

# ComPatAI-konsortio hyödyntää suuria datamääriä oppivan tekoälymallin luomiseksi patologiaan

Turun yliopiston biolääketieteen laitoksen apulaisprofessori Pekka Ruusuvuori johtaa ComPatAI-konsortiota, jossa kehitetään histopatologisten kudoksenäytteiden mallinnusta uutta sisältöä luovien ja ennustavien tekoälymenetelmien avulla.



Histologisen eli kudospillisen näytteen perusteella arvioidaan, tarvitseeko potilas hoitoa. Tavoitteena on kehittää erittäin suuriin data-aineistoihin perustuvia tekoälymalleja, joiden avulla saadaan entistä tarkempaa patologian diagnostiikkaa.

CompPatAI-konsortiossa on lisäksi kehitteillä generatiiviseen tekoälyyn perustuvia kudoksen virtuaalivärjysmalleja. Konsortion muut osapuolet ovat tutkimusjohtaja, dosentti **Leena Latonen** Itä-Suomen yliopistosta sekä patologian osaston ylilääkäri, dosentti **Teemu Tolonen** Fimlab-laboratorioista.

ComPatAI-konsortioissa analysoidaan ensisijaisesti rintasyöpään ja eturauhassyöpään liittyviä kudosomeikkeitä. Digitoitu kuva antaa mahdollisuuden mittauksiin ja erilaisen solutyypin automaattiseen laskentaan.

”Olemme toimineet eturauhassyövän ja rintasyövän parissa. Näistä on ollut dataa tarjolla, koska ne ovat kaikkein yleisimmät syöpätyypit naisilla ja miehillä. Tavoitteena on kuitenkin, että meillä olisi hyvin yleiskäyttöinen malli, jonka päälle voitaisiin rakentaa ratkaisuja erilaisiin ja uusiin käyttökohteisiin.”

Ruusuvuoren mukaan digitalisaatio on tapahtumassa patologiaan nyt ja Suomi on tiettyssä mielessä edelläkävijä.

”Tampereella ja Turussa on siirrytty kokonaan digitaaliseen patologiaan diagnostiikassa. Joka kerta kun näyte otetaan, se skannataan korkearesoluutiiseksi digitaalliseksi. Rutiinidiagnostiikkaa tehdään paljon. Koska väestö ikääntyy, syöpätapaukset ovat nousussa. Dataa saadaan koko ajan kovalla tahdilla.”

## 600 000 kokoleikekuvaa

Skannatut kokolasikuvat saadaan tutkimukseen Fimlabista, joka on Suomen suurin terveydenhuollon laboratorioyhtiö. Sen asiakkaita ovat sairaalat, terveyskeskukset, työterveyshuolto ja yksityiset lääkäriasemat. Lääkealan turvallisuus- ja kehittämiskeskus Fimean lupa käsittää tällä hetkellä 160 050 tapausta eli noin 600 000 kokoleikekuvaa. Koko on yhteensä noin 0,8 petatavua, jolloin yhden tiedoston koko on noin 1,3 GB. Massiivista datamäärää siirretään parhaillaan anonymisoinnin jälkeen Suomen ELIXIR-keskuksen CSC:n LUMI- supertietokoneelle. Se on suurimpia koneelle tehtyjä datan siirtoja.

”Se, että saamme hyödyntää näitä aineistoja tutkimuskäytössä, on valtavan hieno juttu. Tarkoitus on käyttää tätä isoa datamassaa siihen, että pystyttäisiin tekemään mahdolli-

simman hyvin toimivia tekoälyratkaisuja patologioiden käyttöön”, sanoo Ruusuvuori.

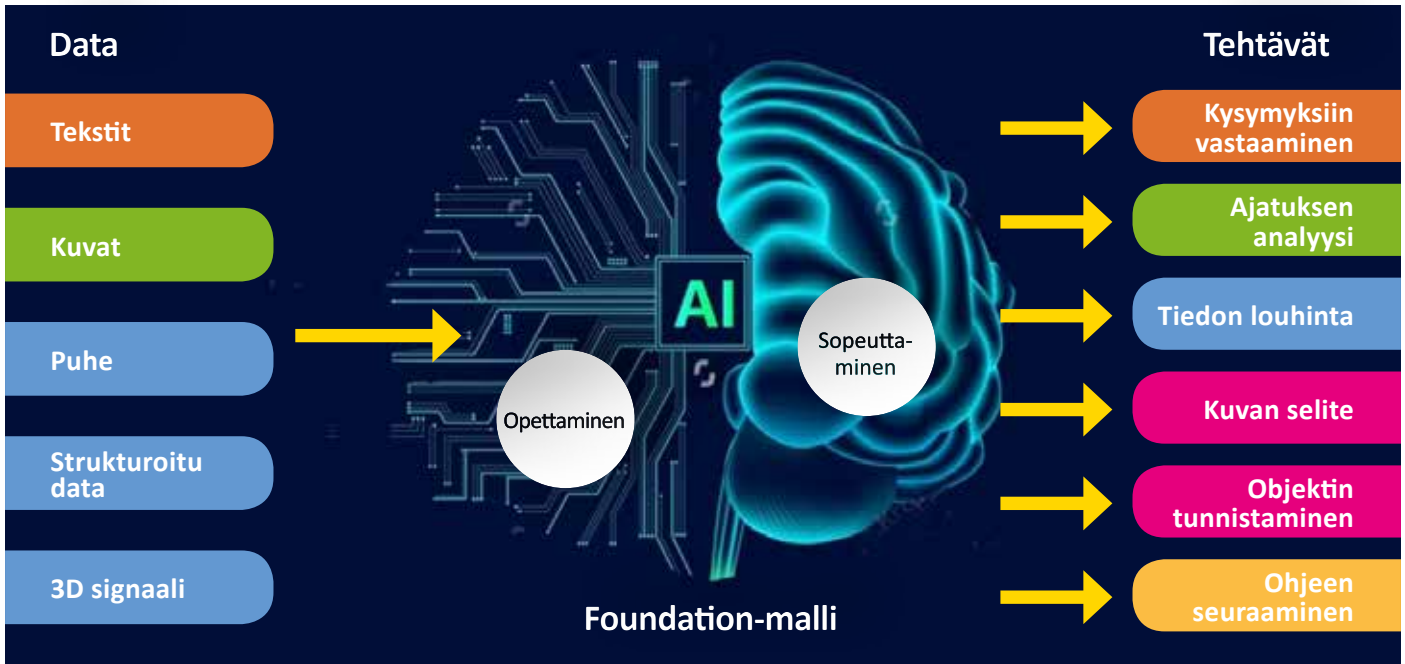
Tavoitteena on, että projektin lopussa tutkijoiden käytössä olisi jopa 2,5 miljoonaa digitoitua kokoleikekuvaa, jolloin dataa olisi kolme petatavua.

”Meillä on lupateknisesti mahdollisuus käyttää kaikkea sitä dataa mitä rutiinisti FIMLABilla tuotetaan digipatologiassa.”

## Neuroverkko oppii kuvasta sellaista, mitä ihmissilmä ei havaitse

Pekka Ruusuvuoren tausta on signaalinkäsittelyssä ja hänen erityisosaamisalueensa on kuva-analyysi. Hän on kiinnostunut siitä, miten tekoälymenetelmissä hyödynnetyistä syvästä neuroverkoista voitaisiin kehittää kohti paremmin erilaisiin käyttötarkoituksiin yleistäviä.

Ruusuvuoren mukaan lähtökohtaisesti kone voidaan opettaa tunnistamaan samoja asioita kuin ihminen. Se voidaan opettaa tunnistamaan erilaisia kudostyyppisiä ja erottamaan syöpäkudosta terveestä kudoksesta. Se voi mitata solusta tai kuvista erilaisia asioita, kuten kuinka aggressiivinen syöpä on ja kuinka pitkälle se on edistynyt. Tekoäly voi



*Foundation-mallin yleisesitys. Foundation-mallien toiminta perustuu syviin neuroverkkoihin ja niiden kykyyn oppia monimutkaisia kuvioita ja rakenteita datasta. Foundation-malleja hyödynnetään yhä enemmän kuvadatan analysoinnissa. Mallit oppivat yhdistämään visuaalisia piirteitä (kuten värit, muodot ja tekstuurit) ja semanttista tietoa (kuvien merkitystä tai tarkoitusta). Mallit hajottavat kuvan pikselitason tietoihin ja oppivat monimutkaisia piirteitä. Ne käyttävät matemaattista tekniikkaa nimeltä itsehuomio (self-attention) ymmärtääkseen, mitkä kuvan osat ovat tärkeitä ja millä tavalla kuvan elementit liittyvät toisiinsa.*

tehdä erottelua ja löytää kudosnäytteestä syöpäalueet ennen kuin patologia alkaa tutkia näytettä. Se voi myös ehdottaa luokitusta. Esimerkiksi eturauhasen syöpäkasvaimesta annetaan ns. Gleason-luokitus, joka kertoo miten aggressiivinen tai edennyt tauti on.

”Tekoälylle on opetettavissa melko tarkasti siis sellaiset tehtävät mitä patologit tekevät”, Ruusuvuori toteaa.

”Perinteisesti koneoppimismenetelmät on rakennettu niin, että meillä on joku kohde- muuttuja ja opetusaineisto, jossa näytetään, että tässä kohtaa tätä kuvaa on tämä objekti ja se kuvaa tätä luokkaa. Sehän on hirveän työlästä, jos meidän pitäisi merkitä kaikkiin satoihin tuhansiin kuviin tätä tietoa.”

Nämä ns. annotaatiotiedot ovat olleet olennaisia, jotta on voitu opettaa tekoälyä automaattisesti tunnistamaan näytteistä esimerkiksi syöpäsolut. Ruusuvuoren mukaan algoritmit ovat kuitenkin kehittyneet siihen suuntaan, että ne pystyvät hyödyntämään raakadataa ilman annotointeja.

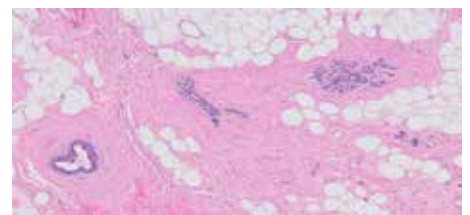
”Mielestäni kaikkein kiinnostavinta onkin se, mitä kaikkea muuta kuvista on irrotettavissa eli ominaisuuksia, mitkä eivät välttämättä

ole itsestään selvästi ihmisen havaittavissa. Ainoa data mitä on nähtävillä, on leikekuvassa. Jos siinä on joku tilastollinen yhteys osoitettavissa, koneoppimisalgoritmi sen löytää – mutta ne yhteydet saattavat olla hyvin kompleksisia. Nykyaikaiset neuroverkot ovat erittäin tarkkoja havaitsemaan kompleksisia yhteyksiä spatiaalisen datan ja ennustettavan muuttujan välillä. Ne voivat olla hyvin vaikeita hahmottaa meille ihmisille.”

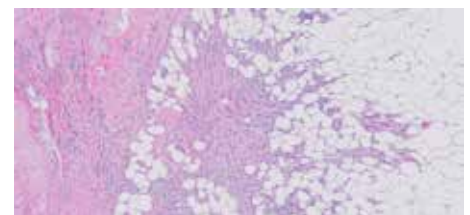
Ruusuvuori on tutkimusryhmänsä kanssa pystynyt koneoppimismallien avulla ennustamaan geeniekspressiota ja mutaatioita suoraan histologisista kuvista. Geenin ekspresio eli ilmentyminen tarkoittaa, että solu tuottaa DNA:n koodaamaa molekyyliä. Geenien ekspresio on erilainen eri kudoksissa. Tekoäly voi havaita kuvasta ihmissilmälle näkymättömiä pieniä muutoksia.

”Kuvissa koneelle on siis näkyvissä jotain, mitä geeniekspresio aiheuttaa soluissa ja kudoksissa. Kone pystyy havaitsemaan erittäin pienenkin eron muuttuneessa ilmiössä. Kone havaitsee sen, mitä ihmissilmä ei ole harjaantunut näkemään. Korostan, että tämä on hyvin suuntaa antavaa ja ei toki toi-

mi kaikille kudoksille tai geeneille. Kaikkien geenien ekspressoituminen ei johda muutokseen kudostasolla sillä tavoin, että se on ennustettavissa kudostenleikekuvasta.”



*Terve kudos: Sidekudoksen joukossa nähdään säännöllisiä maitotiehyitä ja lobuluksia (rauhasliuskvoja). Kuva: Fimlab*



*Rintasyövän kudos: Kasvain muodostaa juosteisia rakenteita ja pieniä saarekkeita, jotka kasvavat epäsäännöllisesti stroomaan (kasvaimen sidekudosverkko) ja rasvaan. Kuva: Fimlab*



ComPatAI-konsortio kehittää suurten datamassojen hyödyntämiseen ns. foundation-mallia. Foundation-malli luo yleiskäyttöisen perustan erilaisille tekoälyratkaisuille oppien histologiaa suuresta näytemäärästä ilman kohdemuuttujia tai annotointeja.

”Kun tälle mallille aletaan opettaa vaikka rintasyövän tai eturauhassyövän tunnistusta, malli alkaa oppimaan pyydettyä tehtävää. Näin pääsemme paljon nopeammin tarkempaan ratkaisuihin. Pystymme hyödyntämään mittavaa data-aineistoa, vaikka meillä ei olisi annotointeja. Se on hieno esitysaskel.”

ComPatAI-konsortio luo omaa foundation-tekoälymallia suomalaisen dataan perustuen.

”Tämä on perustutkimusta, joka mahdollistaa sen, että olemme ensimmäisten joukossa kehittämässä tähän maahan näitä malleja. Toivon, että emme olisi pelkästään isojen ulkomaisten firmojen ja tutkimusryhmien

varassa vaan että meillä rakennettaisiin suomalaisen dataan perustuvaa mallia. Meillä on tässä maassa laadukasta populaatiotason kohorttidataa, jota pitää päästä hyödyntämään. Toivon, että se johtaa siihen, että saadaan Suomeen yrityksiä, joiden kehittämät ratkaisut viedään potilaan hyödyksi rutiinidiagnostiikkaan.”

Tärkeä kysymys on, kuinka nopeasti dataa pystytään siirtämään ja hyödyntämään. Laskentaa ja datan tallennuskapasiteettia tarvitaan koko ajan. Tähän tulevat apuun Suomen ELIXIR-keskuksen CSC:n tarjoamat palvelut.

”Olemme erittäin tyytyväisiä CSC:ltä saamaamme tukeen, kun puhutaan näin poikkeuksellisen isosta hankkeesta ja datamäärästä. Olemme etuoikeutetussa asemassa, koska meillä on apuna CSC:n tapainen toimija, jolta voimme saada resursseja tällaiseen tutkimukseen. Se on selvästi kilpailuetu ja sellainen asia, mistä voi olla valtavan kiitollinen.”

*Pekka Ruusuvooren tutkimus (Towards AI-enabled computational pathology) on Suomen Akatemian rahoittama ja kuuluu LUMI Extreme scale access-projekteihin, jossa pilotoidaan suurteholaskentaa julkisilla datoilla. Ruusuvoorella ja Leena Latosella on lisäksi Suomen Akatemian rahoittama suurteholaskentaan keskittyvä hanke kudosten virtuaalivärjäykseen liittyen. Suomen Akatemian rahoituksella vahvistetaan eurooppalaisen EuroHPC (European High-Performance Computing) -suurteholaskennan resurssien ja LUMI-supertietokoneen hyödyntämistä lippulaivojen aihealueiden tieteelliseen tutkimukseen.*

Digipatologian ja muiden potentiaalistesti sensitiivisten terveysdatan datatyyppien kuten rekisteri- ja omiikkatietovarantojen saatavuus tietoturvalisessä CSC:n käyttöympäristössä kasvaa tulevaisuudessa.

”Kehitys on vasta alussa”, sanoo **Tommi Nyrönen**, joka on Suomen ELIXIR-toimintojen johtaja.

”Suomen ELIXIR on edistänyt CompPatAI-tutkimuksen edellyttämien biolääketieteellisten resurssien muuttamista CSC:n alustapalveluksi. Työn tuloksena syntynyt CSC Sensitive Data-alusta tukee muitakin vastaavia hankkeita. Tällainen on esimerkiksi EU:n digipatologian arkiston rakennushanke bigpicture.eu, joka suunnitelman mukaan alkaa vuonna 2026 tarjota kestävästä ratkaisusta hallita ja tuoda digipatologian data-aineistoja suurteholaskentapalveluihin Euroopan laajuisesti.”

14.12.2024 | **Ari Turunen**

#### LISÄTIETOJA:

##### Ruusuvuorilab

<https://ruusuvuorilab.utu.fi>

##### Fimlab

<https://fimlab.fi>

##### Turun yliopisto

<https://www.utu.fi/fi>

##### CSC – Tieteen tietotekniikan keskus Oy

on valtion omistama, opetus- ja kulttuuriministeriön hallinnoima, voittoa tavoittelematon osakeyhtiö. CSC ylläpitää ja kehittää valtion omistamaa keskitettyä tietotekniikkainfrastruktuuria.

<http://www.csc.fi>

<https://research.csc.fi/cloud-computing>

##### ELIXIR

rakentaa infrastruktuurin bioalan tutkimuksen tueksi. Se yhdistää 21 Euroopan maan ja Euroopan molekyylibiologian laboratorion EMBL:n johtavat organisaatiot yhteiseksi biologisen informaation infrastruktuuriksi. Sen Suomen keskus on CSC Tieteen tietotekniikan keskus Oy.

<http://www.elixir-finland.org>

<http://www.elixir-europe.org>

#### SUOMEN ELIXIR

Puh. +358 9 457 2821 e-mail: [servicedesk@csc.fi](mailto:servicedesk@csc.fi)  
[www.elixir-europe.org/about-us/who-we-are/nodes/finland](http://www.elixir-europe.org/about-us/who-we-are/nodes/finland)

[www.elixir-finland.org](http://www.elixir-finland.org)

#### ELIXIR PÄÄMAJA

EMBL-European Bioinformatics Institute  
[www.elixir-europe.org](http://www.elixir-europe.org)