Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Brigitta Rebane

Rekurrentsed tehisnärvivõrgud inimeste kliiniliste trajektooride ennustamisel

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: Raivo Kolde, *PhD*

Tartu 2020

REKURRENTSED TEHISNÄRVIVÕRGUD INIMESTE KLIINILISTE TRAJEKTOORIDE ENNUSTAMISEL

Bakalaureusetöö

Brigitta Rebane

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on luua tõenäosuslik mudel, mis prognoosib patsientidele potentsiaalseid kliinilisi trajektoore. Esmalt tutvustatakse töös kasutatavate meetodite olemust. Seejärel antakse ülevaade töös kasutatavatest andmetest ning nende ettevalmistamisest tulemuste prognoosimiseks. Lõpuks esitatakse saadud tulemusi ning nende tõlgendusi.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: rekurrentne neurovõrk, kliiniline trajektoor, tõenäosuslik mudel.

RECURRENT NEURAL NETWORKS IN PREDICTING HUMAN CLINICAL TRAJECTORIES

Bachelor thesis

Brigitta Rebane

Abstract

The goal of this thesis is to create a probabilistic model that predicts patients' potential clinical trajectories. Firstly the thesis introduces used methods. In addition, an overview of used data is given with the explaination of how they were prepared for predicting. Ultimately the results are examined and it is shown how to interpret them.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics

Key Words: recurrent neural network, clinical trajectory, probabilistic model.

Sisukord

Sissejuhatus

Haigused on osa elust. Antud töö eesmärk on luua mudel, mis võimaldab ennustada eelneva kliinilise trajektoori põhjal potentsiaalseid terviseprobleeme. Sel viisil saab diagnoose ennetada, vajadusel oma elustiili muuta ning pikemalt kvaliteetset elu elada.

Bakalaureusetöös luuakse e-tervise andmete abil tõenäosuslik mudel, mis suudab genereerida patsiendi eelneva haigusloo põhjal potentsiaalseid personaalseid edasisi trajektoore. Viimaste põhjal saab ülevaate, millal ja millistesse haigustesse on patsientidel suurem tõenäosus haigestuda. Antud töös treenitakse ajaliselt järjestatud andmetel tehisnärvivõrkudega mudel. Andmed pärinevad Haigekassast.

Töö koosneb kahest osast: teoreetilisest ja praktilisest. Teoreetiline osa koosneb kolmest alapeatükist. Esimeses alapeatükis antakse ülevaade masinõppe üldisest taustast ja olemusest. Teises alapeatükis kirjeldatakse tehisnärvivõrke ja neuronite tööpõhimõtet. Kolmandas alapeatükis tutvustatakse rekurrentseid tehisnärvivõrke ning töös kasutatud võrgutüüpe.

Praktilises osas antakse ülevaade töös kasutatavatest andmetest ja avatakse nende sisu. Tutvustatakse kasutatud tunnuseid, diagnoose ja andmete töötlemist. Lisaks kirjeldatakse mudeli loomiseks vajalikke andmete kujusid ja mudeli komponente. Järgnevalt konstrueeritakse tehisnärvivõrkude abil diagnoosiandmetest edasisi ravitrajektoore. Tehakse ülevaade saadud tulemustest, nende kasutamisvõimalustest ja tõlgendamisest.

Andmete töötluseks ning tehisnärvivõrkude treenimiseks kasutati rakendustarkvara R. Töö on kirjutatud ja vormistatud tekstitöötlusprogrammiga LaTex.

Autor tänab bakalaureusetöö juhendajat Raivo Koldet pühendatud aja, rohkete nõuannete ja selgituste eest.

1 Masinõpe

Järgnevas peatükis kirjeldatakse masinõpet, täpsemalt rekurrentseid tehisnärvivõrke ja töös kasutatud meetodeid.

1.1 Ennustamine masinõppes

Chollet ja Allaire [1] väidavad, et masinõppe põhimõte taandub sisendi ja väljundi defineerimisele. Klassikalises programmeerises on sisendiks andmed ja reeglid ning arvuti väljastab nende põhjal soovitud vastuse. Masinõppes on sisendiks aga andmed ja vastused ning arvuti väljastab reeglid. Seetõttu ei nimetata masinõppesüsteemi programmeerimiseks vaid treenimiseks [1, lk. 5].

1.1.1 Treening- ja testhulk

Enne treenimist jagatakse andmestik juhuslikult kaheks: treenimis- ja testhulgaks. Chollet ja Allaire [1] toovad välja, et hea tava järgi moodustab treeninghulk 80% ning testhulk 20% kogu andmestikust. Treenimisandmestiku abil sobitatakse mudel uute andmete prognoosimiseks. Sobitatakse erinevaid funktsioone ning argumente, et saavutada maksimaalne täpsus. Testandmeid mudeli loomisel ei kasutata, nende peal vaid testitakse lõpliku mudeli headust. Nii saab kontrollida, kas mudel töötab hästi ka uutel andmetel [1, lk. 88].

Lisaks koosneb nii treenimis- kui ka testhulk omakorda kahest osast: sisend- ja väljundandmetest. Treeninghulga sisendandmed on kogu treeninghulk ilma prognoositava tunnuseta. Väljundandmed moodustavad tunnused, mida soovitakse sisendandmete põhjal ennustada. Seega treeninghulga abil treenitakse ja sobitatakse mudel, mis sisendandmete põhjal prognoosib väljundandmeid. Rakendades mudelit testhulga sisendandmetel, saadakse prognoositud väljundandmete väärtused [1, lk. 26]. Mida suurem hulk prognoositud väärtuseid

ühtib tegelike testhulga väljundväärtustega, seda täpsem on mudel. Suurema täpsuse korral on ka prognoositud tulemused tähenduslikud.

1.1.2 Hea mudel

On ilmne, et hea mudeli eesmärk on ennustada võimalikult täpselt oodatavaid väärtuseid. Salehinejadi jt [2] sõnul vaadeldakse mudeli headuse hindamisel treeninghulga ja testimishulga kao vahet. Kui treeninghulgal mõõdetud kadu on väiksem testhulga kaost, võib mudel olla ülesobitatud [2, lk. 2]. Viimane tähendab, et mudel õppis ära kindlad seosed sisend-ja väljundandmetel, kuid need seosed ei kehti uutel erinevatel andmetel. Seega tuleb mudeli loomisel kadu jälgida, sest halva mudeli korral pole saadud tulemusted väärtuslikud.

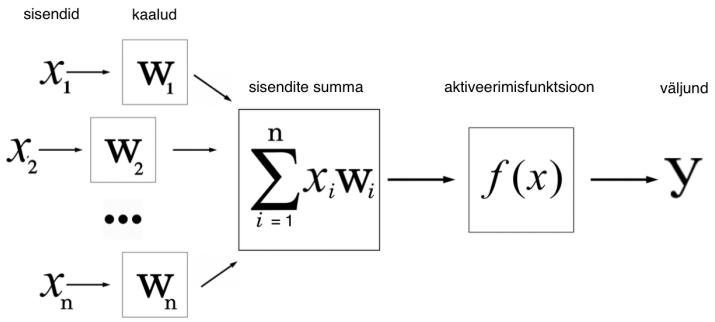
1.2 Tehisnärvivõrgud

Edasises peatükis kirjeldatakse neuroneid ning tehisnärvivõrke. Eelnevate paremaks mõistmiseks antakse ka lühiülevaade kaofunktsioonidest ja optimiseerimisalgoritmidest.

1.2.1 Neuron

Järgnevas alapeatükis tutvustatakse tehisnärvivõrgu peamist komponenti-neuronit. Kirjeldatakse selle tööpõhimõtet ja struktuuri.

Petlenkov [3] toob välja, et tehisnärvivõrk on nimetatud bioloogilise närvivõrgu mudeli järgi, sest see järgib ka viimase tööpõhimõtet. Tehisnärvivõrk koosneb neuronitest ning nendevahelistest ühendustest [3, lk. 3]. Bioloogiline neuron on keeruline süsteem ning edaspidi kirjeldatakse tehisneuronit, mis on bioloogilise neuroni lihtsustatud matemaatiline mudel.



Joonis 1: Tehisneuron.

Joonise 1 põhjal on igal neuronil n sisendit, mis on uuritavat objekti kirjeldavad tunnused. Antud töös on tunnusteks erinevad diagnoosid, vanus ja kvartal patsiendi haigestudes. Igal sisendil on ka kaal, millega korrutades kaalutakse kõik sisendite väärtused [3, lk. 4]. Kaalud leitakse kaofunktsiooni abil, mida kirjeldatakse järgmises alapeatükis.

Pärast sisendite kaalumist ning nende summeerimist rakendatakse saadud tulemusel aktiveerimisfunktsiooni. Viimase eesmärk on neuronite väljundeid mõistlikes piirides hoida. Levinuimad aktiveerimisfunktsioonid on sigmoidfunktsioonid, mille graafikud on s-tähe kujulised. Aktiveerimisfunktsiooni tulemus ongi antud neuroni väljundväärtus, mis viiakse järgmise neuroni sisendiks [3, lk. 5].

1.2.2 Kaofunktsioon ja optimeerimine

Kaofunktsioon väljastab vea oodatud ja saadud vastuste vahel. Õppimisprotsessi käigus, mitme iteratsiooni jooksul uuendatakse neuronites sisendite kaale nii, et viga oleks minimaalne [3, lk. 14]. Seega leitakse kaofunktsiooniga närvivõrgu sobivus andmetega, kuna võrreldakse prognoositud vastust olemasoleva väljundväärtusega.

Kaalukoefitsendid leitakse optimeerimisalgoritmiga nii, et viga oleks vähim.

Optimeerimisalgoritm töötab koos kaofunktsiooniga, sest viimase tulemuse põhjal leitakse tunnustele uued kaalud. Enim tuntud optimiseerimisalgoritm on gradientlaskumise (ingl *gradient descent*) meetod [2, lk. 3].

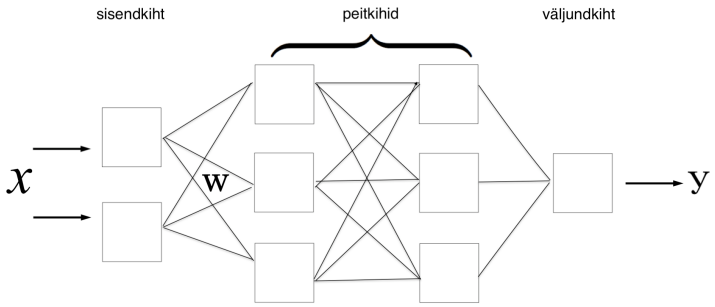
Ruderi [4] sõnul gradientlaskumise meetod uuendab kaalusid igal iteratsioonil kogu andmestikule. Selle eesmärk on uuendada kaale leides tuletise vea iga tunnuse kohta [4, lk. 1, 2]. Salehinejad jt [2] toovad lisaks välja, et optimiseerimise eelduseks on aktiveerimisfunktsiooni diferentseeruvus kõikide tunnuste suhtes. Seega kao minimiseerimiseks uuendatakse kaalud sõltuvalt vigade tuletistest kaalu suhtes [2, lk. 3].

Mudeli treenimisel määratakse ka ploki suurus (ingl *batch-size*) ning epohhid (ingl *epoch*). Chollet ja Allaire'i [1] sõnul moodustatakse treeningandmetest plokisuuruse põhjal väiksemad hulgad, mille kaupa andmete peal mudelit treenima hakatakse. Epohhide arv määrab, mitu korda itereeritakse kogu treenimist treeningandmestikul [1, lk. 32]. Seega pärast iga epohhi uuendatakse kaalud kaofunktsiooni tulemuse järgi, et kadu vähendada.

1.2.3 Tehisnärvivõrgu struktuur

Tehisnärvivõrk koosneb neuronitest, kus viimased on omavahel kihtide kaupa ühenduses. Tavalises tehisnärvivõrgus on üks sisendkiht, üks peitkiht ning üks väljundkiht. Mida rohkem lisatakse peidetud kihte, seda komplekssem on neurovõrk. Kihtide arvukus suurendab võrgu sügavust ning mitmete erinevate kihtidega neurovõrguga treenimist nimetataksegi sügavõppeks [2, lk. 1].

Joonise 2 põhjal on võrgul üks sisendkiht, üks väljundkiht ning peidetud kihid. Sisendkihis jaotatakse sisendid neuronite vahel. Edasi kaalutakse sisendid ning saadetakse peitkihtidesse. Iga kaalukoefitsent on määratud kolme indeksiga: neuroni sisendi number, neuroni järjekorra number kihis ning kihi number [3, lk. 8]. Petlenkovi [3] sõnul pärast sisendi kaalumist ja summeerimist lisatakse tulemusele neuroni nihe θ_{ij}. Saadud summad moodustavad ühe peitkihi neuronite



Joonis 2: Tehisnärvivõrk.

väljundid. Analoogiliselt toimib sama protsess kuni väljundkihi neuronite väljundväärtustest moodustatakse neurovõrgu väljundvektor [3, lk. 8].

1.3 Rekurrentsed tehisnärvivõrgud

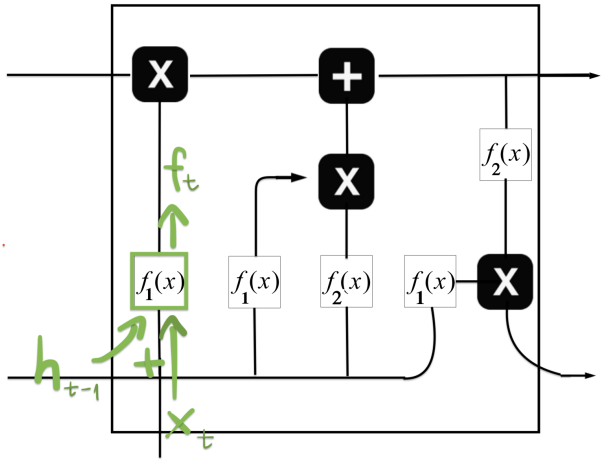
Rekurrentsetes tehisnärvivõrkudes (edaspidi RTN) on samuti kolme eri tüüpi kihte: sisendkiht, rekurrentsed peitkihid ja väljundkiht. Erisus tavalistest tehisnärvivõrkudest seisneb selles, et peitkihtidesse talletatakse läbitud informatsioon. See tähendab, et peitkihi mingi seisund sõltub selle eelmisest seisust [2, lk. 1].

Chollet [1] väidab, et RTN-d on sobivaimad järjestikuste andmetega töötamiseks. Viimasteks on tavaliselt tekstijärjestused, aga antud töös ajaperioodid. Neurovõrgu olek lähtestatakse iga erineva järjestuse, antud töös iga patsiendi haigusloo vahel [1, lk. 180]. Praktilises osas vaadeldi ravi algust kvartali täpsusega ning autor defineeris üheks andmepunktiks ühe patsiendi 12 kvartali pikkuse haigusloo. Järelikult töös loetakse iga patsiendi 12 kvartali pikkust haiguslugu ka üheks võrgu sisendiks.

Järgnevates alapeatükkides kirjeldatakse täpsemalt kahte rekurrentse tehisnärvivõrgu tüüpi. Erinevaid tehisnärvivõrgu tüüpe kasutatakse mudelites kihtidena. Kihid valitakse mudelisse andmete kuju ja soovitud tulemuste kuju põhjal.

1.3.1 LSTM

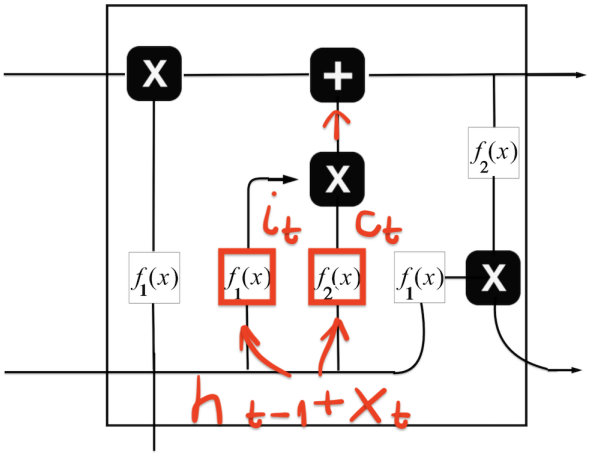
Üheks rekurrentse närvivõrgu tüübiks on pikk lühimälu (ingl *Long Short-Term Memory*), edaspidi tähistatud LSTM. Eelneva eesmärk on mällu talletada sekventsiaalseid seoseid, sest lihtsate RTN võrkude puhul ei suuda mudel õppida järjestikuseid pikki seoseid. Peitkihi neuronites on RTN-ga võrreldes sigmoidfunktsioonide asemel mälurakud. Viimased kontrollivad sisendi ja väljundi infovoogusid, et gradient plahvatuslikult ei suureneks. Eelneva abil suudab närvivõrk hoida meeles pikemat järjestust ning selle põhjal õppida [2, lk. 9]. Järgnevas alapeatükis kirjeldatakse vaadeldava võrgutüübi tööpõhimõtet täpsemalt.



Joonis 3: Mäluraku mälu värav.

Joonisel 3 on kujutatud LSTM mälurakku ning rohelisega mälu väravat (ingl *forget gate*) tööpõhimõtet. Funktsioon tähisega $f_1(x)$ märgib logistilist sigmoidfunktsiooni ning $f_2(x)$ märgib hüperboolset tangensi funktsiooni. Joonistel 3, 4, 5 ja 6 on kujutatud erinevate väravate tööpõhimõtted eri värvidega, sest autori arvates on keeruline protsess selliselt kergemini hoomatav. Kõik tähised tähistavad erinevate operatsioonide sisend- ja väljundvektoreid.

Phi [5] sõnul esmalt mälurakus liidetakse eelmine peitkihi seisund h_{t-1} ning sisend x_t, vt joonis 3. Seejärel saadud summa saadetakse mäluväravasse ehk summal rakendatakse sigmoidfunktsooni. Viimase väärtused jäävad vahemikku $(0, 1)$, seda kirjeldatakse täpsemalt järgmises alapeatükis. Kusjuures, mida lähemal on saadud väärtus arvule 0, seda suurema tõenäosusega jäetakse tunnus mudelist välja. Mida lähemal arvule 1, seda tõenäolisemalt jäetakse tunnus mudeli mällu. Eelnevast tuleb ka vastava värava nimetus [5]. Joonisel 3 on tähistatud mäluvärava väljund f_t rohelisega.

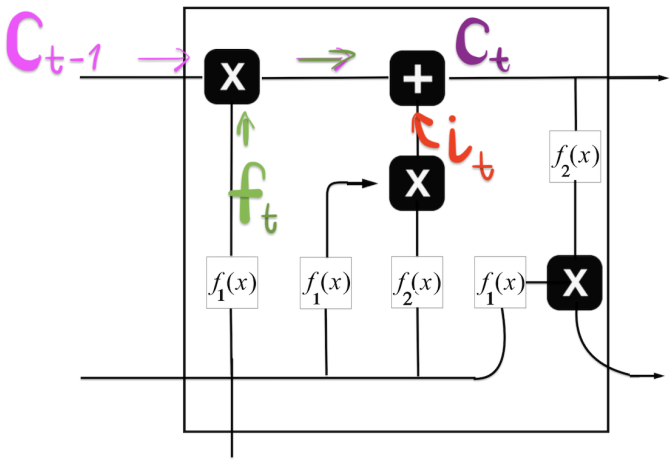


Joonis 4: Mäluraku sisendi värav.

Järgmisena viiakse Phi [5] sõnul sisendi väravasse (ingl *input gate*) eelmise peitkihi seisundi ja sisendi summa, vt joonis 4. Sisendi väravas rakendatakse summal esmalt logistilist sigmoidfunktsiooni, mille abil otsustatakse analoogiliselt eelmise sammuga, kas tunnus on oluline. Teisena rakendatakse summal hüperboolset tangensi funktsiooni, mille väärtused jäävad vahemikku $(-1, 1)$. Salehinejad jt [2] sõnul hoiab funktsioon 1 väärtused ühtlases muutumispiirkonnas $(-1, 1)$. Viimase funktsiooni kuju avaldub

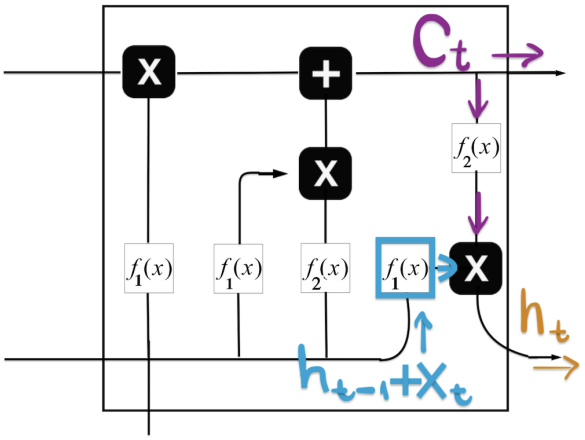
$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{1}$$

Saadud funktsioonide väljundväärtused, vastavalt i_t ja c_t, korrutatakse omavahel ning saadetakse edasi. Eelneva tööpõhimõte on märgitud punasega, vt joonis 4. Korrutise väärtus on samuti vahemikus $(-1, 1)$ ning logistilise sigmoidfunktsiooni väärtus määrab, kas korrutis on mudelis oluline. See tähendab, et kui sigmoidfunktsiooni väljund on nullilähedane, siis on seda ka eelnevate funktsioonide väärtuste korrutis ning tunnust ei jäeta mäluraku mällu.



Joonis 5: Mäluraku olek.

Phi [5] sõnul on pärast kahte eelnevat sammu piisavalt informatsiooni raku oleku (ingl *cell state*) uuendamiseks. Seega joonisel 5 kirjeldatakse raku oleku muutust eelmisest olekust. Esmalt korrutatakse raku eelmine olek c_{t-1} mäluvärava tulemusega f_t [5]. Seega kui mingi mäluvärava väljund on nullilähedane, jäetakse mudelist välja vastavad raku eelmise oleku tunnused. Järgnevalt summeeritakse viimane tulemus ja sisendvärava väärtused i_t ning saadakse raku uus olek c_t, vt joonis 5.



Joonis 6: Mäluraku väljundvärav.

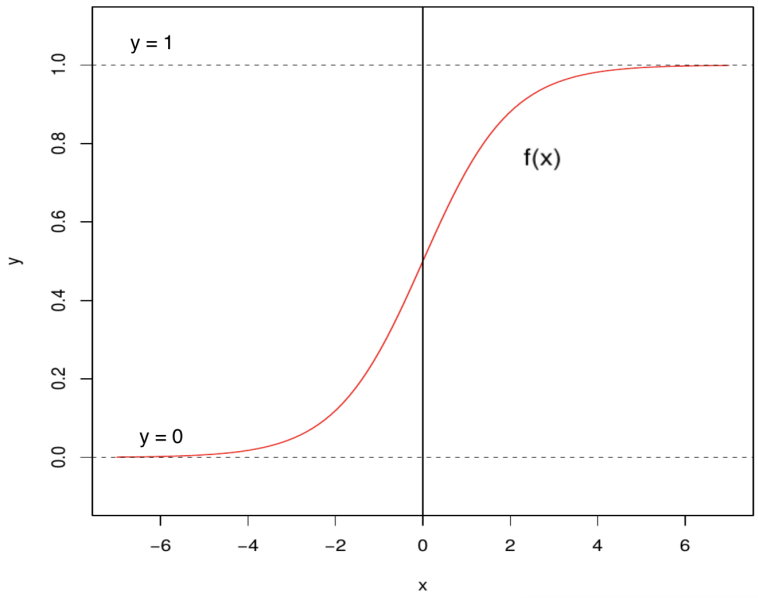
Väljundväravas (ingl *output gate*) suunatakse Phi [5] sõnul esmalt eelmise peitkihi h_{t-1} ja sisendi x_t summa sigmoidfunktsiooni. Edasi uuel rakul c_t rakendatakse tanh funktsiooni. Kaks eelnevat tulemust korrutatakse ning saadakse oluliste tunnustega peitkihi väljund h_t [5]. Joonise 6 põhjal on LSTM-i mäluraku väljundiks raku uus olek c_t ning mudelis oluliste peitväärtuste vektor h_t.

1.3.2 Täissidus kiht

Täissidusa (ingl *densely connected layer*) närvivõrgu eesmärk on leida võimalikult palju seoseid sisendi tunnuste vahel. Viimasest tuleb ka antud võrgu nimetus, sest erinevalt teistest tehisnärvivõrkudest, leitakse seoseid kõikide tunnuste, mitte ainult lokaalsete vahel. Lisaks kasutatakse võrku tavaliselt viimases kihis binaarse klassifikaatorina [1, lk. 114]. Vastavas kihis kasutatakse tavaliselt aktiveerimisfunktsioonina logistilist sigmoidfunktsiooni. Viimase kuju avaldub Salehinejad jt [2, lk. 2] sõnul järgmiselt:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

Sigmoidfunktsiooni 2 kuju järgi võib märgata, et selle väärtused jäävad 0 ja 1 vahele. Tegur $e^{-x} > 0$, sest $e > 0$, seega $e^{-x} > 0 \xrightarrow{+1} 1 + e^{-x} > 1$. Järelikult on funktsiooni nimetaja suurem lugejast ja väärtused iga x korral jäävad 0 ja 1 vahele. Eelnev arutluse tulemus avaldub ka funktsiooni kuju graafikul, vt joonis 7.



Joonis 7: Logistiline sigmoidfunktsioon.

2 Mudeli loomine

Järgnevas peatükis antakse ülevaade antud töö praktilises osas loodud ennustavast mudelist. Kirjeldatakse kasutatud andmeid, nende ettevalmistamist ja töötlemist. Lisaks selgitatakse mudeli loomist ja tutvustatakse selles kasutatud argumente.

2.1 Andmed

Töös kasutatavad ja uuritavad andmed pärinevad Haigekassast ning tegemist on rinnavähi kohordiga. Kogu andmestik koosneb kolmest tabelist: diagnoosi info, ravimise info ning patsiendi info. Eelnevate tabelite liitmisel saadi töös kasutatav koondtabel, mis koosneb 472751 reast ehk patsiendist.

2.1.1 Tunnuste kirjeldus

Tehisnärvivõrkude treenimisel kasutati antud töös järgmiseid tunnuseid:

- patsiendi id
- diagnoosi kood
- ravi algus- ja lõppkuupäev
- patsiendi vanus ravi alguses.

Isikuandmete kaitse seaduse alusel on tunnus patsiendi id mitmekordselt krüpteeritud ning ümber paigutatud [6]. Antud andmestiku põhjal ei ole võimalik ühendada patsiendi identifitseerimisnumbrit konkreetse isikuga. Patsiendi id tunnuse abil eristati mudeli loomisel erinevaid patsiente.

Diagnoose on registreeritud alates 2008. aasta märtsist kuni 2016. aasta detsembri lõpuni. Ravi algus ja lõpp olid märgitud kuupäeva täpsusega. Ravi lõppkuupäeva edaspidi ei uurita, sest antud töös loodud mudelis ravi pikkust ei arvestata.

Ravi alguskuupäeva tunnus kodeeriti ümber üldisemale kujule. Edaspidi vaadeldakse ravi algusena selle algusaastat kvartali täpsusega. Unikaalseid aasta ja kvartali kombinatsioone on kokku 32.

Ravi alguskuupäeva ja sünniaasta abil arvutati lisatunnus- patsiendi vanus ravi alguses. Eelnevat kasutati iga patsiendi kohta lisatunnusena mudeli loomisel.

2.1.2 Diagnoosid

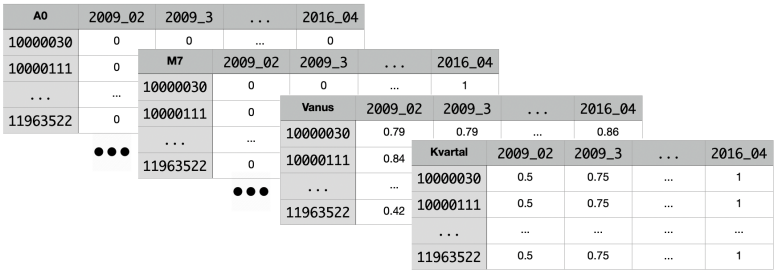
Diagnoosid on kodeeritud rahvusvahelistes RHK-10 diagnoosikoodides. Kood koosneb kolmest osast: tähega märgitud 22 erinevat haiguse gruppi, kahekohalise arvuga häiritud kehaosa ning punktiga eraldatud kahekohaline arv, mis täpsustab konkreetset diagnoosi. Lisa 8 põhjal on kõige üldisemalt diagnoosid grupeeritud 21 rühma. Erinevaid diagnoose on andmestikus kokku 5583.

Antud andmete hulk Eesti rahvastiku põhjal on ebapiisav iga vastava konkreetse diagnoosi prognoosimiseks, sest erinevaid patsiente on andmestikus vaid 12850. VanVoorhise ja Morgani sõnul on mudeli loomiseks vajalik hulk rusikareegli järgi vähemalt $\sim$ 50 rida iga parameetri kohta [7]. Seega vajalik andmehulk oleks täpsete diagnooside prognoosimiseks vähemalt 279150 realine ehk umbes 21 korda suurem. Järelikult prognoosi reliaabluse säilitamiseks uuritakse edaspidi üldistatud diagnoose: täht ning RHK kood kümnendi täpsusega. Sellisel kujul on erinevate diagnooside hulk 194.

Leidus juhtumeid, kus ühel ja samal kuupäeval oli isikule registreeritud mitu erinevat diagnoosi. Korrektsuse mõtte ühe ja sama haiguse korduva diagnoosi välistamiseks kasutati sellistel puhkudel antud töös vaid peamist visiidil registreeritud diagnoosi.

2.1.3 Andmete töötlemine

Andmed valmistati treenimiseks ette viies need kolmemõõtmelisele kujule. Selguse mõttes tähistatakse edaspidi read tähega X, veerud tähega Y ning kolmas mõõde tähega Z. Antud töös on kolmas mõõde erinevate kahemõõtmeliste tabelite kogum, vt joonis 8.



Joonis 8: Andmete kuju.

Mõõtme X iga rea moodustab unikaalse identifitseerimiskoodiga patsient, selliseid ridu on 12850. Joonise 8 järgi on iga erineva koodiga patsiendi kohta igas mõõtme Z dimensioonis rida, milles on andmed iga kvartali kohta.

Mõõde Y tähistab ravi algust kvartali täpsusega. Seega iga veergu märgib unikaalne aastaarv ning kvartal, selliseid veerge on kokku 32. Sissekandeid on 2008. aasta kohta vaid vastava aasta esimesest kvartalist, seetõttu edaspidi jäetakse andmetest välja 2008. aasta andmed. Aasta 2009 kohta on sissekanded kolmest viimasest kvartalist. Aastatel 2010-2016 on registreeritud patsiente igas neljas kvartalis. Joonise 8 põhjal on seega uuritavad kvartalid aasta 2009 teisest kvartalist kuni 2016. aasta neljanda kvartalini.

Mõõde Z koosneb kõikidest diagnoosidest, patsiendi vanusest ravi alguses ning vastavast kvartalist, mil ravi algas. Seega mõõtmel Z on kokku 196 dimensooni, millest 194 on erinevad diagnoosid.

Kolmemõõtmelisele kujule viies kodeeriti diagnooside andmed binaarsesse arvusüsteemi: iga väärtus on kas 0 või 1. See tähendab, et iga registreeritud diagnoosi kohta on uues kolmemõõtmelises tabelis vastaval kohal arv 1, mujal 0, vt joonis 8. Vanuste ja kvartalite väärtused on erinevates vahemikes, mistõttu tuli need normaliseerida [1]. Seega viidi viimased tunnused üle väärtustele vahemikus $0 \ldots 1$.

Joonise 8 põhjal patsiendil koodiga 10000030 aasta 2009 teises kvartalis haigust $A0$ ei diagnoositud, ta oli sel hetkel 79-aastane ning see oli aasta teine kvartal. Küll aga diagnoositi eelneval patsiendil aasta 2016 neljandas kvartalis haigus diagnoosiga $M7$. Ta oli sel hetkel 86-aastane ning see oli aasta viimane kvartal.

2.2 Mudel

Järgnevas alapeatükis antakse ülevaade andmete ettevalmistamisest mudeli loomiseks ja loodud mudeli komponentidest. Kirjeldatakse andmete jaotamist erinevateks alamhulkadeks ja nende kuju muutuseid selle käigus. Viimaseks tutvustatakse töös loodud mudelit, milliseid kihte ja funktsioone selles kasutati.

2.2.1 Treening- ja testhulk

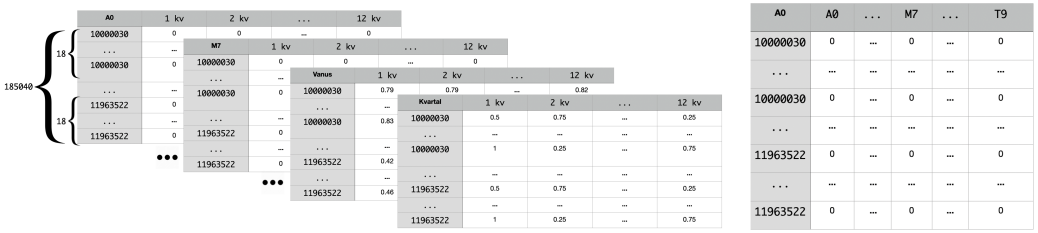
Kasutades tagasipanekuta juhuslikku valikut, leiti juhuslikud arvud 80% üldkogumi reaindeksite ulatuses. Treeningandmestik koostati üldkogumi ridadest, mille indeksid ühtisid juhuslikult leitud arvudega. Testandmestik koostati üldkogumi ülejäänud ridadest. Järgmisena jagati nii treening- kui ka testhulk sisend- ja väljundandmestikeks. Konkreetsuse mõttes kirjeldatakse järgnevalt treeninghulga töötlemist, testhulk jaotatakse analoogiliselt.

2.2.2 Sisend- ja väljundandmed

Järgmisena kirjeldatakse treening- ja testhulga jaotamist sisend-ja väljundandmestikeks. Antud töös on sisendandmestik kolmemõõtmeline tabel. Sarnaselt üldkogumiga on sellel kolm dimensiooni: esimene mõõde koosneb patsientidest, teine mõõde kvartalitest ning kolmas mõõde diagnoosidest, vanusest haigestudes ning kvartalist haigestudes.

Töös loodud mudeli eesmärk on iga rea ehk patsiendi kohta antud kliinilise trajektoori põhjal ennustada järgmise kvartali haigestumised. Defineeriti seeria, mis koosneb 12 kvartalist. Kokku oli algses andmestikus 32 erinevat kvartalit, kuid edaspidi kasutati neist 31, sest esimene oli ainus kvartal 2008. aasta kohta.

Järjestikuses 31 kvartalite hulgas on seega kokku 18 järjestikust 12 kvartali pikkust seeriat. See tähendab, et sisendhulka lisati iga inimese kohta väärtused, kus tema kliiniline trajektoor on esimesest kvartalist (2009 2. kv) kuni 12. kvartalini (2012 1. kv). Järgmisena lisati iga inimese kohta tema trajektoor teisest kvartalist (2009 3. kv) kuni 13. kvartalini (2012 2. kv). Selliselt lisati iga inimese kohta 18 trajektoori kuni lisati trajektoor 19. kvartalist (2013 3. kv) kuni 31. kvartalini (2016 2. kv).



(a) Sisendandmed. **(b)** Väljundandmed

Joonis 9: Treeninghulk.

Sisendandmestikus on järelikult 185040 rida, sest see on põhiandmestiku 80% ridade 18-kordne. Iga patsiendi kohta on 18 rida ehk 18 eri aegadel kliinilist trajektoori. Teises mõõtmes on 32 veeru asemel 12 veergu, sest prognoosimiseks

defineeriti 12 kvartali pikkune seeria, vt joonis 9a.

Väljundandmestik on maatrikskujul, sest teine dimensioon koosneb vaid ühest veerust. Ridade arv on sisendandmestikuga sarnaselt 185040, veerge on 194 ehk iga diagnoosi kohta. Väljundandmestikus patsiendi vanust ja kvartalit ravi alguses ei arvestata, sest ennustatakse vaid diagnoose. Analoogiliselt sisendandmestikule vaadeldakse väljundandmestikus iga inimese kohta 18 kliinilist trajektoori.

Väljundandmestikus on iga inimese kohta diagnoosid esimesena 13. kvartali kohta, järgmisena 14. kvartali ning nii kuni 32. kvartalini, vt joonis 9b. Seega moodustavad sisendandmestiku 12 kvartali pikkused seeriad ning väljundandmestiku vastavalt iga seeriale järgneva kvartali andmed- koos moodustavad kaks andmestikku järelikult 13 kvartali pikkused seeriad. Analoogiliselt moodustatakse ka testhulgast nii sisend- kui ka väljundandmestik.

Joonise 9a esimese rea põhjal patsiendil koodiga 10000030 aasta 2009 teisest kvartalis, 2009. aasta kolmandas kvartalis ega 2012. aasta esimeses kvartalis haigust $A0$ ei diagnoositud. Ta oli selle 12 kvartali pikkuse perioodi alguses 79-aastane ning lõpus 82-aastane. Joonise 9b esimese rea põhjal ei diagnoositud tal järgmises kvartalis ehk 2012. aasta teises kvartalis samuti haigust koodiga $A0$.

2.2.3 Mudeli struktuur

Antud töös loodi selline mudel, millel on kaks kihti: pikk lühimälu kiht ja täissidus kiht. Järgnevas peatükis kirjeldatakse kasutatud argumente mudeli koostamisel.

Tabel 1: Loodud mudeli ülevaade.

Kihi tüüp	**Väljundi kuju**	**Parameetrid**
LSTM	(None, 128)	166400
täissidus	(None, 194)	25026

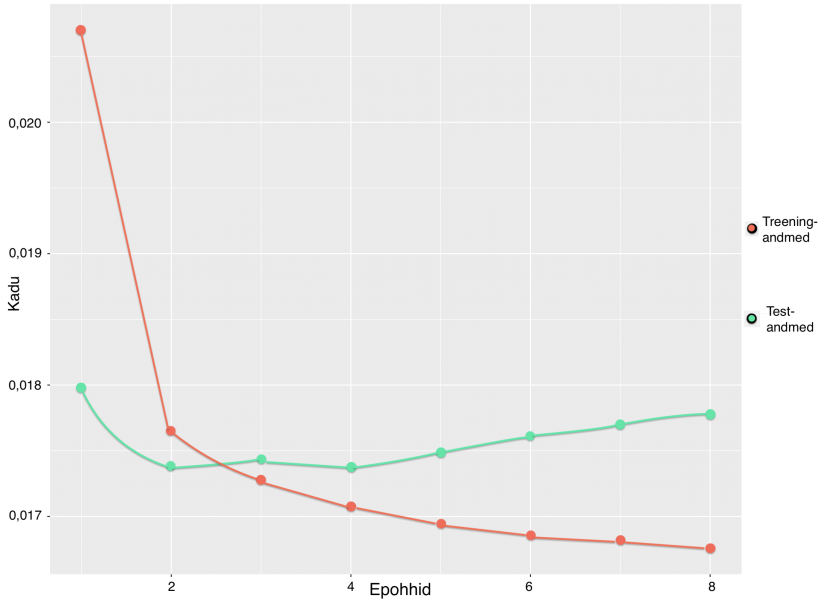
Tabeli 1 põhjal on esimeses kihis kasutatud LSTM võrgutüüpi. Teises tulbas on märgitud sisend ning väljundi kuju. Täpsemalt teine arv märgib väljundvektori pikkust. Seega LSTM kihi väljundiks on 128-elemendiline vektor, kuhu on esialgu mälu reserveerimiseks lisatud nulltüüpi väärtused. Nii saab prognoosimisel need väljundväärtustega üle kirjutada. Mudeli esimese kihi väljundi kuju nimetatakse ploki suuruseks. Järelikult rakendatakse LSTM-i treeninghulga 128-elemendistel plokkidel. Tabeli 1 viimases veerus on märgitud antud kihi kaalukoefitsentide hulk.

Teises kihis on väljundvektori pikkus 194. Enne diagnooside tõenäosuste prognoosimist on selles samuti kohad reserveeritud nullväärtusega, vt tabel 1. Kuna täissidus võrk on ka viimane kiht, siis see väljastab igale patsiendile 194 pikkuse vektori, kus iga element on ühe diagnoosi tekketõenäosus.

Mudeli õpisammuks määrati gradientlaskumise optimiseerimismeetodis $0,01$. Kaofunktsioonina kasutati selles binaarset ristentroopiakahju (ingl *binary crossentropy*) funktsiooni. Viimane töötab kõige optimaalsemalt kasutatavate andmete peal koos sigmoidfunktsiooniga [1, lk 105]. Teises ehk viimases kihis kasutatigi aktiveerimisfunktsioonina logistilist sigmoidi.

2.2.4 Kadu ja epohhid

Järgnevas alapeatükis kirjeldatakse kao ja epohhide arvu suhet. Vaadeldakse eraldi kadu nii treening- kui ka testandmete näitel.



Joonis 10: Kadu ja epohhide arv.

Joonisel 10 kujutatakse kao muutumist epohhide lõikes. Punasega on joonisel märgitud treeningandmetel mõõdetud kadu. Viimane väheneb iga epohhiga. See on oodatav tulemus, kuna gradientlaskumise optimiseerimisalgoritmi eesmärk ongi mudeli treenimisel iga epohhiga kadu minimiseerida.

Samas testandmete kadu tähistav roheline joon alates neljandast epohhist kasvab, vt joonis 10. Arvestades ülesobitamise ohtu oleks võinud mudeli epohhide arvu muuta kas kolmeks või neljaks. Siiski autori arvates kasvab testandmesiku kadu epohhide suurenedes aeglaselt ning vastavas piirkonnas vaid $0,0005$ ühiku võrra. Seega jäeti epohhide hulk mudelis kaheksaks.

3 Tulemused

Mudel treeniti väljastama patsientidele erinevate diagnooside tekketõenäosuseid. Edaspidi kirjeldatakse ennustamise protsessi ning kuidas tõenäosuseid uuel ennustamisel kasutada. Tuuakse tööprotsessi kirjeldamise selgitamiseks näited erinevate patsientide kohta.

3.1 Ennustamise protsess

Kliiniliste trajektooride prognoosimiseks loodi tsükkel, mis itereeris üle unikaalsete patsientide. Esmalt prognoosis mudel igale patsiendile tema kliinilise trajektoori põhjal iga haiguse tekketõenäosuse. Seejärel genereeriti tõenäosusele Bernoulli jaotusest realisatsioon. Viimast protsessi kirjeldatakse järgmises alapeatükis.

Tasub märkida, et treenimisel kasutati nihutatud aegridu, mille moodustavad iga patsiendi 18 kliinilist trajektoori. Iga diagnoosi ennustamise järel lisati andmestikku vastava diagnoosi kohta saadud tulemusega aegrida. Seega sai viimast kasutada järgmisel ennustusel.

3.1.1 Tulemuste tõlgendamine

Saadud diagnooside tõenäosused tõlgendatakse Bernoulli jaotuse abil eduks ja ebaeduks. Antud töös tähistatakse edu arvuga 1 ja see tähistab vastava haiguse tekkimist. Eelnev teisendus on vajalik, et diagnoosi tulemust järgmise aegrea prognoosimisel sisendina kasutada.

Täpsemalt ennustatakse igale patsiendile iga 194 diagnoosi kohta, kui tõenäoliselt talle vastav haigus järgmises kvartalis tekib. Arvestades tehisnärvivõrgu ehitust, vastab loodud mudeli viimases kihis üks neuron ühele diagnoosikoodile ehk ennustatavale tunnusele. Kuna viimased on binaarsel kujul, siis mudeli väljundiks

on tõenäosused vahemikus nullist üheni. Seega iga neuron treenitakse väljastama konkreetse diagnoosi tõenäosust.

Rakendustarkvara R baaskäskluse *rbinom* abil väljastatakse igale diagnoosile selle tekkimistõenäosuse põhjal Bernoulli jaotusest realisatsioon, kas 1 või 0. Seejuures tasub märkida, et genereeritud väärtused on vaid üks stsenaarium, mis on teatud määral juhuslik. Samadel tõenäosustel genereeritud väärtused ei pruugi olla identsed.

Tabel 2: Patsiendi 10021316 diagnooside prognoosid.

Diagnoos \ Kvartal	18. kv	19. kv	20. kv	21. kv	22. kv	23. kv	24. kv	25. kv	26. kv	27. kv	28. kv	29. kv	30. kv ennustus	30. kv tulemus
A0	0	0	0	0	0	0	0	0	0	0	0	0	0,00025	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
C5	1	1	1	1	1	1	1	1	1	1	1	1	0,63	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
E1	1	1	1	1	1	1	0	1	1	1	1	0	0,52	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
G3	0	0	0	0	0	0	0	0	0	0	0	0	0,000005	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
I3	0	0	0	0	0	0	0	0	0	0	0	0	0,00017	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
J9	0	0	0	0	0	0	0	0	0	0	0	0	0,005	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
M7	0	1	0	0	0	0	0	0	0	0	0	0	0,0048	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
S0	0	0	0	0	0	0	0	0	0	0	0	0	0,0026	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Z5	0	1	0	1	0	1	0	1	1	1	1	0	0,46	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
T9	0	0	0	0	0	0	0	0	0	0	0	0	0.00023	0

Tabelis 2 kujutatakse patsiendi 10021316 erinevate diagnooside viimast märgitud kliinilist trajektoori 18. kvartalist kuni 29. kvartalini. Seejärel diagnoosi ennustatud tekkimistõenäosust ning viimases veerus selle realisatsiooni. Välja on toodud patsiendi kõik diagnoosid, mis realiseeriti eduna. Lisaks seitsme erineva diagnoosi tulemused.

Nakkushaiguse diagnoosi pole varasemalt patsiendile ühelgi vastaval kvartalil

märgitud, kuid kasvaja on diagnoositud igas kvartalis. Lisaks on talle diagnoositud eri aegadel sisesekretsiooni-, toitumis- ja ainevahetushaigusi, lihasluukonna ja sidekoe haigus ning terviseseisundit mõjustavaid tegureid ja kontakte terviseteenistusega, vt tabel 2.

Tabeli 2 põhjal prognoositi järgmiseks kvartaliks suurima tekkimistõenäosusega 63% kasvaja. Tulemus on küllaltki kõrge ning antud tõenäosusele realiseeriti ka edu. Lisaks realiseeriti eduks ka kõrge prognoositud tõenäosusega 46% kontakt tervisteenistusega.

Tasub märkida, et tabelis 2 on patsiendil prognoositud kõrge tekkimistõenäosus 52% sisesekretsiooni-, toitumis- ja ainevahetushaigusele. Samas on eelneva tõenäosuse realisatsioon antud juhul 0. Viimane ilmestab realisatsioonide genereerimise juhuslikkust ning see on vaid üks võimalik stsenaarium.

Lisaks ennustati tabeli 2 põhjal diagnoosi J9 ehk hingamiselundite haiguse (lisa 8) tekkimistõenäosuseks $0,5\%$. Vaatamata väiksele tõenäosusele realiseeriti viimane eduks.

3.2 Prognooside varieeruvus

Järgnevas alapeatükis kirjeldatakse ühe noore väheste diagnoosidega patsiendi potentsiaalseid kliinilisi trajektoore. Teises alapeatükis ühe vanema hulgaliste diagnoosidega patsiendi potentsiaalseid trajektoore.

3.2.1 Näide väheste diagnoosidega patsiendi kohta

Patsiendi 10031105 vanus oli andmestiku põhjal eri aegadel 18 kuni 25 aastat. Tabeli 3 põhjal diagnoositi teda sellel ajaperioodil kokku 8 korral.

Antud patsient valiti näiteks oma diagnooside vähesuse tõttu. Praktilises osas uuriti, kas ja kuidas erinesid vastavale patsiendile ennustatud 10 erinevat kahe aasta pikkust kliinilist trajektoori.

Tabel 3: Patsiendi 10031105 diagnooside ajalugu.

Kvartal	**Diagnoos**
4. kv	O8, Z3
5. kv	Z3
6. kv	Z3
16. kv	N6
27. kv	M7
28. kv	M7, N6

Patsiendi esimesed diagnoosid on märgitud 4. kvartalil ehk 2010. aasta esimeses kvartalis. Lisa 8 põhjal on esimene diagnoos raseduse, sünnituse ja sünnitusjärgse perioodi grupist ning teine kontakt terviseteenistusega. Samuti registreeriti järgmisel kahel kvartalil terviseteenuse kontaktid. Aasta 2013 ning 2016. aasta esimeses kvartalis diagnoositi patsiendile haigus kuse-suguelundite grupist. Samuti diagnoositi lihasluukonna ja sidekoe haigus aasta 2015. aasta viimases ning 2016. aasta esimeses kvartalis.

Tabel 4: Patsiendi 10031105 10 erinevat potentsiaalset kliinilist trajektoori.

Prognoos / **Kvartal**	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
30. kv	-	-	-	-	-	M6	-	M7	Z0	-
31. kv	C5	J3	H3, H5	N9, Z0	M6, M7, Z0	B0, B3,C5, N9	-	K5	Z9	H3, Z0
32. kv	C5, R5, Z4	-	K2	B3, J4	C5	C5, D1, H1, J0, M2, N9, Z0, Z5	M1, N6	N6	C5	D4, I1, N6
33. kv	A4, G6, H6, M2, N9, Z0	C4	C4	H8, H9, N3, Z0	C5, F4, M6	C5	N6	C7, Z9	C5, H1, K8, N3, N7	C1, D4, E6, M6

Tabeli 4 põhjal on kümme erinevat potentsiaalset kliinilist trajektoori vaadeldavale patsiendile märgatavalt erinevad. Siiski pole seitsmel korral kümnest mudel patsiendile 30. kvartaliks ehk 2016. aasta 3. kvartaliks midagi diagnoosinud. Kahel korral on viimasesse kvartalisse prognoositud lihasluukonna ja sidekoe haigus ning ühel korral kontakt terviseteenusega, vt tabel 4 ja lisa 8.

Võib märgata, et tabelis 4 on kaheksal prognoosil kümnest ennustatud mingil hetkel kasvaja diagnoos. Viimane on tingitud ilmselt antud andmestikust, sest töös kasutatakse rinnavähi kohordi andmeid. Antud patsiendil pole tabeli 3 põhjal varasemalt vähki diagnoositud. Seega seostas mudel tema teisi diagnoose teiste vähidiagnoosiga patsientide kliiniliste trajektooridega.

Analoogiliselt eelnevaga ennustati mitmel korral silma- ja silmamanuste grupi ning kõrva- ja nibujätkehaigusi, kuigi patsiendil vastavad varasemad diagnoosid puuduvad. Samuti ennustati kolmel korral erinevaid seedeelundite haigusi, vt tabel 4. Tegu võib olla haigustevaheliste seostega, mida võiks täpsemalt edasi uurida.

Lisaks prognoositi patsiendile mitmel korral lihasluukonna ja sidekoe haiguseid, kuid vastava haigusgrupi diagnoos on patsiendil ka eelnevalt kahel korral määratud, vt tabel 4. Sellise korduva andmemustri põhjal võib eeldada, et tegu on kroonilise probleemiga.

3.2.2 Näide paljude diagnoosidega patsiendi kohta

Patsiendi 10344337 vanus on vaadeldaval ajaperioodil andmestikus 77 kuni 84 aastat. Vastavas ajavahemikus on tabeli 5 järgi märgitud talle 143 diagnoosi.

Patsient valiti näiteks oma diagnooside rohkuse tõttu. Edaspidi uuritakse, kuidas käitub prognoosimisel mudel, kui patsiendi kliiniline trajektoor on mitmekesine.

Tabeli 5 põhjal on patsiendil igas andmestikust teadaolevas kvartalis märgitud vähemalt kaks diagnoosi, enamikus üle nelja diagnoosi. Enim on märgitud

Tabel 5: Patsiendi 10344337 diagnooside ajalugu.

Kvartal	Diagnoos	Kvartal	Diagnoos	Kvartal	Diagnoos
4. kv	C5, H4, H5,I1, J3, M1, M8	14. kv	D1, E0, H2, H4, I1, K5	24. kv	D1, E1, H4, I1, M1
5. kv	C5, H4, H5, I1, K4, M1, Z0	15. kv	D1, D3, E1, H2	25. kv	C5, D1, E0, E1, H4
6. kv	C5, H0, Z0	16. kv	D3, E1, I1, K5, Q1, Z0	26. kv	D3, E7, H4, I1
7. kv	C5, H5, I1, L0, L8, M1, Z4	17. kv	C5, D1, D3, E1, H3, I1, L0, M1, S9	27. kv	D3, E1, H4, I1, J3, M1
8. kv	D1, H4, I1, J3, N9	18. kv	D1, D3, E1, H4, K5	28. kv	E1, K8
9. kv	C5, D1, H4, I1, M1, M5, Q1, Q4	19. kv	D3, E1, I1, L8	29. kv	E1, H4, M1
10. kv	D1, E0, H0, H4, M1, Q1	20. kv	D1, E1, H3, H5, I1, Z0		
11. kv	E0, E1, H4, M1	21. kv	C5, D1, E1		
12. kv	C5, D1, E1, E7, H5, I1, L3, M1, Q1	22. kv	D3, E1, H4, M1, Z0		
13. kv	C5, D1, E0, E1, H0, H4, H5, I1, M1, Q4	23. kv	E1, H4, I1, K5		

diagnoose algusega C ja D, mis tähistavad kasvajaid (lisa 8), viimaseid on kokku lisatud 31. Järgnevalt on 29 korda patsiendile registreeritud erinevaid silma- ja silmamanuste haigusi. Samuti on patsiendil diagnoositud vaadeldaval ajaperioodil 24 erinevat sisesekretsiooni-, toitumis- ja ainevahetushaiguse diagnoosi, 16 ja 15 vastavalt vereringeelundite ning lihasluukonna ja sidekoe haigust. Lisaks on erinevaid haigusi veel vähemalt 7 erinevast grupist.

Tabel 6: Patsiendi 10344337 10 erinevat potentsiaalset kliinilist trajektoori.

Kvartal \ Prognoos	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
30. kv	C7, E1, H4	D5, E1	C5, E1, H4, Z0	H3	C5, E1	C5, E1, H3, H4, M1, S5	E1, Z0	E1, Z9	J0	H4, M3, N8
31. kv	H2, H4	E8, H1, H4	E1, H4, I1	C5, E1	-	H4	C5, E1	H4	E1, H4, M1	H4, Z0
32. kv	E0, E1, H4, I4, J3, K5	E1, H4	E1, H4, N8	E1, H4, H9	C5, E1, H4, M1	E1, H4, L5, Z0	C5, E1, H3	E1, H4, I1, Z0	E1, H4	E1
33. kv	H4	E1, H6	C5, E1, H4, H5	C5, E1, H3, H4	C5, E1, M1, N8	E1, H4, L5, M1, M8	E1, H4	C5, E1, H3, H4, I1	-	E1

Tabelist 6 võib esmalt välja tuua, et igal prognoosimisel kümnest on mingil ajahetkel välja tulnud sisesekretsiooni-, toitumis- ja ainevahetushaigus koodiga $E1$, paaril korral ka sama grupi teine variatsioon. Lisaks ilmneb kõigil ennustusel silma- ja silmamanuste haigus. Järelikult on suure tõenäosusega tegu krooniliste haigustega.

Enamikel prognoosimistel ennustatakse kasvaja tekkimist, kus põhiline diagnoos on C5, vt tabel 6. Sotsiaalministeeriumi kodulehe [8] põhjal tähistab viimane rinna pahaloomulist kasvajat. Saadud tulemust selgitab taas andmestiku päritolu ning patsiendi eelnev kliiniline trajektoor.

Patsiendil on eelnevalt korduvalt diagnoositud lihasluukonna ja sidekoe haigusi, vt tabel 5. Samas tabeli 6 põhjal on vastavat diagnoosi prognoositud vähem, alla pooltel ennustustel. Lisaks on vastavast grupist prognoositud kolme erinevat variatsiooni M1, M3, M8, kusjuures diagnoosi M3 pole tal varem märgitud. Vastupidiselt on patsiendile eelnevalt hoopis diagnoositud M5. Sotsiaalministeeriumi kodulehe [8] järgi kuulub M1 artrooside gruppi, tulemuste järgi võib see olla krooniline. Lisaks võib M5 ehk mingi seljahaiguse varasem diagnoos põhjustada ennustatud luuhaigust (M8).

3.3 Haiguste tõenäosuste varieeruvus

Lisaks uuriti ka erinevate haiguste prognoositud tõenäosuste varieeruvust. Kui haiguse ennustatud tekkimistõenäosused on andmestiku põhjal väga erinevad, on alust arvata, et mudeli põhieesmärk on täidetud. See tähendab, et mudel on võimeline erineva taustaga inimestele personaalseid ennustusi väljastama.

Tabel 7: Diagnooside tõenäosuste varieeruvus.

RHK-10	**Diagnoosigrupp**	**tõenäosuse dispersioon**
C5	kasvajad	0,069
H4	silma- ja silmamanuste haigused	0,016
Z5	terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega	0,014
F3	psüühika- ja käitumishäired	0,05
F2	psüühika- ja käitumishäired	0,050
E1	sisesekretsiooni-, toitumis- ja ainevahetushaigused	0,0037
H3	silma- ja silmamanuste haiguse	0,0035
F0	psüühika- ja käitumishäired	0,0031
...	...	...
B6	teatavad nakkus- ja parasiithaigused	$4{,}49 \times 10^{-11}$
N5	kuse-suguelundite haigused	$4{,}29 \times 10^{-11}$
Q0	kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad	$3{,}99 \times 10^{-11}$
B4	teatavad nakkus- ja parasiithaigused	$3{,}88 \times 10^{-11}$
J6	hingamiselundite haigused	$3{,}46 \times 10^{-11}$
A8	teatavad nakkus- ja parasiithaigused	$3{,}33 \times 10^{-11}$
Q1	kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad	$3{,}16 \times 10^{-11}$
J7	hingamiselundite haigused	$2{,}78 \times 10^{-11}$

Tabelist 7 selgub, et kõige enam varieeruvad diagnoosid järgmistest gruppidest: kasvajad, silma- ja silmamanuste haigused, terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega, psüühika- ja käitumishäired, sisesekretsiooni-, toitumis- ja ainevahetushaigused.

Kõige vähem varieeruvad tabeli 7 ja lisa 8 järgi haigused gruppidest nakkus- ja parasiithaigused, kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad, kuse-suguelundite haigused ja hingamiselundite

haigused. Nii tabeli 7 all- kui ka ülaosas korduvad sama diagnoosigrupi erinevad variatsioonid.

Ilmneb tendents, et mudel on ennustanud sarnaseid ja vähem varieeruvaid tõenäosuseid haigustele, mida autori arvates ongi kliinilise trajektoori põhjal keeruline prognoosida. Näiteks nakkushaiguseid, mis tekivad nakkustekitaja sattumisest organismi ning kaasasündinud väärarendid võivad olla pärilikud.

Lisaks varieeruvad rohkem patsienditi diagnoosid, mis on kroonilised ja vajavad tihemini terviseteenistuse ülevaadet. Enim varieerub kasvajagrupi diagnoos, kuna andmetes on tegemist rinnavähi kohordiga. Viimast õppis mudel paremini diagnoosima, kuna selle kohta oli vastava valimi põhjal enim informatsiooni.

Kokkuvõte

Antud bakalaureusetöös loodi Haigekassa andmetel tõenäosuslik mudel. Tulemuste põhjal võib väita, et loodud mudel ennustas antud patsientidele neile personaalseid potentsiaalseid kliinilisi trajektoore. Vaadeldavate diagnooside prognoositud tõenäosused erinesid patsienditi märgatavalt.

Lisaks ennustas mudel tulemuste põhjal vaadeldavatele patsientidele enim haigusi sellistest gruppidest, milliseid oli neil varem diagnoositud. Ilmnesid kliinilised trajektoorid, mis viitasid kroonilistele haigustele. Samas ilmnesid ka prognoosid haigustele, milliseid patsientidele varem diagnoositud polnud. Tegu võis olla seostega, mida tuleks edasi uurida.

Kui patsiendi eelnev kliiniline trajektoor on olnud mitmekesine, on ka mitmekordselt teostatud prognoosid sarnasemad. Viimast põhjendab ka autori intuitiivne eeldus, kui inimesel pole palju tervisekomplikatsioone registreeritud, siis on raske talle konkreetset haigust ennustada.

Kasutatud kirjandus

[1] Chollet, F., Allaire, J. J. (2018). *Deep Learning with R.* New York: Manning Publications Co.

[2] Salehinejad, H., Sharan, S., Barfett, J., Colak, E. ja Valaee, S. (2018). *Recent Advances in Recurrent Neural Networks.* New York: Cornell University.

[3] Petlenkov, E. (2004). *Tehisnärvivõrgud ja nende rakendused.* Õppematerjal. Kasutatud 02.05.2020, https://www.ttu.ee/public/i/infotehnoloogia-teaduskond/Instituudid/automaatikainstituut/oppeained/susteemiteooria/Tehisnarvivorgud-EP_2004.pdf

[4] Ruder, S. (2017). *An overview of gradient descent optimization algorithms.* New York: Cornell University.

[5] Phi, M. (2018). Illustrated Guide to LSTM's and GRU's: A step by step explanation. *Medium: towards data science.* Kasutatud 04.05.2020, https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[6] Isikuandmete kaitse seadus. (04.01.2019). *Riigi Teataja I.* Kasutatud 03.05.2020, https://www.riigiteataja.ee/akt/106012016006

[7] Wilson Van Voorhis, C. R., Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology, 3, 48.* http://doi.org/10.20982/tqmp.03.2.p043

[8] Sotsiaalministeeriumi koduleheküg. (i.a). Kasutatud 09.05.2020, https://rhk.sm.ee/

Lisad

Lisa 1. RHK-10 koodid ja nende tähendused

Tabel 8: Rahvusvaheliste haiguste klassifikatsioon [8]

RHK-10	**Diagnoosi grupp**
A00-B99	teatavad nakkus- ja parasiithaigused
C00-D49	kasvajad
D50-D89	vere ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid
E00-E89	sisesekretsiooni-, toitumis- ja ainevahetushaigused
F01-F99	psüühika- ja käitumishäired
G00-G99	närvisüsteemihaigused
H00-H59	silma- ja silmamanuste haigused
H60-H95	kõrva- ja nibujätkehaigused
I00-I99	vereringeelundite haigused
J00-J99	hingamiselundite haigused
K00-K95	seedelundite haigused
L00-L99	naha- ja nahaaluskoe haigused
M00-M99	lihasluukonna ja sidekoe haigused
N00-N99	kuse-suguelundite haigused
O00-O9A	rasedus, sünnitus ja sünnitusjärgne periood haigused
P00-P96	perinataal- e sünniperioodis tekkivad teatavad seisundid
Q00-Q99	kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
R00-R99	mujal klassifitseerimata sümptomid, tunnused ja kliiniliste ning laboratoorsete leidude hälbed
S00-T88	vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
V00-Y99	välised surmapõhjustajad
Z00-Z99	terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Brigitta Rebane,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Rekurrentsed tehisnärvivõrgud inimeste kliiniliste trajektooride ennustamisel“, mille juhendaja on Raivo Kolde, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Brigitta Rebane
18.05.2020