Eeli Mäkinen

Veren glukoositasojen ennustaminen koneopin avulla

Tietotekniikan kandidaatintutkielma

29. huhtikuuta 2023

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Eeli Mäkinen

Yhteystiedot: eeli.p.p.makinen@student.jyu.fi

Ohjaaja: Tytti Saksa

Työn nimi: Veren glukoositasojen ennustaminen koneopin avulla

Title in English: Predicting Blood Glucose Levels Using Machine Learning

Työ: Kandidaatintutkielma

Opintosuunta: Tietotekniikka

Sivumäärä: 23+0

Tiivistelmä: Tämä kirjoitelma on kirjallisuuskatsaus veren glukoositasojen ennustamiseen koneopin avulla. Tutkielma tarkastelee ennustuksesta tehtyä kirjallisuutta ja käy läpi koneopin käsitteitä liittyen aiheeseen. Reaaliaikainen ennustaminen mahdollistaa paremman glukoosikontrollin ja voi lisätä automaatiota tyypin 1 diabeteksen hoitoon.

Avainsanat: LaTeX, kandidaatintutkielmat, tyypin 1 diabetes mellitus, koneoppi, veren glukoositason ennustus, gradient boosting machine, tukivektorikone, satunnaismetsä, päätöspuu

Abstract: This paper is a literature review on blood glucose level prediction using machine learning. The thesis examines existing literature on prediction and discusses machine learning concepts related to the topic. Real-time prediction enables better glucose control and can increase automation in the management of type 1 diabetes.

Keywords: LaTeX, Bachelor's Thesis, Type 1 Diabetes Mellitus, Machine Learning, Blood Glucose Level Prediction, Gradient Boosting Machine, Support Vector Machine, Random Forest, Decision Tree

Kuviot

Sisällys

1 Johdanto

Diabetes on krooninen sairaus, joka vaikuttaa miljooniin ihmisiin ympäri maailmaa (Maahs ym. 2010). Yksi merkittävimmistä haasteista tyypin 1 diabeetikoille on glukoositasojen hallinta, sillä huonosti hallittu diabetes johtaa vakaviin terveysongelmiin. Tyypin 1 diabeetikoille tärkeä hoitokeino on jatkuva glukoosiseuranta (Facchinetti 2016). Glukoosidataa on seurannan myötä paljon, mutta sen hyödyntäminen on tällä hetkellä vain taudin sairastajalla.

Tämä kirjallisuuskatsaus pyrkii tutkimaan ja ymmärtämään koneopinalgoritmeja, joita käytetään veren glukoositasojen ennustamiseen. Työn motivaationa on parantaa ymmärrystä diabeteksesta ja oppia lisää koneoppimistekniikoista. Reaaliaikainen ennustaminen mahdollistaa paremman glukoosikontrollin ja voi lisätä automaatiota sairauden hoitoon. Tutkielma pyrkii vastaamaan tutkimuskysymykseen: Voidaanko koneoppia hyödyntää veren glukoositasojen ennustamiseen?

Tutkielmassa käsitellään diabetekseen liittyviä lähtötietoja ja glukoosinseurannan konsepteja luvussa 2. Luvussa 3 käsitellään yleisiä koneopin algoritmeja, jotka ovat asiaankuuluvia liittyen tutkimukseen. Luvussa 4 tutkitaan kirjallisuutta tehdystä tutkimuksesta. Luvussa 5 käydään läpi haasteita ja rajoituksia liittyen tutkimukseen. Lopuksi luvussa 6 on yhteenveto tutkielman keskeisimmistä tuloksista.

2 Tyypin 1 diabetes mellitus

Tyypin 1 diabetes mellitus (T1DM) on autoimmuunisairaus, jossa ihmisen immuunijärjestelmä tuhoaa haiman tuottamia beetasoluja, jotka vastaavat insuliinin tuotannosta. Tämä johtaa insuliinin vajaatuotantoon jolloin veren glukoositasot ovat kroonisesti koholla (Roglic 2016). Kaikki T1DM -potilaat aloittavat insuliinihoidon (Mehmood ym. 2020). Insuliinihoidossa diabeetikko injektoi insuliinia insuliiniruiskeella tai insuliinipumpulla. Tyypillisesti insuliineja on lyhyt- ja pitkävaikutteisia. T1DM diagnosoidaan yleensä lapsuudessa tai nuoruudessa, ja sen aiheuttaa yhdistelmä geneettisiä sekä ympäristöllisiä tekijöitä. Tarkkaa syytä taudin kehittymiselle ei kuitenkaan tiedetä. Yleisiä oireita ovat lisääntynyt jano sekä virtsaaminen, painon lasku, väsymys ja sumea näkö. Hoitokeinoina tavallisesti käytetään insuliinihoitoa, jatkuvaa glukoosinseurantaa, ruokavalion optimointia ja säännöllistä liikuntaa. Hoidon tavoitteena on estää komplikaatioita jotka voivat olla hengenvaarallisia, jos diabetes jää hoitamatta (Maahs ym. 2010).

T1DM kehittyy yleensä lapsena tai nuorena aikuisena. Sairauden kehittymismekaniikoista ei tiedetä paljoa ja yleensä sairaudentekijät lasketaan ympäristöllisiin ja geneettisiin tekijöihin. Virossa on kolme kertaa vähemmän T1DM -potilaita kuin Suomessa, mihin ei ole löytynyt järkevää selitystä. Samanlainen yhteys on havaittu myös Puerto Ricossa ja Yhdysvalloissa (Atkinson 2012).

2.1 Hypo- ja hyperglykemian vaarat

Hypoglykemiassa veren glukoositaso tippuu alle 3,9 mmol/L (DiMeglio, Evans-Molina ja Oram 2018; Mujahid, Contreras ja Vehi 2021). Se voi johtaa tajunnan menetykseen, kohtaukseen, sekavuuteen ja lopulta kuolemaan. Diabeetikolle hypoglykemia on erittäin epämukava olotila ja sitä usein pelätään.

Hyperglykemiassa veren glukoositaso nousee yli 10,0 mmol/L (DiMeglio, Evans-Molina ja Oram 2018). Yleisiä oireita on janoisuus, jatkuva virtsaamisen tarve ja pitkäkestoisena tajuttomuus tai kooma. Se on myös pääsyy sydän- ja verisuonisairauksille diabeetikoilla.

Varsinkin nuorilla lapsilla veren glukoositasot heittelevät enemmän kuin aikuisilla, jolla on negatiivinen vaikutus heidän aivojen kehittymiselle (Nevo-Shenker ja Shalitin 2021). Kroonisella hyperglykemialla ja epätasaisella veren glukoosipitoisuudella on yhteys aivojen rakenteen muunnoksiin ja toiminnan häiriintymisiin nuorilla T1DM -potilailla.

2.2 Jatkuva glukoosinseuranta

Jatkuva glukoosinseuranta (engl. Continuous Glucose Monitoring, CGM) on menetelmä tyypin 1 diabeteksen hallinnassa (Facchinetti 2016). CGM-teknologia koostuu kolmesta pääkohdasta. Ihonalainen sensori kerää jatkuvasti tietoa veren glukoosipitoisuudesta. Sensorissa on myös lähetin, jonka tarkoitus on kommunikoida signaalinvastaanottimen kanssa, joka on yleensä langattomasti toimiva laite. Langaton laite vastaanottaa signaalin, joka kertoo reaaliaikaisen veren glukoosipitoisuuden. Laite myös visualisoi pitoisuuden ja tallentaa sen. Tällä hetkellä suurin osa älypuhelimista pystyy vastaanottamaan lähettimen signaalin ja toimimaan kuten edellä mainittu laite. CGM tarjoaa yksityiskohtaisempaa tietoa kuin perinteiset glukoosipitoisuuden itsevalvontamenetelmät (sormen päistä otettava verinäyte).

Meta-analyysi jatkuvasta glukoosiseurannasta analysoi 19:n satunnaisesti kontrolloidun tutkimuksen tuloksia ja havaitsi, että CGM:n käyttö voi johtaa merkittäviin parannuksiin glukoosikontrollissa ja vähentää hypoglykemian riskiä (Gandhi ym. 2011). Tutkimuksessa havaittiin, että CGM alensi HbA1c -tasoja (pitkäaikainen veren glukoositaso) 0,4 - 0,5 prosenttiyksikköä. CGM:n käyttö nuorilla lapsilla on etenkin suositeltua, jolloin saadaan pienennettyä hypo- ja hyperglykemian aikaa (Nevo-Shenker ja Shalitin 2021).

2.3 Veren glukoositasojen ennustus

Diabeteksen hoidossa tärkeää on hypo-/hyperglykeemisten tapahtumien ehkäiseminen. Yksi ensimmäisistä tutkimuksista aiheen parissa selvitti glukoositasojen ennustamisen mahdollisuutta (Sparacino ym. 2007). CGM:n avulla seurattiin 48 tunnin ajan veren glukoositasoja tyypin 1 diabeetikoilla, joita osallistui kokeeseen 28 henkilöä. Tulokset osoittivat, että glukoosin ennustaminen menneistä tiedoista on mahdollista ja algoritmien suorituskyky on riittävä hypo-/hyperglykeemisten tapahtumien estämiseksi. Tapahtumat pystyttiin ennusta-

maan 20–25 minuuttia ennen kynnysarvon ylitystä. Tutkimuksessa tuotiin esiin useita ongelmia ja haasteita, jotka ratkaistiin seuraavan kerran vuonna 2016 luomalla älysensoritekniikkaa (Facchinetti 2016).

Yhdistämällä kaksi olemassa olevaa teknologiaa on mahdollista kehittää keinotekoinen haima, joka toimisi normaalin haiman lailla tuottaen insuliinia sitä tarvittaessa (Mehmood ym. 2020). Teknologian takana ovat insuliinipumppu ja jatkuva glukoosinseuranta, jotka toimisivat yhdessä. Tämä vaatii algoritmin, joka pystyy täydellisesti ennustamaan veren glukoositasoja datan virheistä huolimatta.

3 Yleisiä koneopin algoritmeja

Koneoppiminen (engl. machine learning) on tietokoneiden kyky oppia ja tehdä päätöksiä datan perusteella ilman, että niitä erityisesti ohjelmoitaisiin näihin tarkoituksiin (Bishop 2006). Koneoppiminen on erityisen hyödyllinen työkalu aloilla, joissa käsitellään suuria datamääriä. Koneopin avulla datan perusteella voidaan luoda ennusteita ja auttaa päätöksentekoa eri konsepteissa. Yksi tällainen ala on diabeteksen hoito, jossa glukoositasojen ennustaminen on tärkeää oikeanlaisen hoidon suunnittelussa.

Tässä luvussa käsitellään tarkemmin viittä koneoppimisen algoritmia tai mallia, joita on käytetty glukoositasojen ennustamisessa. Tarkasteltavat algoritmit on valittu, koska niiden tehokkuudesta glukoositasojen ennustamisessa on vahvaa näyttöä (Afsaneh ym. 2022). Lisäksi käydään läpi algoritmien vahvuuksia ja heikkouksia.

3.1 Puupohjaiset mallit

Puupohjaiset algoritmit ovat yleisiä ja tehokkaita koneopin malleja, joita käytetään luokittelussa ja regressiossa. Nämä algoritmit oppivat sarjan pääsääntöjä datasta ja muodostavat puumaisen rakenteen, jota käytetään ennustamaan uusia tietoja.

Päätöspuu on suosittu ja laajasti käytetty koneopin algoritmi. Breimanin kirja ”Classification and Regression Trees” on perusteellinen teos aiheesta (1984). Teos tarjoaa yksityiskohtaisen kuvauksen päätöspuista ja niiden käytöstä ennustavassa mallinnuksessa. Algoritmilla on hierarkkinen rakenne, joka jakaa piirteiden tilan alueisiin, jotka ovat homogeenisia kohdennettavan muuttujan suhteen (Breiman 1984, luvut 1-4). Puu rakennetaan rekursiivisesti valitsemalla ominaisuus, joka parhaiten jakaa datan kahteen tai useampaan homogeeniseen osaan. Tämä jakamisprosessi jatkuu, kunnes pysäytyskriteeri täyttyy. Esimerkkejä pysäytyskriteereistä ovat maksimisyvyys ja vähimmäismäärä näytteitä lehtisolmussa.

Breiman (1984, luku 1) kuvaa päätöspuiden keskeisien osien olevan jakamissäännöt, pysäytyskriteerit, karsinnat ja muuttujien valinta. Jakamissääntö määrittää miten data jaetaan kunkin sisäsolmun kohdalla. Jakamissääntöjä on useita ja niiden päätehtävä on mitata jaon

informaatioarvoa (Breiman 1984, luku 4). Pysäytyskriteerit määrittävät milloin jakaminen on lopetettava (Breiman 1984, luku 3). Karsinta puolestaan on tekniikka päätöspuun koon pienentämiseksi, ennustavan tarkkuuden parantamiseksi ja monimutkaisuuden vähentämiseksi. Breiman (1984, luku 10) kuvailee useita menetelmiä päätöspuiden karsimiseksi, kuten kustannusmonimutkaisuus karsinta ja pienimmän virheen karsinta, jotka perustuvat tarpeettomien solmujen ja alipuiden poistamiseen.

Vuonna 2001 Breiman ehdotti ”satunnaismetsää”, joka on osoittautunut hyvin yleiskäyttöiseksi luokittelu- ja regressiomalliksi (Breiman 2001). Menetelmä yhdistää useita satunnaisia päätöspuita ja keskiarvoistaa niiden ennusteet. Parhaisiin tuloksiin päästään, kun muuttujien määrä on huomattavasti suurempi kuin havaintojen määrä. Breiman (2001) kuvailee bagging-tekniikkaa menetelmänä, jota käytetään yhdessä satunnaismuuttujan valinnassa. Bagging tulee sanoista bootstrap aggregating ja se on keskeinen osa satunnaismetsää. Menetelmää käytetään useiden päätöspuiden luomiseen eri bootstrap-näytteistä koulutusdatassa. Jokainen puu on koulutettu eri bootstrapilla käyttäen satunnaisesti valittua muuttujaa, joka auttaa hajauttamaan puita ja estämään ylisovitusta. Ylisovitus tapahtuu, kun koneoppimismalli alkaa sovittamaan opetusdataa liian monimutkaisesti (Demšar ja Zupan 2021). Tällöin tulosten ennustamistarkkuus heikentyy.

Satunnaismetsä on monipuolinen algoritmi, joka soveltuu suurten aineistojen käsittelyyn, kuten esimerkiksi oppimistehtäviin (Biau ja Scornet 2016). Algoritmi on erinomainen valinta reaaliaikaiseen ennustamiseen, koska se kykenee tehokkaaseen suurten aineistojen käsittelyyn, sopeutuu ajallisesti muuttuvaan dataan ja tarjoaa tarkkoja ennustuksia nopeasti muuttuvissa ympäristöissä. Veren glukoositasojen ennustaminen vaatii algoritmilta reaaliajassa suoriutumista.

3.2 Tehostamismallit

Tehostamisen pääidea on parantaa perusmallien ja heikkojen mallien tarkkuutta yhdistämällä useita perusmalleja iteratiivisesti vahvemmiksi malleiksi (Freund ja Schapire 1997). Perusajatuksena on kouluttaa sarja heikkoja malleja (päätöspuu, lineaarinen malli ja neuroverkko) koulutusdatalla, joista jokainen perättäinen malli keskittyy oppimaan edellisten mallien

virheistä. Tämä saavutetaan antamalla suurempia painoarvoja väärin luokitetuille näytteille koulutusdatassa, mikä pakottaa peräkkäiset mallit priorisoimaan niitä näytteitä oppimisprosesseissaan.

Gradient Boosting Machine (GBM) on koneoppimisalgoritmi, joka käyttää tehostamista. Friedmanin (2001) mukaan pääajatuksena on käyttää gradienttimenetelmää tappiofunktion minimoimiseen. Tappiofunktio on ennustetun ja todellisen tuloksen välinen ero. Algoritmi lisää iteratiivisesti päätöspuita malliin, jossa jokainen puu on sovitettu edellisen puun tekemille jäännösvirheille. Lopullinen malli on päätöspuiden painotettu summa. GBM:n käytöllä on useita etuja muihin algoritmeihin verrattuna (Friedman 2001). GBM pystyy käsittelemään erilaisia syötetietoja, kuten kategoria-, teksti- ja numerodatoja. Se pystyy myös käsittelemään puuttuvia tietoja ja poikkeamia, ja se on vähemmän altis ylisovitukselle kuin muut algoritmit. Sen käyttö on laskennallisesti vaativaa ja saattaa vaatia useiden hyperparametrien hienosäätöä.

Extreme Gradient Boosting (XGBoost) on vakinaistettu muoto GBM:stä ja käytännössä tehokkaampi (Chen ja Guestrin 2016). XGBoost lisää säännöllisyyttä tappiofunktioon, joka rajaa mallin kertoimien kokoa. Tämä estää ylisovitusta ja lisää mallin yleistämiskykyä. XGBoostissä on lisäksi sisäänrakennettu mekanismi puutuvien arvojen käsittelemiseksi syötteessä. Puuttuvien arvojen satunnaisen täydentämisen sijaan XGBoost huomioi ne, ja tekee niistä omat johtopäätökset. XGBoost käyttää rinnakkaiskäsittelyä koulutusprosessin nopeuttamiseksi jakamalla työmäärän useille suorittimille tai näytönohjaimille.

3.3 Tukivektorikoneet

Tukivektorikoneet ovat valvotun oppimisen algoritmeja. Tukivektorikoneet toimivat löytämällä parhaan mahdollisen päätösrivin kahden tietojoukon välillä, joka maksimoi tietojoukkojen välisen marginaalin (Burges 1998). Marginaali on kahden lähimmän pisteen välimatka, jota kutsutaan tukivektoriksi. Tukivektorikoneita on lineaarisia, polynomisia ja radiaalisia perusfunktioita. Jokainen variaatio käyttää erilaista ydinfunktiota, joka muuntaa syötetiedot hypertasoksi korkeammassa ulottuvuudessa, jossa on helpompi löytää lineaarinen päätösrivi.

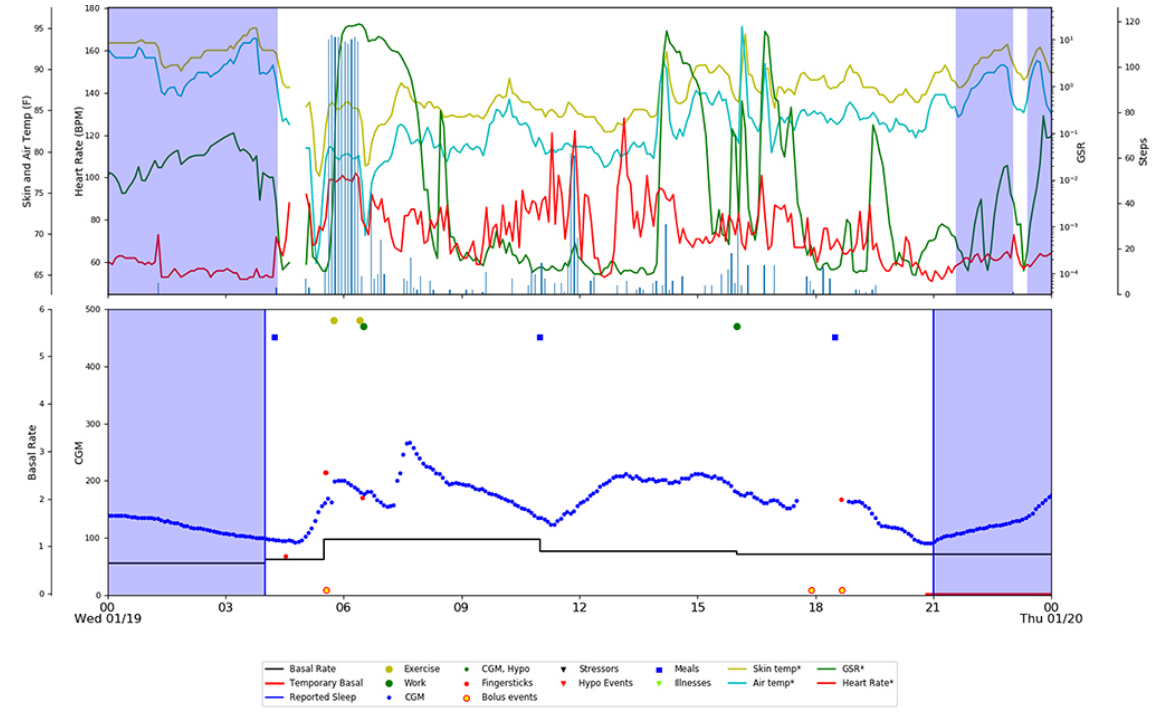
Tukivektorikoneet kärsivät samankaltaisista rajoitteista kuten tehostamis- ja puupohjaiset

mallit (Cortes ja Vapnik 1995). Ne ovat laskennallisesti vaativia, erityisesti isoille aineistoille. Aineiston koon kasvaessa, tukivektorikoneen koulutusaika kasvaa huomattavasti enemmän. Suoritus heikkenee, jos hyperparametrit ovat valittu väärin, kuten ydinfunktio ja säännöllisyysparametri. Hyperparametrit ovat vaikea valita ja säätää optimaalisesti. Tukivektorikoneet ovat herkkiä meluisille ja virheellisille aineistoille, mikä johtaa alioptimaalisiin päätösriveihin. Sama pätee myös epätasapainoisille aineistoille, joissa yhdellä luokalla on huomattavasti vähemmän näytteitä kuin toisella. Tämä johtaa vinoihin päätösriveihin, jotka suosivat enemmistöluokkaa.

4 Kirjallisuuden tulkinta

4.1 OhioT1DM datasetti

OhioT1DM datasetti (Kuvio 1) on luotu veren glukoositasojen ennustamisen mahdollistamiseen (Marling ja Bunescu 2020). Se sisältää kahdeksan viikon ajalta tietoa 12 potilaan veren glukoositasoista. Datasetti julkaistiin vuonna 2018 ensimmäiseen Blood Glucose Prediction Challenge -haasteeseen, jonka tarkoitus oli lisätä tutkimusta aiheesta. Tämä on ensimmäinen julkinen datasetti, joka sisältää tietoa glukoositasoista, insuliiniannoksista sekä tiedot aterioista ja fyysisestä toiminnasta. Potilailta kerättiin aktiivisuusrannekkeilla tietoa heidän elintoiminnoista kuten sydämen sykkeestä (Kuvio 1). Tässä luvussa keskitytään veren glukoositasoon, jota on merkitty sinisellä poikkiviivalla (Kuvio 1).



Kuvio 1. Datasetin visualisointi (Marling ja Bunescu 2020)

Aineisto sisältää CGM-mittarin tiedot viiden minuutin välein, veren glukoositasot sormenpäämittauksista, pikainsuliinin ja perusinsuliinin annokset, itse raportoidut ateria-ajat hiilihydraattiarvioineen sekä itse raportoidut ajat liikunnalle, unelle, työlle, stressille ja sairauk-

sille. Nämä tiedot löytyvät datasta XML-tiedostomuodossa. Data visualisoidaan OhioT1DM Viewer -sovelluksella. Alla on mainittu tärkeimmät tietokentät datasetistä, jotka voidaan myös havaita Kuviosta 1.

- <patient> Tiedot insuliinin tyypistä sekä potilaan tiedot.
- <basal> Perusinsuliinin annostus.
- <temp basal> Tilapäinen insuliinin annostus, joka korvaa normaalin perusinsuliinin annostuksen.
- <glucose level> Glukoositaso CGM-laitteesta 5 minuutin välein.
- <fingerstick> Glukoositaso verensormenpäämittauksesta.
- <bolus> Pikainsuliinin annostelu, yleensä ennen ateriaa tai hyperglykemiaa.
- <sleep> Potilaan itse raportoima unen määrä ja laatu.
- <exercise> Potilaan itse raportoima liikunnan määrä ja laatu.
- <work> Tietoja potilaan työelämästä.
- <meal> Potilaan itse raportoima aterian aika ja arvioima aterian hiilihydraattien määrä.

XML-tiedostossa on paljon tietokenttiä, kuten edellä mainittiin, jotka vaikeuttavat koneoppimismallien opetusta. Suuri määrä tietokenttiä johtaa yleensä korkeampaan ulottuvuuteen, joka vaikeuttaa mallin hahmontunnistuskykyä. Myös ylisovitus on yleinen ongelma.

4.2 Tutkimuksia glukoositasojen ennustamisesta

Tutkimuksessa (Georga ym. 2013) suoritettiin perusteellinen tutkimus glukoosipitoisuuden ennustamisesta tyypin 1 diabeetikoilla. Tutkimuksessa ehdotettu menetelmä perustuu ”tukivektoriregressioon” (engl. Support Vector Regression, SVR) ja käsittelee ennustamista monimuuttujaregressioprosessina. Mallissa huomioidut muuttujat ovat plasman insuliinipitoisuus, ruokailusta peräisin olevan glukoosin määrä verenkierrossa, sekä energiankulutus fyysisen toiminnan aikana. Tutkimuksessa oli 27 potilasta. Tulokset osoittavat, että monimuuttujadata voi merkittävästi lisätä sekä lyhyen että pitkän aikavälin glukoosin ennustamisen tarkkuutta.

Tutkimuksessa (Gadaleta ym. 2019) vertailtiin regressio- ja luokittelualgoritmeja hyper- ja hypoglykeemisten taphtumien ennustamiseksi CGM-signaalien perusteella. Sekä staattisia, että dynaamisia lähestymistapoja tutkittiin. Tulokset osoittivat, että staattiset menetelmät

suoriutuivat paremmin kokonaisuutena, joista lineaarinen ja bayesilainen regressio antavat parhaan tapahtumaennusteen. Tukivektorikone suoriutui lähes yhtä hyvin kuin paras regressori, kun sitä koulutettiin tiettyyn tapahtumaan, kuten hyperglykemiaan. Erikoistunut tukivektorikone antoi parhaan kokonaissuorituksen.

Tutkimuksessa (Mosquera-Lopez ym. 2020) keskityttiin yölliseen hypoglykemiaan. Yöllinen hypoglykemia on yksi vaarallisimmista hypoglykemian muodoista, koska potilas ei välttämättä herää siihen. Tutkimuksessa koulutettiin SVR-malli käyttäen 124 potilaan CGM-dataa ja insuliniannoksia. Tuloksissa osoitettiin, että SVR-malli ennusti 94 prosenttia yöllisistä hypoglykemioista.

Tuore tutkimus yhdisti SVR:n ja syväopin neuroverkkomallin moniajoperiaatteella (Multitask Learning), jolla saatiin tarkkoja ennustuslukemia OhioT1DM datasetistä (Daniels, Herrero ja Georgiou 2022). Yhdistämällä koneopin ja syväopin malleja voidaan tehdä tehokkaita ennustuksia pienemmällä datan määrällä, joka yleensä on haastavaa.

Neljää koneopin mallia, jotka koulutettiin 104 potilaan CGM-datasta ja insuliiniannoksista, käytettiin ennustamaan aterianjälkeisiä hypoglykemioita (Seo ym. 2019). Satunnaismetsä suoritui parhaiten ennustaen 91,3 prosentin tarkkuudella aterianjälkeisiä hypoglykemioita 30 minuutin ennustusaikavälillä.

Kuten useissa tutkimuksissa mainittiin, nämä koneoppimismallit on opetettu dataseteillä, jotka perusteellisesti listaavat insuliini-, hiilihydraatti- ja liikuntamäärät. Rreaaliaikaisessa ennustamisessa ei välttämättä saada näitä tietoja mallin käytettäväksi. Tutkimuksessa (Dave ym. 2020) saatiin luetettavia tuloksia ilman hiilihydraattidataa. Optimoitu satunnaismetsämalli pystyi tunnistamaan hypoglykemian 30 minuuttia etukäteen 94 prosentin varmuudella.

5 Haasteet ja rajoitukset

Jatkuva glukoosinseuranta on elintärkeää diabeteksen hoidossa, koska se mahdollistaa terveellisen elämäntyylin suunnittelun. Invasiiviset ja jatkuvat mittaukset aiheuttavat kipua ja rasitusta yksilöille ja voivat täten estää seurannan (Lekha ja M. 2018). Näiden haasteiden käsittelemiseksi on alettu kehittää ei-invasiivisia menetelmiä, jotka voivat voittaa mainitut rajoitukset ja estää kivun. Tutkimus esittää ei-invasiivisen Monte Carlo -fotonisimulaatiopohjaisen mallin sormenpään valoabsorptiospektroskopian avulla mittaamaan veren glukoosipitoisuutta (Haque ym. 2021). Yksinkertaisesti sanottuna menetelmä perustuu infrapunavaloon. Tämän jälkeen käytettiin XGBoostia analysoimaan saatua glukoosidataa. Tuloksissa osoitettiin menetelmän olevan vähemmän kompleksi muihin ei-invasiivisiin menetelmiin, mutta tarkempi ennustamaan glukoositasoja (Haque ym. 2021).

Suurin osa tutkimuksiin käytetyistä dataseteistä hyödyntää potilaiden itse arvioimia muuttujia, kuten hiilihydraattien ja liikunnan määrää (Marling ja Bunescu 2020; Dave ym. 2020). Luottaessa itsearvioituihin muuttujiin on tärkeää huomioida mittausvirheiden ja vinouman mahdollisuus, joka voi vaikuttaa ennustemallien tarkkuuteen ja yleistettävyyteen. Lisäksi datasetit, joista löytyy muuta kuin CGM-dataa, ovat suhteessa pieniä. Tämä voi rajoittaa kattavampien ennustemallien kehittämistä ja validointia. Tämä ei kuitenkaan tarkoita, että jo olemassa oleva data olisi hyödytöntä.

Yksi haaste on datan epätasapainoisuus, jossa normaalit glukoositasot ovat yleisempiä kuin hypoglykeemiset tapahtumat. Tämä aiheuttaa tilanteen, jossa koneoppimismallit voivat vääristyä enemmistöluokan suuntaan ja suoriutua huonosti vähemmistöluokan ennustamisessa (Mujahid, Contreras ja Vehi 2021). Datan laatuun vaikuttaa CGM:n tarkkuus ja luotettavuus, mikä ei ole täydellistä. Koneoppimismallien ymmärrettävyys voi olla vaikeaa kliinisessä käytössä, jossa läpinäkyvyys ja selitettävyys on erityisen tärkeää. Myös aiemmin mainittu ylisovitus on usein ongelmallista.

CGM-sensoriteknologia on mullistunut viime vuosikymmenen aikana, joka on mahdollistanut uusien teknologioiden kehittämistä. Huomio on siirtynyt kohti käytettävyyden parantamista ja CGM-laitteiden tuottaman runsaan datan hyödyntämistä. Kaksi tärkeää tutkimusa-

luetta liittyen CGM-sensoreiden käyttöön ovat insuliiniannostelusääntöjen optimointi CGM-datan perusteella ja kalibrointien tarpeen vähentäminen. CGM-laitteet tarjoavat jatkuvaa tietoa perinteisen sormenpäämittauksen lisäksi, kuten glukoositasojen trendeistä ja aiemmasta historiasta. Näitä voidaan hyödyntää insuliiniannostelusääntöjen parantamiseksi. Lisäksi kalibrointien tarpeen vähentäminen tai poistaminen on ratkaisevaa CGM-laitteiden käytettävyyden parantamiseksi (Cappon ym. 2019).

Vikaantumisen havaitsemiseen liittyvät algoritmit ovat myös tärkeitä hetkellisten ja pysyvien vikojen tunnistamiseksi CGM-sensori-insuliinipumppujärjestelmässä, erityisesti suljetun silmukan automatisoiduissa insuliininannostelujärjestelmissä (keinotekoinen haima). Vaikka joitain algoritmeja on ehdotettu, lisätutkimusta tarvitaan niiden tehokkuuden määrittämiseksi ennen kuin ne voidaan ottaa käyttöön kliinisissä ympäristöissä (Facchinetti 2016).

6 Yhteenveto

Tämä tutkielma analysoi tyypin 1 diabeetikoiden verenglukoositasojen ennustamista koneopin avulla. Glukoositasojen ennustaminen on haastavaa, johtuen taudin monimutkaisuudesta. Tulokset osoittivat, että koneoppimisalgoritmit voivat ennustaa glukoositasoja tehokkaasti.

Tutkielmassa käsiteltiin diabeteksen riskialttiutta ja potilaan kokemaa henkistä rasitusta sairauden kamppailemisen kanssa. Tyypin 1 diabetes on krooninen sairaus, joka vaatii potilaalta jatkuvaa huomiota ja hoidon säätelyä. Reaaliaikaisen ennustamisen ja automatisoidun hoidon kehitys voivat tarjota uusia työkaluja diabeteksen hoitoon ja auttaa lievittämään potilaan kokemaa kuormitusta. Voitaisiin myös mahdollistaa potilaan parempi omahoito, ehkäistä diabeteksen aiheuttamia komplikaatioita ja ylipäätään parantaa potilaan elämänlaatua.

Tutkielmassa käytiin läpi tehokkaimmat koneoppimisalgoritmit, jotka ovat tutkimusten perusteella suoriutuneet parhaiten glukoositasojen ennustuksessa. Optimoitu satunnaismetsä ja erilaiset tukivektorikoneet nousivat tuloksissa esiin eniten. Tehostamismalleja käsiteltiin myös. Mainittakoon, että menetelmät haluttiin rajata koneopin näkökulmasta, jolloin osa tehokkaista menetelmistä karsiutui automaattisesti pois.

Tulevaisuuden tutkimusta voidaan parantaa kehittämällä suurempia datasettejä ja huomioimalla elämäntapatekijöiden vaikutusta glukoositasoihin tarkemmin ja isommalla skaalalla. Monitorointilaitteiden tarkkuutta tulee myös kehittää. Vaikka lisätutkimuksia tarvitaan ennustamisen tarkkuuden ja yleistettävyyden parantamiseksi koneoppi on arvokas työkalu diabeteksen hoidossa nyt ja tulevaisuudessa.

Lähteet

Afsaneh, Elaheh, Amin Sharifdini, Hadi Ghazzaghi ja Mohadeseh Zarei Ghobadi. 2022. "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review". *Diabetology & Metabolic Syndrome* 14, numero 1 (27. joulukuuta 2022): 196. ISSN: 1758-5996. https://doi.org/10.1186/s13098-022-00969-9.

Atkinson, Mark A. 2012. "The Pathogenesis and Natural History of Type 1 Diabetes". *Cold Spring Harbor Perspectives in Medicine* 2, numero 11 (marraskuu): a007641. ISSN: 2157-1422. https://doi.org/10.1101/cshperspect.a007641.

Biau, Gérard, ja Erwan Scornet. 2016. "A random forest guided tour". *TEST* 25, numero 2 (1. kesäkuuta 2016): 197–227. ISSN: 1863-8260. https://doi.org/10.1007/s11749-016-0481-7.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer. ISBN: 978-0-387-31073-2.

Breiman, Leo, toimittanut. 1984. *Classification and regression trees*. Wadsworth statistics/probability series. New York: Chapman & Hall. ISBN: 978-0-534-98053-5 978-0-534-98054-2 978-0-412-04841-8.

———. 2001. "Random Forests". *Machine learning* 45 (1): 5–32. ISSN: 0885-6125. https://doi.org/10.1023/A:1010933404324.

Burges, Christopher J. C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition". *Data mining and knowledge discovery* 2 (2): 121–. ISSN: 1384-5810. https://doi.org/10.1023/A:1009715923555.

Cappon, Giacomo, Martina Vettoretti, Giovanni Sparacino ja Andrea Facchinetti. 2019. "Continuous Glucose Monitoring Sensors for Diabetes Management: A Review of Technologies and Applications". *Diabetes & Metabolism Journal* 43 (4): 383. ISSN: 2233-6079, 2233-6087. https://doi.org/10.4093/dmj.2019.0121.

Chen, Tianqi, ja Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System”. Teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. KDD ’16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 13. elokuuta 2016. ISBN: 978-1-4503-4232-2. https://doi.org/10.1145/2939672.2939785.

Cortes, Corinna, ja Vladimir Vapnik. 1995. “Support-vector networks”. *Machine Learning* 20, numero 3 (syyskuu): 273–297. ISSN: 0885-6125, 1573-0565. https://doi.org/10.1007/BF00994018.

Daniels, John, Pau Herrero ja Pantelis Georgiou. 2022. “A Multitask Learning Approach to Personalized Blood Glucose Prediction”. Conference Name: IEEE Journal of Biomedical and Health Informatics, *IEEE Journal of Biomedical and Health Informatics* 26, numero 1 (tammikuu): 436–445. ISSN: 2168-2208. https://doi.org/10.1109/JBHI.2021.3100558.

Dave, Darpit, Daniel J. DeSalvo, Balakrishna Haridas, Siripoom McKay, Akhil Shenoy, Chester J. Koh, Mark Lawley ja Madhav Erraguntla. 2020. “Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction”. *Journal of Diabetes Science and Technology* 15, numero 4 (1. kesäkuuta 2020): 842–855. ISSN: 1932-2968. https://doi.org/10.1177/1932296820922622.

Demšar, Janez, ja Blaž Zupan. 2021. “Hands-on training about overfitting”. *PLoS computational biology* 17, numero 3 (maaliskuu): e1008671. ISSN: 1553-7358. https://doi.org/10.1371/journal.pcbi.1008671.

DiMeglio, Linda A, Carmella Evans-Molina ja Richard A Oram. 2018. “Type 1 diabetes”. *The Lancet* 391, numero 10138 (kesäkuu): 2449–2462. ISSN: 01406736. https://doi.org/10.1016/S0140-6736(18)31320-5.

Facchinetti, Andrea. 2016. “Continuous Glucose Monitoring Sensors: Past, Present and Future Algorithmic Challenges”. *Sensors* 16, numero 12 (9. joulukuuta 2016): 2093. ISSN: 1424-8220. https://doi.org/10.3390/s16122093.

Freund, Yoav, ja Robert E Schapire. 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. *Journal of Computer and System Sciences* 55, numero 1 (elokuu): 119–139. ISSN: 00220000. https://doi.org/10.1006/jcss.1997.1504.

Friedman, Jerome H. 2001. “Greedy function approximation: A gradient boosting machine.” *The Annals of Statistics* 29, numero 5 (1. lokakuuta 2001). ISSN: 0090-5364. https://doi.org/10.1214/aos/1013203451.

Gadaleta, Matteo, Andrea Facchinetti, Enrico Grisan ja Michele Rossi. 2019. “Prediction of Adverse Glycemic Events From Continuous Glucose Monitoring Signal”. Conference Name: IEEE Journal of Biomedical and Health Informatics, *IEEE Journal of Biomedical and Health Informatics* 23, numero 2 (maaliskuu): 650–659. ISSN: 2168-2208. https://doi.org/10.1109/JBHI.2018.2823763.

Gandhi, Gunjan Y, Michelle Kovalaske, Yogish Kudva, Kristin Walsh, Mohamed B Elamin, Melody Beers, Cathy Coyle ym. 2011. “Efficacy of Continuous Glucose Monitoring in Improving Glycemic Control and Reducing Hypoglycemia: A Systematic Review and Meta-Analysis of Randomized Trials”. *Journal of Diabetes Science and Technology* 5, numero 4 (1. heinäkuuta 2011): 952–965. ISSN: 1932-2968. https://doi.org/10.1177/193229681100500419.

Georga, Eleni I., Vasilios C. Protopappas, Diego Ardigò, Michela Marina, Ivana Zavaroni, Demosthenes Polyzos ja Dimitrios I. Fotiadis. 2013. “Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression”. Conference Name: IEEE Journal of Biomedical and Health Informatics, *IEEE Journal of Biomedical and Health Informatics* 17, numero 1 (tammikuu): 71–81. ISSN: 2168-2208. https://doi.org/10.1109/TITB.2012.2219876.

Haque, Chowdhury Azimul, Shifat Hossain, Tae-Ho Kwon ja Ki-Doo Kim. 2021. “Noninvasive In Vivo Estimation of Blood-Glucose Concentration by Monte Carlo Simulation”. *Sensors* 21, numero 14 (19. heinäkuuta 2021): 4918. ISSN: 1424-8220. https://doi.org/10.3390/s21144918.

Lekha, S., ja Suchetha M. 2018. “Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network”. Conference Name: IEEE Journal of Biomedical and Health Informatics, *IEEE Journal of Biomedical and Health Informatics* 22, numero 5 (syyskuu): 1630–1636. ISSN: 2168-2208. https://doi.org/10.1109/JBHI.2017.2757510.

Maahs, David M., Nancy A. West, Jean M. Lawrence ja Elizabeth J. Mayer-Davis. 2010. “Epidemiology of Type 1 Diabetes”. *Endocrinology and Metabolism Clinics of North America* 39, numero 3 (syyskuu): 481–497. ISSN: 08898529. https://doi.org/10.1016/j.ecl.2010.05.011.

Marling, Cindy, ja Razvan Bunescu. 2020. “The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020”. *CEUR workshop proceedings* 2675 (syyskuu): 71–74. ISSN: 1613-0073. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/.

Mehmood, Sohaib, Imran Ahmad, Hadeeqa Arif, Umm E. Ammara ja Abdul Majeed. 2020. “Artificial Pancreas Control Strategies Used for Type 1 Diabetes Control and Treatment: A Comprehensive Analysis”. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, *Applied System Innovation* 3, numero 3 (syyskuu): 31. ISSN: 2571-5577. https://doi.org/10.3390/asi3030031.

Mosquera-Lopez, Clara, Robert Dodier, Nichole S. Tyler, Leah M. Wilson, Joseph El Youssef, Jessica R. Castle ja Peter G. Jacobs. 2020. “Predicting and Preventing Nocturnal Hypoglycemia in Type 1 Diabetes Using Big Data Analytics and Decision Theoretic Analysis”. *Diabetes Technology & Therapeutics* 22, numero 11 (1. marraskuuta 2020): 801–811. ISSN: 1520-9156, 1557-8593. https://doi.org/10.1089/dia.2019.0458.

Mujahid, Omer, Ivan Contreras ja Josep Vehi. 2021. “Machine Learning Techniques for Hypoglycemia Prediction: Trends and Challenges”. *Sensors* 21, numero 2 (14. tammikuuta 2021): 546. ISSN: 1424-8220. https://doi.org/10.3390/s21020546.

Nevo-Shenker, Michal, ja Shlomit Shalitin. 2021. “The Impact of Hypo- and Hyperglycemia on Cognition and Brain Development in Young Children with Type 1 Diabetes”. *Hormone Research in Paediatrics* 94 (3): 115–123. ISSN: 1663-2818, 1663-2826. https://doi.org/10.1159/000517352.

Roglic. 2016. “WHO Global report on diabetes: A summary”, kesäkuu. https://www.ijncd.org/text.asp?2016/1/1/3/184853.

Seo, Wonju, You-Bin Lee, Seunghyun Lee, Sang-Man Jin ja Sung-Min Park. 2019. “A machine-learning approach to predict postprandial hypoglycemia”. *BMC Medical Informatics and Decision Making* 19, numero 1 (6. marraskuuta 2019): 210. ISSN: 1472-6947. https://doi.org/10.1186/s12911-019-0943-4.

Sparacino, Giovanni, Francesca Zanderigo, Stefano Corazza, Alberto Maran, Andrea Facchinetti ja Claudio Cobelli. 2007. “Glucose Concentration can be Predicted Ahead in Time From Continuous Glucose Monitoring Sensor Time-Series”. Conference Name: IEEE Transactions on Biomedical Engineering, *IEEE Transactions on Biomedical Engineering* 54, numero 5 (toukokuu): 931–937. ISSN: 1558-2531. https://doi.org/10.1109/TBME.2006.889774.