryhmään elektrodien sijainnin perusteella, ja kustakin ryhmästä laskettiin signaalin amplitudin keskiarvo. Piirteenekstraktointi tapahtui sovittamalla esiprosessoiduilla epookeilla erilaisia DDE-malleja, jotka vaihtelivat viiveparametrin arvon ja termien korkeimman asteen suhteen. Jokaista koehenkilöä kohden sovitettiin 50 erilaista mallia. Varsinaisina piirteinä käytettiin mallien kertoimia ja vastaavia pienimmän neliösumman virhetermejä. Luokittelijana käytettiin tilastotieteellistä lineaarista sekamallia (engl. linear mixed model, ks. esim. [54]). Luokittelijan sovittamiseen käytettiin kuuden henkilön dataa kummastakin luokasta (yh-

---

<sup>7</sup>Tarkemmin sanottuna epookit eivät olleet kerätty täysin lepotilassa, vaan aktiivisen koeasetelman yhteydessä koehenkilöiden odottaessa merkkiä tarttumisliikkeen tekemiseen kädellään.

teensä 600 epookkia) ja testaamiseen kolmen henkilön dataa kummastakin luokasta (yhteensä 300 epookkia). Luokittelija erotti PD-potilaat kontrollihenkilöistä hyvällä tarkkuudella, saavuttaen lääkitsemättömien potilaiden datalla AUC-estimaatin, joka vaihteli välillä 0.95 – 1.0, ja lääkittyjen potilaiden datalla 0.97 – 1.0. Luokittelija lisäksi erotti lääkityt potilaat lääkitsemättömistä saavuttaen AUC-estimaatin, joka vaihteli välillä 0.86 – 1.0. Vastaavat AUC-estimaatit olivat hieman matalampia testattuna eri epookkien välillä.[55]

Yuvaraj et al. (2016) on koneoppimispohjainen EEG-analyysi, jossa aineistona on 20:n varhais- ja keskivaiheen PD-potilaan ja 20:n samanikäisen terveen kontrollihenkilön 5-minuuttia pitkät 14-kanavaiset EEG-nauhoitteet. Nauhoitteista poistettiin silmänliikkeiden aiheuttamat häiriöt kynnystystekniikalla, ja muut häiriökomponentit eliminoitiin kuudennen kertaluvun Butterworth-kaistanpäästösuodattimella, raja-arvoina 1 Hz ja 49 Hz, jonka jälkeen nauhoitteet ositettiin kahden sekunnin epookeihin jatkokäsittelyä varten. Epookeista ekstraktoitiin joukko bispektripohjaisia<sup>8</sup> tunnuslukupiirteitä. Piirteiden ekstraktoinnin jälkeen piirteet asetettiin niiden erottelukyvyn perusteella tärkeysjärjestykseen Studentin t-testin avulla. Tätä järjestystä seuraten malliin sisällytettiin piirteitä yksi kerrallaan kunnes korkein mahdollinen luokittelutarkkuus on saavutettu. Näitä piirteitä käyttäen sovitettiin joukko erilaisia koneoppimismalleja, jotka olivat päätöspuu, sumea k:n lähimmän naapurin menetelmä, k:n lähimmän naapurin menetelmä, naiivi Bayesin luokitin, probabilistinen hermoverkkomalli sekä tukivektorikone lineaarisella ytimellä, 2. ja 3. asteen polynomisella ytimellä sekä radiaaliantafunktioytimellä (SVM-RBF). Mallin suorituskyvyn raportointiin käytettiin sensitiivisyyttä, spesifisyyttä, presisiota, F-arvoa sekä luokittelutarkkuutta, ja koneoppimismallien suorituskyvyn arviointiin annetuil-

---

<sup>8</sup>Bispektri on signaalin epälineaaristen interaktioiden analysointiin käytetty menetelmä. Korkeamman asteen spektrien (engl. Higher order spectra, lyh. HOS) laskemisesta ja käytöstä signaalien analysoinnissa tarkemmin esim. [56]

la metriikoilla käytettiin 10:n laskoksen ristiinvalidointialgoritmia. Kaikki koneoppimismallit saavuttivat huomattavan korkean luokittelutarkkuuden (0.887 - 0.996), parhaana SVM-RBF. Artikkelin kirjoittajat mainitsevat mahdollisena rajoitteena tutkimuksen otoskoon, joka koostui vain 20:stä PD-positiivisesta koehenkilöstä ja 20:stä terveestä kontrollihenkilöstä.[57] Koska analyysissä suoritetaan datan partioiduista edeltävänä vaiheena piirteidenvalintamenettely, joka vaikuttaa implisiittisesti käytettävän tietoa kohdemuuttujista, on epäselvää voiko raportoituja tuloksia pitää täysin edustavina mallin todellisen yleistyskyvyn suhteen.

Liu et al. (2017) on koneoppimispohjainen EEG-analyysi, jonka aineisto koostuu 17:n PD-potilaan ja 25:n kontrollihenkilön silmät auki lepotilassa näytteenottotaajuudella 250 Hz kerätyistä 10-kanavaisista EEG-tallenteista. Analyysistä kirjoitettussa artikkelissa ei kuvailla tallenteiden esiprosessointimenettelyä. Tallenteille suoritettiin kolmitasoinen diskreetti aallokemuunnos (engl. discrete wavelet transform, jatkossa DWT. ks. esim. [58]), jonka avulla tallenteista erotettiin joukko taajuuskaistoja. DWT jakaa signaalin kahteen taajuuskaistaan taajuusspektrin keskipisteen suhteen niin, että niin kutsuttu detail-komponentti (lyh. "D") sisältää taajuusspektrin keskipistettä korkeammat taajuudet ja approximation-komponentti (lyh. "A") sisältää taajuusspektrin keskipistettä matalammat taajuudet. Näin kolmitasoinen DWT jakaa signaalin, jonka näytteenottotaajuus on 250 Hz, kolmeen detail- ja kolmeen approximation-komponenttiin. Kolmitasoisien DWT-muunnoksen komponentit olivat  $D_1 = 125 - 250$  Hz,  $D_2 = 62.5 - 125$  Hz,  $D_3 = 31.25 - 62.5$  Hz ja  $A_3 = 0 - 31.25$  Hz. Artikkelin pohjalta jää osin epäselväksi, mitä kaikkia DWT:llä ekstraktoiduista komponenteista käytettiin piirteinä. Ilmeisesti piirteinä on käytetty ainoastaan jokaisen EEG-tallenteen kanavan  $A_3$ -komponentin (0 - 31.25 Hz) näyte-entropiaa, joten DWT toimii tässä yhteydessä ikään kuin alipäästösuodattimen tavoin. DWT:n avulla käsitellyistä aikasarjoista laskettiin sitten näyte-entropia -kompleksisuusarvot, joita analyysissä käytettiin piirteinä. Luokitteluun käytettiin uudenlaista kolmiarvoista

Optimal center constructive covering algorithm -päätosalgoritmia (*O\_CCA*), joka luokittelee annetun tapauksen joko positiiviseksi, negatiiviseksi tai rajatapaukseksi, jolloin tapausta ei lasketa kumpaankaan luokkaan kuuluvaksi. Päätosalgoritmia testattiin 10:n laskoksen ristiinvalidointialgoritmillä sekä datalla, jonka piirteenekstraktoinnin välivaiheena oli käytetty kolmitasoista DWT-muunnosta, että datalla jonka piirteenekstraktointiin ei ollut käytetty DWT:tä. *O\_CCA*-päätosalgoritmin ristiinvalidoituja testituloksia verrattiin sitten neljään perinteiseen luokittelualgoritmiin (Tukivektorikone, K:n lähimmän naapurin menetelmä, naiivi Bayesin luokitin ja satunnaismetsäluokitin). DWT-muunnosta välivaiheena käyttäen *O\_CCA*-päätosalgoritmi saavutti 92.86%:n luokittelutarkkuuden, siinä missä vastaava luokittelutarkkuus kun DWT-muunnosta ei oltu käytetty piirteenekstraktoinnin välivaiheena jäi 88.10%:iin. Verrokkialgoritmeista tukivektorikone saavutti 97.62%:n luokittelutarkkuuden, K:n lähimmän naapurin menetelmä saavutti 90.48%:n luokittelutarkkuuden, Naiivi Bayesin luokitin saavutti 88.10%:n luokittelutarkkuuden ja satunnaismetsäluokitin saavutti 83.33%:n luokittelutarkkuuden. On syytä huomioda, että luokittelutulokset kaikilla menetelmillä ovat varsin korkeita, ja luokittelualgoritmeista parhaiten suoriutunut tukivektorikone luokitteli vain yhden tapauksen väärin. Artikkelissa ei mainita tutkimukseen osallistuneiden PD-potilaiden sairauden edistymisastetta, joten on mahdollista että kyse on suhteellisen pitkälle edistyneestä sairaudesta, jolloin luokittelun voi odottaa onnistuvan huomattavasti korkeammalla tarkkuudella. *O\_CCA*-päätosalgoritmin etu paremmin suoriutuneeseen tukivektorikonealgoritmiin verrattuna on siinä, että se mahdollistaa epäselvien tapausten luokittelun omaan kategoriaansa, jolloin voidaan välttää vääriä päätöksiä, tosin jää epäselväksi miten näitä tapauksia on kohdeltu mallin testisuorituskykyä arvioitaessa. [59]

Chaturvedi et al. (2017) on koneoppimispohjainen EEG-analyysi, jossa tutkimuskysymyksenä oli paitsi luokittelutarkkuuden havainnointi, myös luokittelun kan-

nalta tärkeimpien piirteiden tunnistaminen. Koska analyysin suhteellisen yksinkertaisella perinteisten taajuuskaistojen suhteellisiin amplitudeihin perustuvalla piirteekstraktointimenettelyllä tuotetut piirteet ovat helposti tulkittavissa, analyysi voi myös auttaa ymmärtämään Parkinsonin taudin aiheuttamia muutoksia aivojen toiminnassa. Tutkimuksen aineisto koostui 50:n PD-potilaan<sup>9</sup> ja 41:n vastaavan ikäisen kontrollihenkilön levossa, silmät kiinni kerätyistä 256-kanavaisista EEG-nauhoitteista. 12:n minuutin kestoista EEG-nauhoitteista valittiin yhteensä kolmen minuutin verran vähintään 30 sekunnin mittaisia artefaktittomia segmenttejä, jotka alinäytteistettiin 500 Hz:iin.<sup>10</sup> 214:n kanavan (mm. kaulan ja poskien elektrodit jätettiin tutkimuksen ulkopuolelle) signaalit suodatettiin välille 0.5 – 70.0 Hz, lyhyet segmentit yhdistettiin käyttäen käänteistä Hanning-ikkunaa, ja data uudelleenreferoitiin signaalien keskiarvoon. Esiprosessoinnin päätteeksi datasta poistettiin loputkin artefaktit itsenäiskomponenttianalyysillä. Edellä mainitusti esiprosessoitujen kanavien voimaspektri analysoitiin Welchin metodilla<sup>11</sup>, ja voimaspektristä laskettiin delta- (1 – 4 Hz), theta- (4 – 8 Hz), alpha1- (8 – 10 Hz), alpha2- (10 – 13 Hz) ja betakaistojen (13 – 30 Hz) suhteelliset amplitudit jakamalla kunkin kaistan absoluuttinen amplitudi koko signaalin (1 – 30 Hz) amplitudilla, minkä jälkeen samoille aivoalueille (kummankin hemisfääriin etulohko, keskialue, päälakilohko, ohimolohko, ja takaraivolohko) kuuluvien kanavien taajuuskaistojen suhteelliset amplitudit aggregoitiin häiriön eliminoimiseksi. Piirteitä ekstraktoitiin lopulta 79, mukaan lukien jokaisen taajuuskaistan suhteellinen kokonaisamplitudi, jokaisen taajuuskaistan aluekohtainen suhteellinen amplitudi, aluekohtaiset alpha1/theta-suhdeluvut, sekä mediaanit ja maksimit. Analyysiin käytettiin neljää koneoppimisalgoritmia, jotka oli-

---

<sup>9</sup>Alun perin tutkimukseen osallistui 68 PD-potilasta, joista yhdeksän karsiutui muiden terveydentilaan liittyvien seikkojen ja yksi onnettomuuden takia. Kahdeksan PD-potilasta karsiutui heikkolaatuisten EEG-nauhoitteiden takia.

<sup>10</sup>Alkuperäistä näytteenottotaajuutta ei mainita tutkimusartikkelissa.

<sup>11</sup>Menetelmä signaalin spektritiheyden estimointiin. ks. esim. [60]

vat l1-normilla regularisoitu logistinen regressio, päätöspuuluokittelija<sup>12</sup>, tukivektorikone ja satunnaismetsäluokittelija. Mallien suorituskyky estimoitiin 10:n laskoksen ristiinvalidointialgoritmillä, minkä lisäksi sisäkkäistä 10:n laskoksen ristiinvalidointialgoritmia käytettiin logistisen regressiomallin hienosäätöparametrin optimointiin. Malleista logistinen regressio ja satunnaismetsäluokittelija käyttävät sisäänrakennettua piirteenvälyntää, ja niiden avulla voi siten vastata analyysin toissijaiseen tutkimuskysymykseen tärkeimpien piirteiden tunnistamisesta. Logistisen regression ristiinvalidoitu suorituskyky oli joukon matalin (ACC: 0.56, AUC 0.63), ja satunnaismetsäluokittelijan joukon korkein (ACC: 0.78, AUC 0.8), siinä missä päätöspuu (ACC: 0.68, AUC: 0.67) ja tukivektorikone (ACC: 0.747, AUC: 0.73) asettuivat ääripäiden väliin. Molemmat piirteenvälyntää suorittavista menetelmistä nostivat esiin vasemman aivopuoliskon ohimolohkon theta-alueen sekä vasemman keskialueen ja vasemman ohimolohkon alueen  $\alpha_1/\theta$ -suhdeluvut.[61]

Anjum et al. (2019) on koneoppimispohjainen EEG-analyysi, jossa tutkittiin uudenlaisen LEAPD-luokittelualgoritmin suorituskykyä Parkinsonin taudin tunnistamisessa. Analyysin aineisto koostui kahdessa eri laboratorioissa 500 Hz:n näytteenottotaajuudella nauhoitetuista kaksi minuuttia pitkistä 64-kanavaisista EEG-tallenteista, joita kaikkiaan oli 82:lta testihenkilöltä (41 PD-positiivista ja 41 tervettä kontrollitapausta). Tallenteista 54 (27 PD-positiivista ja 27 tervettä kontrollitapausta) oli nauhoitettu New Mexicon yliopiston laboratorioissa, ja 28 (14 PD-positiivista ja 14 tervettä kontrollitapausta) Iowan yliopiston laboratorioissa. Vain New Mexicon yliopiston laboratorioissa nauhoitettiin koehenkilöiden EEG-tallenteet sekä silmät auki että silmät kiinni, siinä missä Iowan yliopiston laboratorioissa nauhoitettiin koehenkilöiden EEG-tallenteet vain silmät auki. Laboratoriot myös käyttivät osin eri standardeja elektrodien sijoituksen suhteen, joten poikkeavia elekt-

---

<sup>12</sup>Tarkemmin sanottuna J48-toteutus C4.5-pätöspuualgoritmista. Mallin dokumentaatio on nähtävissä <https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html>

rodeja oli karsittava yhteisen standardin määrittämiseksi, minkä jälkeen jäljellä oli 62 kanavaa. Esikäsittelyn yhteydessä tallenteista poistettiin silmänräpäykset yms. artefaktit käyttäen itsenäiskomponenttianalyysiä. Ennen piirteenekstraktointia aineistolle suoritettiin signaalien energian normalisointi <sup>13</sup> ja kaistanpäästösuodatus taajuuksille 2.5 Hz – 14.0 Hz.<sup>14</sup> Piirteenekstraktointi perustui lineaariseen ennustekoodaukseen (engl. Linear predictive coding, lyh. LPC) käyttäen Burgin metodia (maksimaalisen entropian menetelmä spektritiheyden estimointiin, ks. Esim. [62]) Tätä piirteenekstraktointimenetelmän etuna voidaan pitää sitä, että menetelmä käyttää koko taajuusspektriä ad hoc -tyyppisesti määriteltyjen taajuuskaisojen sijaan. Näin tuotettujen piirrevektorien dimensionaalisuutta alennettiin vielä pääkomponenttianalyysin avulla. (engl. Principal component analysis, lyh. PCA. ks. esim. [63]) Yksikanavaisen nauhoitteen luokittelu perustuu ns. LEAPD-indeksin laskemiseen siten, että kynnyksarvona toimi 0.5. Kynnyksarvoa pienemmät indeksin arvot merkitsivät luokittelua negatiiviseen ja suuremmat positiiviseen luokkaan. Monikanavaisen aineiston tapauksessa laskettiin kanavien LEAPD-indeksien keskiarvo, joka toimi kokonaisuusteena. Mallin hyperparametrit, kuten pääkomponenttien lukumäärä ja LPC-mallin kertaluku, valittiin sisäkkäisellä ristiinvalidointimenettelyllä. Koneoppimismallin sovittamiseen käytettiin New Mexican yliopistossa talletettua aineistoa, ja mallin evaluointi toteutettiin sekä toistetun k:n laskoksen ristiinvalidointialgoritmeilla, käyttäen vuorollaan viittä ja kymmentä laskosta, leave-one-out -ristiinvalidointialgoritmeilla, sekä lopuksi käyttämällä Iowan yliopistossa nauhoitettua aineistoa testiaineistona. Mali sovitettiin ja testattiin vuorollaan sekä silmät auki että silmät kiinni nauhoitettuja tallenteita, vain silmät auki nauhoitettuja tallentei-

---

<sup>13</sup>Diskreetin signaalin energia lasketaan kaavalla

$$E = \sum_{n=-\infty}^{\infty} |x(n)|^2 \quad (3.1)$$

<sup>14</sup>Myöhemmin tutkimusraportissa mainitaan päästökaistan ylärajaksi 12 Hz.

ta ja vain silmät kiinni nauhoitettuja tallenteita. Sekä silmät auki että silmät kiinni nauhoitetuilla tallenteilla sovitettua mallia testattiin vielä erikseen käyttäen testiaineistona Iowan yliopistolla nauhoitettuja tallenteita. Molemmiin tavoin (sekä silmät auki että silmät kiinni) nauhoitetulla aineistolla sovitettu ja testattu malli saavutti ristiinvalidoidun 0.853 testiluokittelutarkkuuden ja 0.933 testi-AUC-estimaatin, siinä missä vain silmät auki nauhoitetulla aineistolla sovitettu ja testattu malli saavutti ristiinvalidoidun 0.787 testiluokittelutarkkuuden ja vain silmät kiinni nauhoitetulla aineistolla sovitettu ja testattu malli saavutti ristiinvalidoidun 0.822 testiluokittelutarkkuuden. Molemmiin tavoin nauhoitetulla aineistolla sovitettu malli saavutti Iowan yliopistolla nauhoitetulla aineistolla testattaessa 0.857 testiluokittelutarkkuuden ja 0.852 testi-AUC-estimaatin. Oikean puolen ohimo-pääläenlohkon alueille, pääläenlohkon alueille, pääläenlohkon-takaraivohkon alueille ja takaraivohkon alueille kuuluvien sekä vasemman puolen etulohkon-keskialueen ja keskialueen-pääläenlohkon alueelle kuuluvien elektrodien sisällyttämisen malliin havaittiin lisäävän eniten luokittelijan suorituskykyä. [64]

## 3.2 Lievä kognitiivinen heikentyminen ja EEG

Huomattava osa viimeaikaisesta Parkinsonin tautia käsittelevästä EEG-tutkimuksesta näyttää kohdistuvan taudin ei-motoriseen lievänä kognitiivisena heikentymisenä tunnettuun oireeseen. Lievä kognitiivinen heikentyminen (engl. mild cognitive impairment, lyh. MCI) tarkoittaa tilaa, jossa kognitiiviset kyvyt ovat iän ja koulutustaustan asettamia odotuksia heikommat, mutta eivät siinä määrin että päivittäinen elämä häiriintyisi merkittävästi.[65] Tutkimuksissa on sekä pyritty erottamaan luokittelemaan observaatiot MCI-tapauksiin ja tavallisiin PD-potilaisiin[66][67], että arvioimaan kognitiivisen heikentymän astetta.[68][69] MCI:n tunnistamiseen liittyvä koneoppimispohjainen EEG-tutkimus näyttää käyttävän paljolti samankaltaisia piirteitä kuin Parkinsonin taudin tunnistamiseen liittyvä, kuten EEG-kanavien

taajuuskaistojen amplitudit [67][69] ja konnektiivisuuspiirteet[66]. MCI:n tunnistaminen vaikuttaa tulosten perusteella (AUC = 0.74 - 0.77)[66][67] vaativammalta tehtävältä kuin Parkinsonin tautia sairastavien koehenkilöiden erottaminen terveistä kontrollihenkilöistä. Kognitiivisen heikentymisen asteen ennustaminen puolestaan vaikuttaa tulosten perusteella (ACC = 0.86 - 0.92), jossakin määrin helpommalta tehtävältä, tosin korkein luokittelutarkkuusarvio oli tuotettu tilanteessa, jossa ennustettiin 20:n vähiten kognitiivisesti heikentyneen ja 20:n eniten kognitiivisesti heikentyneen väliltä.[69][68] Otokoot vaihtelivat analyyseissä 40:n ja 118:n välillä, ollen samaa mittakaavaa PD:n tunnistamistehtävän otoskokojen kanssa.

### 3.3 Kirjallisuuden yhteenveto ja keskustelua

Edellä referoitu Parkinsonin taudin tunnistamiseen keskittyvä kirjallisuus sisältää sekä EEG- että MEG-analyysejä, joiden menetelmät ja tulokset vaikuttavat olevan keskenään samantyyppisiä. Aineistojen koot vaihtelevat 18:n ja 91:n välillä.<sup>15</sup> Vaikuttaa siltä, että analyyseissä on pyritty suurin piirtein tasapainottamaan PD-positiivisten tapausten ja terveiden kontrollihenkilöiden lukumäärät aineistossa, mutta tarkalleen yhtä montaa positiivista ja negatiivista tapausta on tarkastelluista analyyseistä käyttäneet vain Lainscsek et al. (2013), Yuvaraj et al. (2016) ja Anjum et al. (2019) Luokkaepätasapainon vallitessa suorituskykyestimaatin ilmoittamiseen on syytä kiinnittää huomiota, sillä luokittelutarkkuus määriteltynä oikein luokiteltujen tapausten suhteena sekä oikein että väärin luokiteltuihin tapauksiin ei ole välttämättä täysin

---

<sup>15</sup>Stoffers et al. (2008) aineisto on koostaan huolimatta niin epätasapainoinen luokkien suhteen (70 PD-positiivista, 21 tervettä kontrollihenkilöä), että ainoastaan analyysissä C, jossa tarkastellaan vain aineiston PD-positiivisia tapauksia on käytössä koko PD-positiivisten tapausten joukko. Kaikissa analyyseissä joissa jollakin tavoin verrataan PD-positiivisia tapauksia terveisiin kontrollihenkilöihin aineiston koko on korkeintaan 40.

Tutkimus	POS/NEG	Tyyppi	Piirteet	Luokittelija	Suorituskyky
Stoffers et al. (2008)	70/21	MEG	SL	N/A	N/A
Gomez et al. (2011)	18/20	MEG	LZC	LDA	ACC 0.84
Lainscsek et al. (2013)	9/9	EEG	DDE	LMM	AUC 0.95-1.0
Yuvaraj et al. (2016)	20/20	EEG	Useita	Useita	ACC 0.89 - 1.0
Liu et al. (2017)	17/25	EEG	SampEn	O_CCA (ym.)	ACC 0.93 (0.83 - 0.98)
Chaturvedi et al. (2017)	50/41	EEG	Useita	Useita	ACC 0.56 - 0.78 AUC 0.63 - 0.8
Anjum et al. (2019)	41/41	EEG	LPC.	LEAPD	ACC 0.85 AUC 0.93

Taulukko 3.1: Yhteenvedot käsitellyistä EEG-tutkimuksista sisältäen otoskoon ja aineiston jakauman, aineiston tyyppin (EEG/MEG), piirteekstraktointialgoritmin, luokittelualgoritmin ja luokittelijan suorituskykyestimaatit.

luotettava metriikka epätasapainoisessa tilanteessa.<sup>16</sup> Gomez et al. (2011) näyttääkin mahdollisesti ilmentävän jonkinasteista enemmistöluokkaan kohdistuvaa vinoumaa, sillä raportoiduissa tuloksissa ilmenee yli kymmenen prosenttiyksikön ero testiaineiston sensitiivisyys- (0.778) ja spesifisyysmetriikoiden (0.90) välillä. AUC sitä vastoin soveltuu käytettäväksi tilanteissa, joissa luokkien välillä vallitsee epätasapaino, ja lähes kaikissa (pl. edellä mainittu Gomez et al. (2011)) referoiduissa analyyseissä onkin raportoitu testiaineiston AUC-estimaatti.

Referoiduista tutkimusraporteista Stoffers et al. (2008), Gomez et al. (2011), Lainscsek et al. (2013) ja Yuvaraj et al. (2017) ilmoittavat PD-positiivisten koehenkilöiden taudin vaiheen tai vakavuuden, siinä missä loput referoiduista tutkimusraporteista eivät niin tee. Tieto voi olla sikäli oleellinen, että pitkälle kehittyneen tai vaikean Parkinsonin taudin tunnistaminen EEG-tallenteen perusteella voi olla merkittävästi helpompaa kuin varhaisessa vaiheessa olevan taudin. On osin epäselvää missä määrin analyyseiden tulokset ovat vertailukelpoisia tilanteessa, jossa vain osassa raporteista on ilmoitettu PD-positiivisten koehenkilöiden taudin edistyminen tai vakavuus.

Analyyseissä käytettiin hyvin vaihtelevia piirteenekstraktointimenetelmiä, mikä näyttää ilmentävän vakiintuneen menettelyn puutetta tutkimusalalla. Kahdessa analyyseissä piirteenekstraktointi perustui signaalin taajuusspektrin analysointiin, Chaturvedi et al. (2017) hyödyntäen perinteisiä taajuuskaistoja ja Anjum et al. (2019) holistiseen tapaan. Niin ikään kahdessa analyyseissä piirteenekstraktointi perustui signaalin jollakin tavoin määriteltyyn kompleksisuuteen, Gomez et al. (2011) hyödyntäen Lempel-Ziv -kompleksisuutta ja Liu et al. (2017) näyte-entropiaa. Jäljelle jäävien analyyseiden piirteenekstraktointi perustui tila-avaruusesityksen rekonstruointiin (Lainscsek et al. (2013)), aivoalueiden synkronisaation analysointiin (Stoffers

---

<sup>16</sup>Tasapainotettu luokittelutarkkuus (engl. balanced accuracy) on käyttökelpoinen metriikka tilanteessa, jossa luokat ovat epätasapainossa. ks. esim. [70].

et al. (2008)), ja korkeamman kertaluvun spektrin tunnuslukuihin (Yuvaraj et al. (2016)). Huomio kompleksisuusmetriikoita käyttävien menetelmien onnistumisesta on osin vaikuttanut tämän tutkielman yhteydessä toteutetun analyysin metodologiaan. Yhdessäkään tarkastelluista tutkimuksista ei ole eroteltu signaalin komponentteja taajuuskaistojen perusteella ennen kompleksisuusmitan laskemista, joten tämä tutkielma tarjoaa osin uudenlaisen näkökulman piirteengenerointiin EEG-nauhoitteesta.

Analyysit käyttivät laajaa kirjoa erilaisia luokittelualgoritmeja tilastollisista malleista (esim. lineaaridiskriminanttianalyysi, logistinen regressio) edistyneempiin koneoppimismalleihin (esim. tukivektorikone, satunnaismetsäluokittelija). Kahdessa analyysissä oli käytetty uudenlaista luokittelualgoritmia (Liu et al. (2017) O\_CCA-luokittelija, sekä Anjum et al. (2019) LEAPD-luokittelija). Näissä tutkimusraporteissa raportoitiin myös varsin korkeita testisuorituskykyestimaatteja (n. 0.85–0.92), joten on mahdollista että erikoisvalmisteiset ad hoc-luokittelijat suoriutuvat hyvin Parkinsonin taudin tunnistamisessa EEG-aineistosta. Ylipäätään testiaineiston luokittelutarkkuusmetriikat vaihtelivat välillä 0.56 – 0.996 ja AUC-estimaatit välillä 0.63 – 1.00. Korkeimpiin raportoituihin suorituskykyometriikoihin on syytä suhtautua sikäli skeptisesti, että ne on saavutettu huomattavan pienillä otosko'oilla, ja on epäilyksenalaista yleistysivätkö tulokset suuremmille aineistoille. Analyyseissä oli käytetty pääosin asianmukaista ristiinvalidointimenettelyä, mutta ristiinvalidointikoetta ei oltu yhdessäkään analyysissä toistettu, mikä voi herättää kysymyksen esitettyjen testiaineiston suorituskykyarvioiden varianssista. Poikkeuksena Yuvaraj et al. (2016) herätti huomiota aineiston partitiointia edeltävällä piirteentalintamenettelyllä. Ne analyysit, jotka tarkastelivat eri aivoalueilla sijaitsevien elektrodien vaikutusta luokittelijan suorituskykyyn (Gomez et al. (2011), Chaturvedi et al. (2017), Anjum et al. (2019)), raportoivat oikean aivopuoliskon ohimolohkon alueiden, päälaenlohkon alueiden ja takaraivolohkon alueiden sekä jossakin määrin vasemman

aivopuoliskon keskialueiden elektrodien korostuneen analyysissä. Taulukossa 3.1 on esitettyinä yhteenvedot kaikista käsitellyistä EEG-analyyseistä.

MCI:stä kärsivien ja tavallisten Parkinsonin tautia sairastavien koehenkilöiden erottaminen on ollut Parkinsonin taudin tunnistamisen lisäksi kiinnostava koneoppimispohjaisen EEG-tutkimuksen kohde. Analyyseistä on raportoitu hieman vaatimattomampia tuloksia kuin PD:n tunnistamiseen keskittyvissä analyyseissä, mikä voi olla intuitiivista, sillä Parkinsonin tyyppin tunnistaminen (MCI:n sisällyttämisen oireisiin) voidaan tulkita erikoistuneemmaksi tehtäväksi kuin pelkästään Parkinsonin taudin toteamisen.

# 4 Analyysin aineisto, vaiheet ja tulokset

Tässä luvussa käydään läpi analyysin aineisto, vaiheet ja tulokset käyttämällä metodologia-luvussa esiteltyä käsitteistöä. Analyysistä on aikaisemmin raportoitu tutkielman kirjoitushetkellä julkaisemattomassa artikkelissa.<sup>1</sup>

Analyysin aineisto koostui kolmessa eri laboratoriossa toisistaan riippumatta kerätyistä monikanavaisista lepotilan EEG-aineistoista, joista kukin sisältää sekä silmät auki (jatkossa EO), että silmät kiinni (jatkossa EC) kerätyt tallenteet. Aineistot on kerätty Turun yliopistolla (20 Parkinson-positiivista ja 20 tervettä kontrollihenkilöä), New Mexicon yliopistolla (28/28) ja Iowan yliopistolla (EO: 41/41, EC: 20/15). Kokonaisuudessaan silmät auki kerättyjä havaintotapauksia on 178, joista 89 PD-positiivisia ja 89 kontrollitapauksia, ja silmät kiinni kerättyjä havaintotapauksia on 128, joista 66 PD-positiivisia ja 62 kontrollitapauksia. Iowan yliopistolla kerätty EC-aineisto erottuu muista aineistoista sikäli, että sen keräämiseen on osallistunut merkittävästi vähemmän koehenkilöitä kuin muiden aineistojen keräämiseen. Tämän lisäksi Turun yliopistolla kerätyssä EC-aineistossa on yksi positiivinen havaintotapaus vähemmän kuin vastaavassa EO-aineistossa. Turun yliopistolla tutkimukseen osallistuneiden PD-positiivisten koehenkilöiden taudin keskimääräinen kesto oli 6.4

---

<sup>1</sup>Suuronen, I; Airola, A; Pahikkala T; Murtojärvi M; Kaasinen V; Railo H: Budget-based classification of Parkinson's disease from resting state EEG (2021)

Turku (EO/EC)	Kontrolli	PD
N	20	20
Sukupuoli (M/N)	8/12	9/11
Ikä (ka. +/- SD)	67.8 (6.2)	69.8 (7.2)
Taudin kesto vuosina (ka. +/- SD)	n/a	6.4 (4.9)

Taulukko 4.1: Turun yliopistolla kerätyn aineiston kuvaus.

New Mexico (EO/EC)	Kontrolli	PD
N	28	28
Sukupuoli (M/N)	18/10	18/10
Ikä (ka. +/- SD)	69.2 (9.0)	69.7 (8.4)
Taudin kesto vuosina (ka. +/- SD)	n/a	5.5 (4.1)

Taulukko 4.2: New Mexicon yliopistolla kerätyn aineiston kuvaus.

vuotta (standardipoikkeama 4.9), New Mexicon yliopistolla vastaava luku oli 5.5 vuotta (SD 4.1) ja Iowan yliopistolla 5.5 vuotta (SD 3.0). Yhteenvedot analyysissä käytetyistä aineistoista sisältyvät taulukoihin 4.1 - 4.4.

Aineistot on esikäsitelty PREP-prosessointiputken avulla, minkä jälkeen ne yhdenmukaistettiin myöhempää käsittelyä varten poistamalla aineistojen välillä poikkeavat kanavat, ja järjestämällä kanavat yhteiseen järjestykseen. Yhdistettyyn aineistoon jäi 59 kanavaa, jotka sisältyivät jokaiseen aineistoon. Iowassa kerätystä

Iowa (EO)	Kontrolli	PD
N	41	41
Sukupuoli (M/N)	23/18	28/13
Ikä (ka. +/- SD)	71.3 (7.4)	66.1 (8.0)
Taudin kesto vuosina (ka. +/- SD)	n/a	5.4 (3.9)

Taulukko 4.3: Iowan yliopistolla kerätyn EO-aineiston kuvaus.

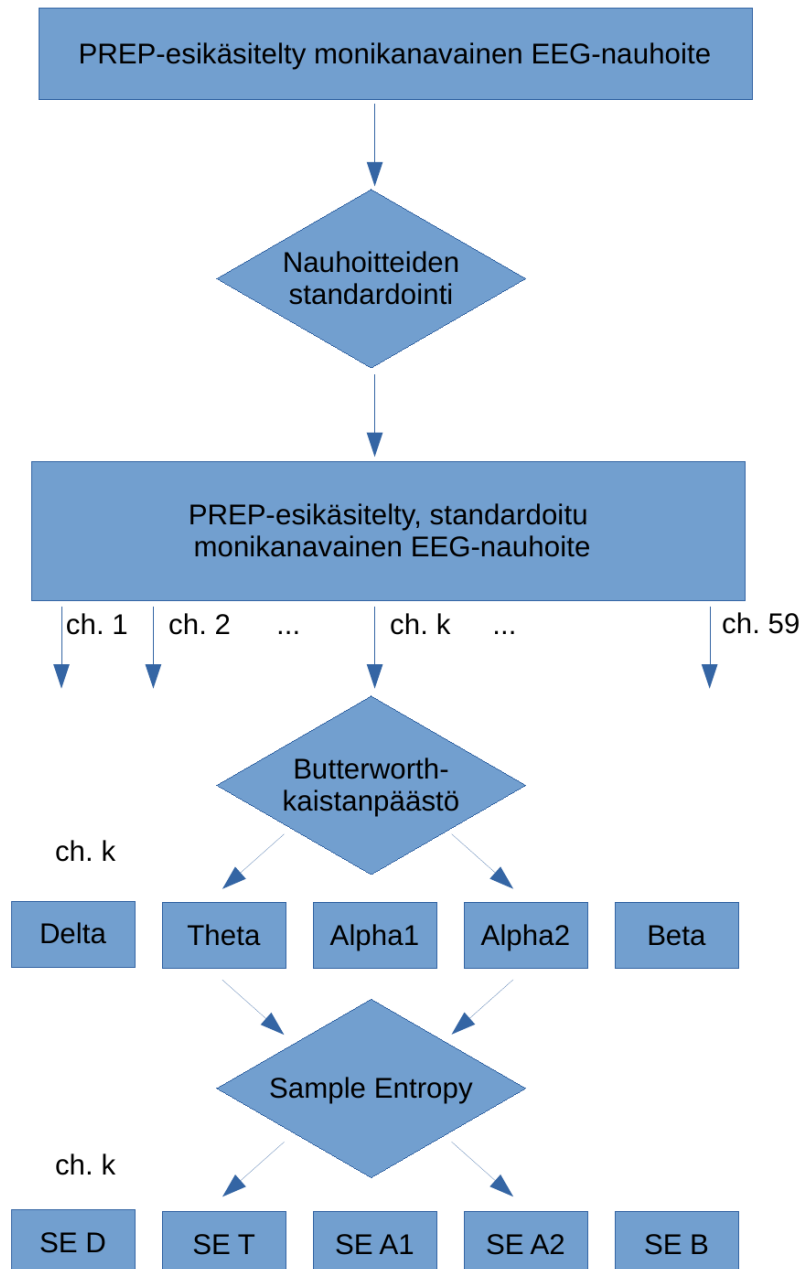
Iowa (EC)	Kontrolli	PD
N	15	20
Sex (M/N)	7/8	12/8
Age (ka. +/- SD)	70.8 (5.1)	67.2 (7.8)
Taudin kesto vuosina (ka. +/- SD)	n/a	5.5 (3.0)

Taulukko 4.4: Iowan yliopistolla kerätyn EC-aineiston kuvaus.

EO-aineistosta poistettiin tässä vaiheessa käsittelyä neljä positiivista tapausta ja kaksi negatiivista, sillä ne sisälsivät poikkeavan määrän kanavia, eivätkä näin ol-  
 len soveltuneet kanavapohjaiseen analyysiin ja lisäksi kolme negatiivista tapausta  
 jotta yhdistetyn kokonaisdatajoukon positiivisten ja negatiivisten tapausten luku-  
 määrät olisivat yhtä suuret. Analyysin aineistoksi päätyi tämän vaiheen jälkeen EO-  
 aineiston 168 tapausta (84 PD-positiivista ja 84 kontrollitapausta), ja EC-aineiston  
 128 tapausta (66 PD-positiivista ja 62 kontrollitapausta).

Yllä kuvatun esikäsittelyn jälkeen yhdistetystä aineistosta ekstraktoitiin piirre-  
 vektorit jokaista kanavaa kohden. Piirteenekstraktioinnin ensimmäisessä vaiheessa  
 kanavan signaalista erotettiin konventionaaliset delta-, theta-, alfa- (sekä matala et-  
 tä korkea) ja beta-taajuuskaistat käyttäen 3. kertaluvun digitaalista Butterworth-  
 kaistanpäästösuodatinta. Näille taajuuskaistan mukaan erotetuille viidelle signaalille  
 laskettiin näyte-entropiamitta käyttäen aiemmin esiteltyä näyte-entropia-algoritmia.  
 Lopputuloksena jokaista observaatiota kohden saatiin 295 numeerista piirrettä, joi-  
 ta ei prosessoitu pidemmälle. Visuaalinen esitys piirteengenroinnin vaiheista kuvassa  
 4.1.

Kuva 4.1: Vuokaavio, joka esittää piirteengenrointiprosessin vaiheita aineiston mielivaltaiselle tapausesimerkille.



Edelläkuvatulla menettelyllä tuotettua piirretensoria<sup>2</sup> käytettiin logistisen regressiomallin suorituskyvyn arviointiin, hyödyntäen sisäkkäistä 10:n laskoksen ristiinvalidointialgoritmia mallin budjettiperustaiseen optimointiin ahnetta hakua käyttävän, ryhmittäisen piirteervalinta-algoritmin avulla, sekä optimoidun ja sovitetun mallin yleistyskyvyn arviointiin. Malliin sisällytettävien piirreryhmien määrää kasvatettiin budjettia hallitsevan muuttujan arvoa kasvattamalla silmukkarakenteessa siten, että kaikki budjetit yhdestä piirreryhmästä täyteen 59:n piirreryhmän valikoimaan kokeiltiin vuorollaan. Näin voitiin havainnoida luokittelijan suorituskyvyn kehittymistä piirreryhmien lisäämisen yhteydessä. Koska piirreryhmät vastaavat EEG-nauhoitteen kanavia, voidaan tältä pohjalta esittää arvio siitä, miten suuri määrä elektrodeja tarvitaan EEG:n mittaamiseen hyvin suoriutuvaa Parkinsonin taudin tunnistamiseen käytettävää luokittelumallia varten.

Luokittelukykyä arvioitiin luokittelutarkkuus- ja AUC-metriikoiden perusteella. Ristiinvalidointiprosessi toistettiin 10 kertaa vaihtelevilla satunnaislukugeneraattorin siemenluvuilla ristiinvalidoinnin yhteydessä tehtävää datan partitiointia varten. Näin saadut 10 suorituskykyarviota keskiarvoistettiin lopullisen keskiarvoestimaatin tuottamiseksi, jotta sattuman<sup>3</sup> osuus olisi mahdollisimman pieni. Iteraatioiden lukumääräksi valikoitui 10 laskenta-ajan sekä varianssin vähentämisen välisenä kompromissina. Tämä iteroitu ja keskiarvoistettu ristiinvalidointiprosessi suoritettiin sekä EO- että EC-datalle.

---

<sup>2</sup>Tensori on matemaattinen objekti, jota voidaan tietojenkäsittelytieteen kontekstissa väljästi ajatella sklaaarien, vektorien ja matriisien kaltaisten objektien yleistyksenä. Siinä missä esimerkiksi matriisilla on kaksi indeksiä, jotka määräävät alkion sijainnin matriisin pysty- ja vaakariveillä, voi tensorilla olla mielivaltainen määrä indeksejä.

<sup>3</sup>Etenkin suhteellisen pienellä datajoukolla toteutunut datan partitiio voi vaikuttaa merkittävästi luokittelijan performanssiestimaattiin

## 4.1 Analyysin tulokset

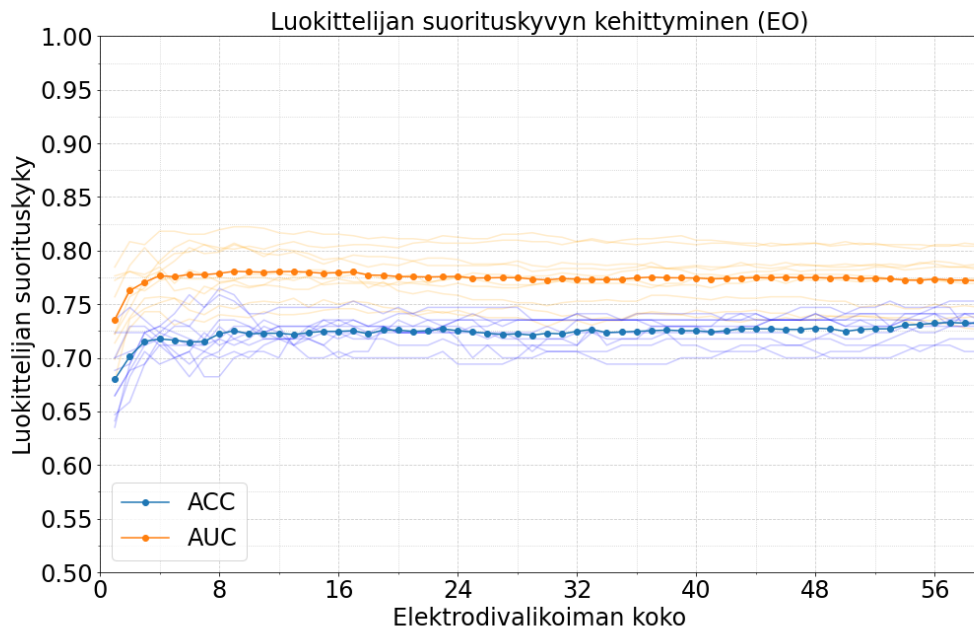
Luokittelijan suorituskyky vaihteli huomattavasti EO- ja EC-aineistojen välillä. Huomioiden Parkinsonin taudin tunnistamisen lähtökohtaisen haastavuuden, sekä käytettävissä olevan aineiston luonteen useasta lähteestä yhdistettynä kokoelmana, luokittelu onnistui EO-aineistolla keskimäärin suhteellisen hyvin. Sitä vastoin EC-aineistolla keskimääräinen luokittelutarkkuus oli huomattavasti matalampi.

EO-aineiston keskimääräisen budjettikohtaisen luokittelutarkkuuden minimi oli 0.68 yhdellä kanavalla, ja vastaavasti maksimi 0.733 57:llä kanavalla. Kanavien lisäämisestä aiheutuva hyöty on varsin pientä: vain kymmenellä kanavalla saavutettiin keskimäärin luokittelutarkkuus 0.722, jonka ero havaittuun keskimääräiseen maksimiin oli vain 0.011. Täydellä 59:n kanavan budjetilla luokittelutarkkuus oli 0.732. AUC-arvon keskimääräinen minimi 0.736 saavutettiin niinkään yhdellä kanavalla, ja keskimääräiset maksimit 0.781 yhdeksällä ja 13:lla kanavalla, minkä jälkeen se alkaa hieman vähentyä. Täydellä 59:n kanavalla joukolla keskimääräinen AUC-arvo oli 0.772. Luokittelijan suorituskykyä EO-aineistolla vaihtelevilla budjeteilla voi tarkastella perusteellisemmin kuvassa 4.2. Kuvassa janoilla yhdistetyt pisteet kuvaavat kaikkien kymmenen iteraation budjettikohtaisia keskiarvoja. Läpikuultavat käyrät kuvaavat performanssiestimaattien iteraatiokohtaista kehitystä.

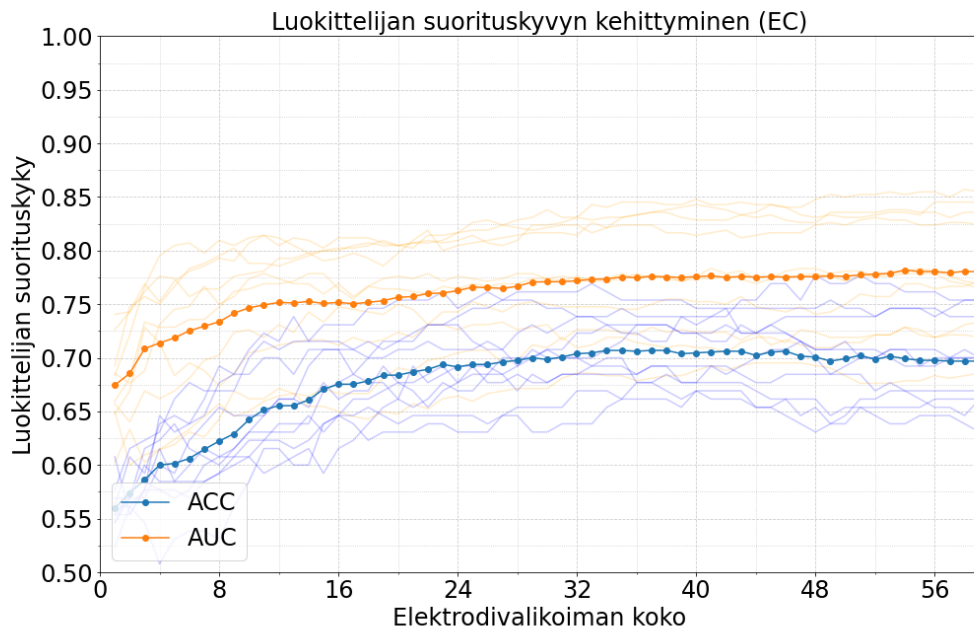
EC-aineiston keskimääräisen budjettikohtaisen luokittelutarkkuuden minimi oli 0.56 yhdellä kanavalla, ja maksimit 0.707 34:llä, 35:llä, 37:llä ja 38:lla kanavalla, minkä jälkeen keskimääräinen luokittelutarkkuus alkaa hieman vähentyä. Täydellä 59:n kanavan joukolla keskimääräinen luokittelutarkkuus oli 0.697. AUC-arvon keskimääräinen minimi oli 0.675 yhdellä kanavalla, ja keskimääräinen maksimi 0.782 54:llä kanavalla. Täydellä 59:n kanavan valikoimalla keskimääräinen AUC-arvo oli 0.78. Huomionarvoisesti EC-aineiston keskimääräinen luokittelutarkkuus ja AUC-arvo paranivat budjettia kasvattaessa pidemmälle kuin EO-aineistolla, jonka tapauksessa kasvu on ensimmäisten lisäkanavien jälkeen hyvin vähäistä. Siitä huolimatta,

että keskimääräisen luokittelutarkkuuden kasvu jatkui pidempään kanavia lisättäessä, pysyi EC-aineistolla saavutettu keskimääräinen luokittelutarkkuus huomattavasti EO-aineistolla saavutettua keskimääräistä luokittelutarkkuutta matalampana. Sitä vastoin keskimääräiset AUC-arvot olivat parhaimmillaan samaa tasoa. Luokittelijan suorituskyyä EC-aineistolla vaihtelevilla budjeteilla voi tarkastella perusteellisemmin kuvassa 4.3.

Kuva 4.2: EO-aineiston luokittelutarkkuuden ja AUC-estimaattien kehittymistä elektrodien lukumäärän funktiona kuvaava graafi.



Kuva 4.3: EC-aineiston luokittelutarkkuuden ja AUC-estimaattien kehittymistä elektrodien lukumäärän funktiona kuvaava graafi.



Tulosten tilastollisen merkitsevyyden osoittamiseksi suoritettiin permutaatiotestit sekä EO- että EC-aineistolla, käyttäen täyttä 59:n kanavan valikoimaa. Tilastollisella merkitsevyydellä tarkoitetaan sitä, kuinka epätodennäköistä on että esitetyt tulokset voitaisiin saavuttaa nollahypoteesin vallitessa. Nollahypoteesilla tarkoitetaan puolestaan tutkimuksessa ideaalisesti falsifioitavaa hypoteesia, että kahden muuttujan - tämän tutkielman tapauksessa EEG-aineistosta ekstrahoitujen piirteiden ja Parkinsonin taudin esiintymistä merkitsevän kohdemuuttujan - välillä ei ole merkitsevää yhteyttä. Nollahypoteesin negaatio - että muuttujien välillä vallitsee yhteystunnetaan puolestaan vaihtoehtoisena hypoteesina. Nollahypoteesin mukaista tilannetta voidaan simuloida esimerkiksi permutoimalla kohdemuuttujaa satunnaisesti, jolloin vaihtoehtoisen hypoteesin mukainen muuttujien välinen mahdollinen yhteys keinotekoisesti häivytetään. Käyttämällä satunnaisesti permutoituja kohdemuuttujan arvoja mallin sovittamiseen ja evaluointiin voidaan tuottaa nollahypoteesin

mukaisia luokitteluperformanssiestimaatteja. Toistamalla permutointimenettelyä ja tallettamalla nollahypoteesin mukaiset performanssiestimaatit voidaan arvioida sitä todennäköisyyttä, että vähintään samantasoiset tulokset kuin permutoimattomilla kohdemuuttujan arvoilla saavutetut voitaisiin saavuttaa nollahypoteesin vallitessa. Tätä menettelyä kutsutaan permutaatiotestiksi.[71]

Sellaisten permutaatioiden osuutta kaikista permutaatioista, joilla luokittelijan performanssiestimaatti on vähintään yhtä hyvä kuin permutoimattomassa tapauksessa kutsutaan p-arvoksi (engl. p-value). Yksikäsitteisen p-arvon tuottamiseksi olisi teoriassa suotavaa sisällyttää testiin kaikki mahdolliset kohdemuuttujan permutaatiot, mutta tämä on usein käytännössä mahdotonta tai epäkäytännöllistä laskentaajan näkökulmasta, joten tavallisesti käytetään jotakin “riittävän suurta” määrää permutaatioita. P-arvoa estimoidaan tavallisesti lausekkeella, jossa jaettavaan tai sekä jaettavaan että jakajaan on lisätty 1, jotta vältettäisiin nollan arvoinen p-arvo:

$$p(est, X, y, N) = \frac{|y' \in \{y'_0, \dots, y'_N\} : (est, X, y') \leq err(est, X, y)| + 1}{N + 1} \quad (4.1)$$

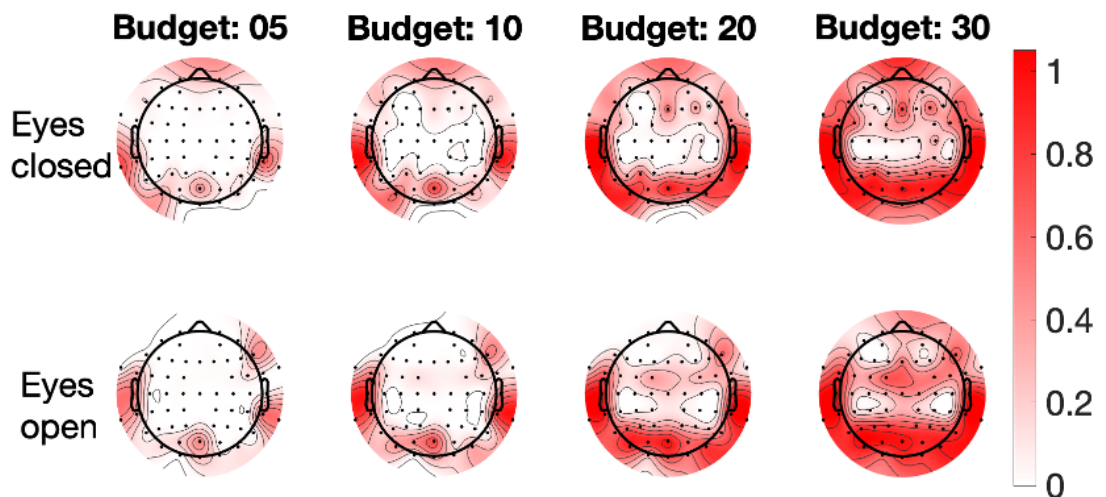
missä “est” on luokittelualgoritmi, “err” luokittelijan performanssin estimoiva funktio, X observaatiot, y observaatioita vastaavien kohdemuuttujien joukko,  $y'_0, \dots, y'_N$  kohdemuuttujan satunnaisia permutaatioita ja N permutaatioiden lukumäärä.

Koska edellä esitelty kaava asettaa p-arvon alarajaksi  $1/N + 1$ , on aiheellista sisällyttää testiin riittävä määrä permutaatioita jotta tilastollista merkittävyyttä implikoiva taso voidaan saavuttaa. Usein tilastollisen merkittävyyden ylärajana pidetään 0.05 tai 0.01, jolloin testattavien permutaatioiden vähimmäismäärä on 20 tai 100. Tässä analyysissä suoritettiin permutaatiotestit 10,000:llä satunnaisella permutaatiolla, käyttäen evaluointiin determinististä 10:n laskoksen litteää ristiinvalidointia. EO-datalla saavutettiin näin luokittelutarkkuudeksi 0.725 p-arvolla 0.0001. EC-datalla puolestaan saavutettiin luokittelutarkkuudeksi 0.686 niin ikään p-arvolla 0.0001. Saavutetut p-arvot ilmentävät tilastollista merkitsevyyttä korkealla toden-

näköisyydellä että nollahypoteesi voidaan hylätä.

Analyysissä tutkittiin paitsi kanavien lukumäärän vaikutusta luokittelijan suorituskykyyn, myös kanavien valintajärjestystä. Itoimalla budjettia yhdestä kanavasta täyteen 59:n kanavan joukkoon oli mahdollista tarkastella eri budjeteilla sovitettujen mallien kanavavalikoiman koostumusta. Kiinnittämällä mielivaltaisen budjetin voitiin tarkastella kuinka usein mikäkin kanava valikoitui malliin valitun budjetin ollessa voimassa, ja näin tutkia sekä vertailla eri kanavien merkitystä mallin luokitelukyvylle. Havaittiin, että ohimo- ja takaraivolohkon alueet korostuvat elektrodien valinnassa, siinä missä päälakilohkon alueilta elektrodeja ei valittu usein. Kanavien valintatiheyttä voi tarkastella perusteellisemmin kuvassa 4.4, jossa kanavaa vastaavan elektrodin sijaintiin on yhdistetty sen normalisoitu valintakertojen lukumäärä.

Kuva 4.4: Lämpökartta, joka yhdistää kunkin elektrodin sijainnin ja sen normalisoidun valintakertojen lukumäärän. Valinnat budjeteilla 5, 10, 20 ja 30 elektrodia.



Kanavien valinnan merkitystä arvioitiin vertaamalla optimoitujen kanavavalikoimien kanssa sovitetun mallin testisuorituskykyestimaatteja satunnaisilla samankokoisilla kanavavalikoimilla sovitettujen mallien suorituskykyarvioihin. Koska suurilla budjeteilla satunnaiset kanavavalikoimat ovat suurella todennäköisyydellä usein

koostumukseltaan samankaltaisia optimoitujen kanavavalikoimien kanssa, keskityttiin tässä analyysin vaiheessa tarkastelemaan yhden, viiden ja kymmenen kanavan suuruisia satunnaisia kanavavalikoimia, joita verrattiin analyysissä saatuihin testituloksiin samansuuruisilla budjeteilla. Satunnaisten kanavavalikoimien tapauksessa mallien ristiinvaldointiin käytettiin samoja datan partitioita kuin elektrodivalinnan sisältävässä analyysissä.<sup>4</sup> Jokaista partitiota kohden tarkasteltiin 100:aa satunnaisesti valittua kanavien osajoukkoa, joilla tuotetuista testiaineiston suorituskykyarvioista laskettiin keskiarvot partitiokohtaisen testiaineiston suorituskykyestimaatin tuottamiseksi. Tämän vaiheen tarkoituksena oli tuottaa robustimpia suorituskykyarvioita kuin jos olisi tarkasteltu vain yhtä satunnaista kanavien osajoukkoa. Näin saaduista partitiokohtaisista testiaineiston suorituskykyarvioista laskettiin niin ikään keskiarvot testiaineiston kokonaissuorituskykyarvioiden tuottamiseksi.

Satunnaisten kanavavalikoimien kanssa sovitettujen mallien keskimääräisistä testiaineiston suorituskykyarvioista huomattiin, että mallien kanavavalikoiman optimoimisella oli yllättävän pieni vaikutus mallin testiaineiston suorituskykyarvioon. Optimoitujen mallien testiaineiston luokittelutarkkuusarviot ja AUC-arviot olivat EO-aineiston tapauksessa kaikilla tarkasteltavilla budjeteilla (1, 5 ja 10 kanavaa) korkeintaan noin kaksi prosenttiyksikköä parempia kuin satunnaisilla kanavavalikoimilla sovitettujen mallien testiaineiston luokittelutarkkuusarviot. EC-aineistolla erot olivat samassa mittakaavassa, mutta viidellä ja kymmenellä kanavalla satunnaiset valikoimat suoriutuivat hieman optimoituja kanavavalikoimia paremmin, mikä voi implikoida mallien ylisovittumista. Tarkasteltujen optimoitujen ja satunnaisten kanavavalikoimien keskiarvoistettuja testiaineiston suorituskykyarvioita voi tarkastella perusteellisemmin taulukossa 4.5.

---

<sup>4</sup>Tarkemmin sanottuna satunnaisten elektrodiselektioiden yhteydessä käytetyn litteen ristiinvaldointiprosessin partitiot olivat samat kuin elektrodivalinnan sisältävän analyysin sisäkkäisen ristiinvaldointiprosessin ulomman rakenteen partitiot. Satunnaisen valinnan kanssa ei luonnollisesti ollut tarvetta käyttää sisäkkäistä ristiinvaldointia.

Elektrodivalikoima	ACC	AUC
EO 1 Optimoitu	0.680	0.736
EO 1 Satunnainen	0.666	0.725
EO 5 Optimoitu	0.716	0.776
EO 5 Satunnainen	0.707	0.755
EO 10 Optimoitu	0.722	0.780
EO 10 Satunnainen	0.720	0.761
EC 1 Optimoitu	0.560	0.675
EC 1 Satunnainen	0.557	0.694
EC 5 Optimoitu	0.602	0.719
EC 5 Satunnainen	0.619	0.748
EC 10 Optimoitu	0.642	0.747
EC 10 Satunnainen	0.660	0.762

Taulukko 4.5: Keskiarvoistetut testiaineiston luokittelutarkkuusarviot ja AUC-arviot optimoiduille ja satunnaisille yhden, viiden ja kymmenen kanavan valikoimille

## 4.2 Keskustelua analyysin tuloksista

Analyysin ensisijaisena tuloksena voi pitää huomiota, että monikanavaisesta EEG-datasta ekstraktoiduilla kompleksisuus pohjaisilla piirteillä sovitettu logistinen regressiomalli suoriutuu Parkinsonin taudin tunnistamisesta kohtuullisen hyvällä suorituskyvillä (EO-data: testiaineiston luokittelutarkkuus  $> 0.7$ , AUC  $> 0.75$ ), sekä sitä, että luokittelijan suorituskky melko suppealla kanavavalikoimalla (jopa alle kymmenen kanavaa) on keskimäärin samaa tasoa täyden 59:n kanavan valikoiman kanssa. Toissijaisina tuloksina voidaan pitää huomioita, että eri tutkimusryhmien nauhoittamia potentiaalisesti systemaattisesti toisistaan eroavia monikanavaisia EEG-nauhoitteita voidaan kootusti käyttää kohtuullisen suorituskkyisen koneoppimismallin sovittamiseen, että tutkimushenkilön silmät kiinni tallennettu EEG-nauhoite ei sovellu yhtä hyvin Parkinsonin taudin tunnistamiseen kuin silmät auki tallennettu, sekä sitä, että kanavavalikoiman koostumuksella ei näytä olevan suurta merki-

tystä luokittelusuorituskyvyn suhteen. Tulokset eivät kuitenkaan välttämättä yleisty muun laisille piirteengenerointimenetelmille tai luokittelijoille kuin tässä analyysissä käytetyt.

Analyysin tuloksena raportoitu koneoppimismallin suorituskyky on suhteellisen vaatimaton verrattuna suureen osaan aikaisemmasta tutkimuksesta, mikä voi osin selittyä käytetyn aineiston koolla ja alkuperän epäyhtenäisyydellä - tyypillisesti vastaavien analyysien aineistot ovat suppeampia ja homogeenisiä alkuperän suhteen. Budjettiperustaisen hakualgoritmin suorittama kanavavalinta suosi samoja aivoalueita kuin aiemmassa tutkimuksessa on havaittu.[64] Havainto silmät kiinni tallennetun EEG:n heikommasta luokittelutarkkuudesta on yllättävä aiemman kompleksisuusperusteista piirteengenerointia soveltaneen tutkimuksen valossa, jossa on käytetty silmät kiinni nauhoitettua EEG-dataa. [59] Mainitun tutkimuksen otoskoko on kuitenkin huomattavan pieni, ja on mahdollista että siinä raportoidut tulokset eivät ole edustavia silmät kiinni nauhoitetulle EEG-aineistolle yleisemmin. On myös syytä huomata, että EC-aineiston tapauksessa raportoitu heikompi luokittelusuorituskyky voi ainakin osin aiheutua myös EO-aineiston pienemmästä otoskoosta.

Huomio lokaalisti optimaalisten elektrodiselektioiden ja samankokoisten satunnaisten elektrodiselektioiden suhteellisen pienestä erosta mallin luokittelusuorituskyvyssä antaa aiheutta tutkia laskennallisesti halvempien, esim. ohjaamattomien piirteervalintamenetelmien mahdollisuuksia (esim. suodatinperusteiset piirteervalintamenetelmät) Myös tämän tutkielman yhteydessä tehdyn analyysin toteutunutta elektrodivalintaa kuvaavan lämpökartan mukaisen budjetoidun elektrodien sijoittelun testaaminen uudella EEG-aineistolla olisi mielenkiintoista. Kompleksisuusperusteisten piirteengenerointimenetelmien lisäksi ainakin konnektiivisuusperusteinen piirteengenerointi (esim. Stoffers et al.[21]) vaikuttaa lupaavalta, ja rajallisen budjetin käyttäminen voisi huomattavasti pienentää joihinkin potentiaalisesti hyviä tuloksia tuottaviin piirteervalintamenetelmiin (esim. synkronisaatiokattavuus) liittyvää huo-

mattavan raskasta laskenta-aikaa.

# Lähdeluettelo

- [1] T. Zesiewicz, K. Sullivan ja R. Hauser, ”Nonmotor symptoms of Parkinson’s disease”, 2006.
- [2] K. Beyer, D. Montserrat ja A. Ariza, ”Molecular Pathology of Lewy Body Diseases”, 2009.
- [3] R. Erro ja M. Stamelou, ”The Motor Syndrome of Parkinson’s Disease”, 2017.
- [4] R. Postuma ja T. Berg, ”The New Diagnostic Criteria for Parkinson’s Disease”, 2017.
- [5] N. Titova, N. Qamar ja K. Chaudhuri, ”Biomarkers of Parkinson’s Disease: An Introduction”, 2007.
- [6] M. Pramanik, R. Pradhan ja P. Nandy, ”Biomarkers for Detection of Parkinson’s Disease Using Machine Learning—A Short Review”, 2021.
- [7] C. Goetz et al., ”Movement Disorder Society-Sponsored Revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results”, 2008.
- [8] K. Bhattacharyya, ”Hallmarks of Clinical Aspects of Parkinson’s Disease Through Centuries”, 2017.
- [9] C. Filley, *Neurobehavioral Anatomy, Third Edition*. University Press of Colorado, 2011.
- [10] B. Postle, *Essentials of Cognitive Neuroscience*. John Wiley Sons, 2015.

- 
- [11] R. Dondelinger, "Electroencephalographs", 2009.
- [12] A. Biasiucci, B. Franceschiello ja M. Murray, "Electro-encephalography", 2019.
- [13] E. Reilly, "EEG Recording and operation of the apparatus", teoksessa *Electroencephalography : Basic Principles, Clinical Applications, and Related Fields*, E. Niedermeyer ja F. H. L. da Silva, toim., Wolters Kluwer, 2004.
- [14] G. Deuschl ja A. Eisen, *Recommendations for the Practice of Clinical Neurophysiology: Guidelines of the International Federation of Clinical Physiology*. Elsevier Science B.V., 1999.
- [15] J. Newson ja T. Thiagarajan, "EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies", 2019.
- [16] W. Tatum, *Handbook of EEG Interpretation. Second edition*. Demos Medical Publishing, 2014.
- [17] R. Soikkeli et al., "Slowing of EEG in Parkinson's Disease", 1991.
- [18] D. Cohen ja E. Halgren, "Magnetoencephalography", teoksessa *Encyclopedia of Neuroscience*, L. R. Squire, toim., Academic Press, 2009.
- [19] D. Stoffers et al., "Slowing of oscillatory brain activity is a stable characteristic of Parkinson's disease without dementia", 2007.
- [20] C. Stam ja B. van Dijk, "Synchronization Likelihood: An Unbiased Measure of Generalized Synchronization in Multivariate Data Sets", 2002.
- [21] D. Stoffers et al., "Increased cortico-cortical functional connectivity in early-stage Parkinson's disease: An MEG study", 2008.
- [22] G. V. Rossum ja F. Drake, *Python 3 Reference Manual*. CreateSpace, 2009.
- [23] C. Harris et al., "Array programming with NumPy", *Nature*, 2020.
- [24] P. Virtanen et al., "Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python", 2020.

- 
- [25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, 2011.
- [26] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing", *Behavior Research Methods*, 2021.
- [27] R. Vallat, *AntroPy*, versio 0.1.4. url: <https://github.com/raphaelvallat/antropy>.
- [28] A. Delorme ja S. Makeig, "EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics", 2004.
- [29] T. Ferree, "Spline interpolation of the scalp EEG", 2000.
- [30] N. Bigdely-Shamlo et al., "The PREP pipeline: standardized preprocessing for large-scale EEG analysis", 2015.
- [31] S. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*.
- [32] D. Sundararajan, *Digital Signal Processing, an Introduction*. Springer, 2021.
- [33] P. Flach, "ROC Analysis", teoksessa *Encyclopedia of Machine Learning*, C. Sammut ja G. Webb, toim., Springer, 2011.
- [34] D. Rickles, P. Hawes ja A. Shiell, "A Simple Guide to Chaos and Complexity", 2007.
- [35] J. Richman ja J. Moorman, "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy", 2000.
- [36] S. Pincus, "Approximate Entropy as a Measure of System Complexity", 1991.
- [37] G. Forman ja M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement", 2010.
- [38] A. Airola et al., "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve", 2010.

- 
- [39] G. Cawley ja N. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", 2010.
- [40] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, Inc., 2010.
- [41] J. Lever, M. Krzywinski ja N. Altman, "Regularization", 2016.
- [42] R. Fan et al., "LIBLINEAR: A Library for Large Linear Classification", 2008.
- [43] S. Huang, "Supervised feature selection. A tutorial", 2015.
- [44] P. Cassarà, A. Rozza ja M. Nanni, "A Cross-Entropy-based Method to Perform Information-based Feature Selection", 2016.
- [45] J. Li, K. Cheng, S. Wang ja F. Morstatter, "Feature Selection: A Data Perspective", 2016.
- [46] M. Chen, "A Greedy Algorithm with Forward-Looking Strategy", 2008.
- [47] M. Murtojärvi et al., "Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets", 2020.
- [48] S. Lee et al., "A convolutional-recurrent neural network approach to resting state EEG classification in Parkinson's disease", 2021.
- [49] M. Shaban ja A. Amara, "Resting-state electroencephalography based deep-learning for the detection of Parkinson's disease", 2022.
- [50] P. Diniz, *Digital signal processing : system analysis and design*. Cambridge University Press, 2002.
- [51] C. Gomez et al., "Complexity Analysis of Resting-State MEG Activity in Early-Stage Parkinson's Disease Patients", 2011.
- [52] C. Lainscsek ja T. Sejnowski, "Delay Differential Analysis of Time Series", 2015.

- [53] A. Hyvärinen, "Independent component analysis: recent advances", 2013.
- [54] A. Gałecki ja T. Burzykowski, *Linear mixed-effects models using R : a step-by-step approach*. Springer, 2013.
- [55] C. Lainscsek et al., "Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals", 2013.
- [56] W. Collis, P. White ja J. Hammond, "Higher-Order Spectra: The Bispectrum and Trispectrum", 1997.
- [57] R. Yuvaraj, U. Acharya ja Y. Hagiwara, "A novel Parkinson's Disease Diagnosis Index using higher-order spectra features in EEG signals", 2016.
- [58] D. Sundararajan, *Discrete wavelet transform : a signal processing approach*. Wiley, 2015.
- [59] G. Liu et al., "Complexity Analysis of Electroencephalogram Dynamics in Patients with Parkinson's Disease", 2007.
- [60] J. Proakis, *Digital Signal processing: Principles, Algorithms and Applications*. Prentice-Hall, 1996.
- [61] M. Chaturvedi et al., "Quantitative EEG (QEEG) Measures Differentiate Parkinson's Disease (PD) Patients from Healthy Controls (HC)", 2017.
- [62] D. Manolakis, *Statistical and adaptive signal processing : spectral estimation, signal modeling, adaptive filtering, and array processing*. Artech House, 2005.
- [63] H. Abdi ja L. Williams, "Principal component analysis", 2010.
- [64] M. Anjum et al., "Linear predictive coding distinguishes spectral EEG features of parkinson's disease", 2020.
- [65] G. Smith ja M. Bondi, "Mild Cognitive Impairment and Dementia: Definitions, Diagnosis and Treatment", 2013.

- 
- [66] M. Chaturvedi et al., "Phase lag index and spectral power as QEEG features for identification of patients with mild cognitive impairment in Parkinson's disease", 2019.
- [67] J. Zhang et al., "Identifying Parkinson's disease with mild cognitive impairment by using combined MR imaging and electroencephalogram", 2020.
- [68] V. Geraedts et al., "Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease", 2021.
- [69] M. Berouni et al., "Electroencephalography-Based Machine Learning for Cognitive Profiling in Parkinson's Disease: Preliminary Results", 2018.
- [70] K. Brodersen, C. Ong, K. Stephan ja J. Buhmann, "The balanced accuracy and its posterior distribution", 2010.
- [71] P. Golland, F. Liang, S. Mukherjee ja D. Panchenko, "Permutation Tests for Classification", 2000.