TURUN YLIOPISTO

Koneoppimismenetelmät tähtitieteen survey-tutkimuksissa

Tietojenkäsittelytiede
Tietotekniikan laitos, Teknillinen tiedekunta
Kanditutkielma

Laatija:
Evelina Suominen

Maaliskuu 2026

Kandidaatintutkielma
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Tutkinto-ohjelma: Tietojenkäsittelytiede
Tekijä: Evelina Suominen
Otsikko: Koneoppimismenetelmät tähtitieteen survey-tutkimuksissa
Sivumäärä: 26 sivua, 1 liitesivu
Päivämäärä: Maaliskuu 2026

Tähtitieteen toinen toistaan suurempiskaalaiset tutkimukset ovat johtaneet käsiteltävänä olevien datajoukkojen räjähdysmäiseen kasvuun. Nämä survey-tutkimukset tuottavat jatkuvaa datavirtaa useista kiinnostavista taivaan tutkimuskohteista, eikä kaikkea dataa voida tutkijoiden toimesta käsin käydä läpi. Tarpeen on soveltaa uusia, tehokkaampia ja automatisoidumpia menetelmiä datan analyysiin. Tutkielmassa tarkastellaan kirjallisuuskatsauksena tähtitieteen datajoukkojen erityispiirteitä ja -haasteita sekä erilaisia koneoppimiseen perustuvia data-analyysimenetelmiä, joita voidaan tähtitieteen dataan soveltaa. Johtopäätöksiä tehdään menetelmien keskinäisestä suoriutumisesta sekä yleisestä soveltuvuudesta tähtitieteen dataan.

Avainsanat: tähtitiede, koneoppimismenetelmät, astroinformatiikka

Sisällysluettelo

1 Johdanto

Modernit tähtitieteen tutkimukset ovat johtaneet tutkimusaineistojen koon räjähtävään kasvuun viime vuosikymmeninä. Tätä myötä onkin alettu puhumaan ”big datan aikakaudesta”, jossa osittain tai kokonaan automatisoiduilla havaintolaitteistoilla tuotetaan valtavia datamääriä, joita tutkijoiden toimesta ei mitenkään voida kaikkia käsitellä perinteisillä analyysimenetelmillä. [1] Lähiaikoina aloittava Legacy Survey of Space and Time (LSST) -tutkimuksen on projektoitu tuottavan jopa 200 petatavun kokoisen tutkimusaineiston [2]. Lisäksi tähtitieteessä voidaan hyödyntää valtavaa määrää tietokantojen tai aikaisempien tutkimuksien datajulkaisujen sisältämää tietoa, jota voidaan hyödyntää uudessa tutkimuksessa. Uusia, automaattisempia ja tehokkaampia analyysimenetelmiä on sovellettava, jotta näistä supermassiivisista tutkimusjoukoista löydetään hyödyllisiä tutkimuskohteita ja johtopäätöksiä. [3]

Luonteva valinta tehokkaaseen datankäsittelyyn ovat erilaiset koneoppimismenetelmät. Koneoppimiseksi kutsutaan algoritmisia malleja, jotka itsenäisesti oppivat koulutusaineiston ominaisuuksia, soveltaen näitä uusien data-alkioiden käsittelyyn. Koneoppimismenetelmiä on useita, ja usein erikoistuvat tietynlaisen tehtävän suorittamiseen. Klassiset koneoppimistehtävät ovat yleensä luokittelu- tai klusterointitehtäviä, mutta esimerkiksi regressio-ongelmiin voidaan myös soveltaa koneoppimista. Tähtitieteen alalla on runsaasti tutkimusta koneoppimismenetelmien soveltuvuudesta, ja tällaisia algoritmeja ollaan enenevässä määrin ottamassa käyttöön uuden datan analyysitehtävissä. [1]

Tässä tutkielmassa käsitellään tähtitieteellisen datan erityisominaisuuksia ja siihen liittyviä haasteita. Tämän jälkeen tarkastellaan koneoppimismenetelmiä ja niiden ominaisuuksia sekä soveltuvuutta tähtitieteellisen datan analyysiin. Kirjallisuudesta tarkastellaan esimerkkitutkimuksia, jotka tutkivat tiettyä tai vertailevat useiden koneoppimismenetelmien tehokkuutta tähtitieteen tutkimustehtävissä. Keskittymiskohteena on erityisesti survey-tutkimusten datajoukkojen luokittelu, kategorisointi ja poikkeamien havaitseminen. Tutkielmassa vastataan seuraaviin tutkimuskysymyksiin:

Tk1: Millaisia erityispiirteitä tähtitieteellisissä datajoukoissa on data-analytiikan kannalta?

Tk2: Millaisia koneoppimismenetelmiä näihin datajoukkoihin voidaan soveltaa, ja millaisia tuloksia niillä saadaan?

Tutkielma on kirjallisuuskatsaus. Toteutusmenetelmänä on aiheeseen liittyvien aineistojen haku tieteellisistä tietokannoista ja näiden julkaisujen analyysi.

Tiedonhaku toteutettiin keskinäisesti soveltaen hakulausetta *astronom* AND "machine learning" AND ("survey" OR "big data")*. Haku suodatettiin kohdistuen tutkimus- ja review-artikkeleihin, sekä julkaisuajan perusteella vuodesta 2010 alkaen. Tietokantoina käytettiin IEEE Xplore -tietokantaa, josta alustavia tuloksia 146, otsikon ja tiivistelmän perusteella poimittiin 8, Web Of Science -tietokantaa, josta alustavia tuloksia 360, otsikon ja tiivistelmän perusteella poimittu 8, ja utu Volter-kirjastopalvelua, josta käsiteltäväksi poimittiin 16 artikkelia. Aineistohaussa alustavasti poimittuja lähteitä yhteensä 36, joista sisällön perusteella käsitellään tutkielmassa 20:ä. Käsiteltävänä olevien tutkimusartikkeleiden lisäksi viitataan näissä mainituista survey-tutkimuksista, tähtitieteen tietokannoista sekä monimutkaisista koneoppimismalleista alkuperäisiin lähteisiin lisätietoa ja taustoitusta varten.

Tutkielma jakautuu kolmeen käsittelylukuun. Luvussa 2 käsitellään tähtitieteen survey-datan ominaisuuksia ja datan käsittelyyn liittyviä haasteita. Lisäksi taustoitetaan tutkimuksia, joilla tätä dataa tuotetaan. Luvussa 3 esitellään yleisiä tähtitieteen dataan sovellettavia koneoppimismalleja, niiden ominaisuuksia ja soveltuvuutta tähtitieteen dataan. Luvussa 4 käsitellään yksittäisiä poimintoja kirjallisuudesta koneoppimismenetelmien sovelluksista, keskittyen erityisesti koneoppimismenetelmillä saatuihin tuloksiin. Luku 5 on yhteenveto.

2 Tähtitieteen data ja sen ominaisuuksia

Tähtitieteen alan suurempiskaalaisten tutkimusten myötä puhutaan nykyään monesti tähtitieteen ”big datan aikakaudesta”. Tällaisia suuria datamääriä tuottavat tutkimukset ovat pääasiassa laaja-alaisia survey-tutkimuksia, joiden tarkoituksena on kartoittaa useita kohteita tiettyjen havaintokohteiden ominaisuuksien tutkimisen sijaan. [2]

2.1 Survey-tutkimukset

Käsitteellä ”survey-tutkimus” tarkoitetaan tässä tutkielmassa sellaisia havaitsevan tähtitieteen tutkimuksia, jossa tarkoitus on kartoittaa koko taivasta tai tiettyä laajaa osa-aluetta siitä. Survey-tutkimuksia tehdään pääasiassa eri sähkömagneettisen säteilyn aallonpituusalueilla; usein joko näkyvän valon, infrapunasäteilyn tai radioaaltojen taajuuksilla. Tutkimusten tiedetarkoituksissa on jonkin verran vaihtelua – tarkoituksena voi olla yleinen kohteiden tilastointi, tietyntyyppisten kohteiden etsiminen, tai uusien havaintolaitteistojen tai -menetelmien testaaminen [3, p. 226]. Survey-tutkimusten ero ”perinteisiin” yksittäisiin kohteisiin kohdistuviin tähtitieteen tutkimuksiin on siis laaja-alaisuus ja tarkan havaintokohdelistan puute. Seuraavaksi esitellään muutama poiminta huomattavista survey-tutkimuksista (kts. Liite 1 laajempaa listausta varten).

Palomar Digital Sky Survey (lyh. DPOSS, vuosi 1997) voidaan ajatella olevan ensimmäisiä survey-tutkimuksia – se on digitoitu versio valokuvauslevyillä tehdystä pohjoisen taivaan POSS-II tutkimuksesta [4]. Toinen vuosituhannen vaihteessa toteutettu tutkimus oli Two Micron All Sky Survey (lyh. 2MASS, vuosi 1997–2001), joka kartoitti mikrometrialueella eli infrapuna-aallonpituuksilla lähes koko taivaanpallon alan [5]. Modernimman aikakauden ensimmäisiä tutkimuksia taas on esimerkiksi Sloan Digital Sky Survey -tutkimus (lyh. SDSS, vuodesta 2000 alkaen), joka kartoitti yli 35% koko taivaan alasta, tallentaen samanaikaisesti spektridataa ja kuvia eri aallonpituuksilla [7, 8]. SDSS-tutkimus on merkittävä myös open access -datan näkökulmasta – tutkimuksessa tuotettu data on avoimesti ladattavissa ja hyödynnettävissä edistäen data-analyysiteknologian kehitystä. Viimeiseksi mainittakoon vielä 2020-luvun survey-tutkimuksena Legacy Survey of Space and Time [8] (lyh. LSST), joka on lähiaikoina aloittamassa modernilla teknologialla, varustetun tutkimuksen keskittyen erityisesti aurinkokunnan ja linnunradan kartoitukseen, transienttien kohteiden havaitsemiseen sekä pimeän energian tutkimukseen. LSST-tutkimus on merkittävä erityisesti myös ennennäkemättömän suuren odotetun datamäärän vuoksi. Gravitational-wave Optical

Transient Observer (GOTO) [9] on korkeatasoista automatiikkaa ja useita havaintoteleskooppeja hyödyntävä tutkimusprojekti, joka keskittyy erityisesti gravitaatioaaltohavaintojen optisten vastineiden etsimiseen ja transienttien kohteiden havaitsemiseen muuna havaintoaikana. GOTO-kollaboraatio perustettiin vuonna 2014 ja nykyisessä muodossaan tutkimus on ollut toiminnassa vuodesta 2023.

Survey-tutkimukset tuottavat dataa koskien useita ja vaihtelevia kohteita, joita ei kaikkia ole ennalta kartoitettu tai luokiteltu. Modernit tutkimukset sisältävät lisäksi valtavan määrän kohteita ja raakaa dataa, mitä kaikkea ei voida ihmistutkijatyöllä käsitellä. Tästä johtuu nimitys ”big datan aikakausi”. [2]

2.2 Big Data

Käsite ”big data” määritellään yleisesti ”neljän V:n” kautta: Volume, Variety, Velocity ja Veracity [10, p. 3]. Nämä V:t kuvaavat erilaisia datan ominaisuuksia, jotka hankaloittavat sen käsittelyä perinteisillä menetelmillä. Volume kuvaa datan määrää tai kokoa, yleensä vaadittua tallennustilaa. Variety taas viittaa datan vaihteleviin muotoihin. Velocity viittaa siihen, kuinka nopeasti uutta dataa muodostuu. Ja viimeisenä veracity tarkoittaa datan puhtautta sekä luotettavuutta. Seuraavana on eritelty esimerkkejä näiden ominaisuuksien ilmenemisestä tähtitieteen datassa. [10]

Volume – Survey-tutkimusten tuottama datamäärä on valtava (katso liite 1), ja kasvaa edelleen uusien tutkimusten myötä. Datamäärä vanhemmissa tutkimuksissa on kymmeniä teratavuja, moderneissa petatavuja. Datamäärät muodostavat haasteita tallennustilan ja algoritmiseen data-analyysiin vaadittavan suoritustehon vuoksi, sekä datajoukon keräämiseen vaaditun tiedonsiirtoon. [2]

Variety – Tähtitieteen dataa on hyvin vaihtelevissa muodoissa. Survey-tutkimukset tuottavat pääasiassa kuva- ja/tai spektroskopiadataa. Nämä voivat kattaa vaihtelevia aallonpituusalueita. Kuvista ja spektreistä voidaan muodostaa aikasarjoja ja sekvenssejä. Data on myös vaihtelevan strukturoitua: kohteen tietoja kuten tyyppi, sijainti, tähden spektriluokka, ym. voidaan tallentaa ja hakea erinäisistä tähtitieteellisistä tietokannoista, mutta näitä ei välttämättä ole olemassa tai saatavilla kaikille kohteille. [1]

Velocity – Datan nopeus peilaa datamäärän käsitettä; mutta se keskittyy nimenomaan uuden datan tuotantonopeuteen olemassa olevan datajoukkojen koon tarkastelun sijasta. Datan nopeuteen liittyvät keskinäiset haasteet ovat jatkuva esikäsittelyn ja prosessoinnin vaatimus,

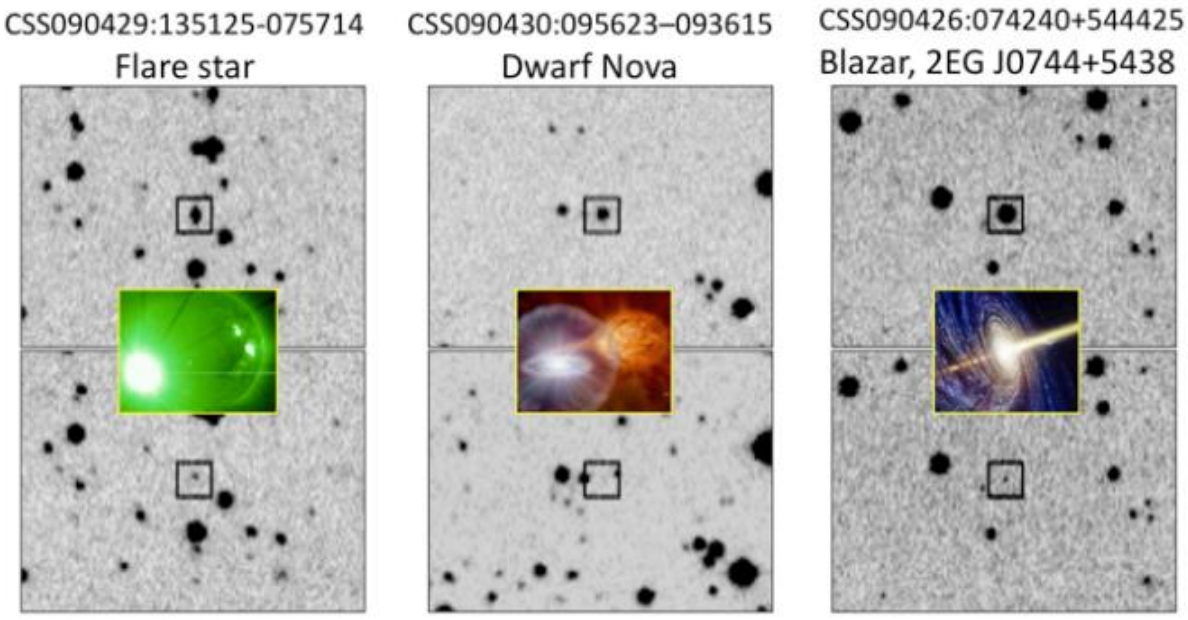
erityisesti kun tavoitellaan tieteellisesti relevantteja löydöksiä reaaliajassa. Esimerkiksi LSST-tutkimuksen odotetaan tuottavan 20TB käsittelemätöntä dataa jokaisena havaintoyönä [11].

Veracity – Datan luotettavuus. Tähtitieteen data sisältää huomattavasti haasteita puhtauden ja luotettavuuden suhteen. Päähaasteina ovat erinäiset kohinan lähteet, sekä datajoukkojen data-alkioiden epätäydellisyys. Epätäydellisyydellä tarkoitetaan tässä yhteydessä data-alkiolta puuttuvia tai käsittelykelvottomia arvoja. Havaintolaitteisto tuottaa aina kohinaa, minkä lisäksi teleskooppien sijainti, fyysinen rakenne ja optiikka sekä vaadittu datan esikäsittely voivat lisätä artefakteja dataan. [1]

2.3 Datan monimuotoisuuteen ja määrään liittyvät haasteet

Tähtitieteellistä dataa kootaan monimuotoisesti. Tämän tutkielman tarkastelun kannalta olennaisimpia tähtitieteen datamuotoja ovat strukturoidut katalogit tai tietokannat, teleskoopeilla tehdyt kuvahavainnot sekä näiden kuvien sekvenssit ja muut aikasarjat. Käsitettä ”tietokanta” käytetään tässä tutkielmaksi väljälti: tietokannalla tarkoitetaan mitä tahansa tähtitieteen dataa koneluettavassa muodossa sisältävää tietojärjestelmää.

Merkittävä modernien survey-tutkimusten kuten PanSTARRS:n ja LSST:n tiedetehtävistä on aikasarja-analytiikka. Aikasarjan analyysissä tarkastellaan havaintokohteen muuttuvuutta ajan suhteen. Näin voidaan havaita nk. transientteja kohteita, eli hetkellisesti muuttuneita kohteita. Muutos on yleisimmin joko kohteen kirkastuminen tai himmentyminen. Tällaisia ovat esimerkiksi supernova- ja kilonovaräjähdykset, sekä jotkut aktiivigalaksit. Modernien havaintotekniikan odotetaan nostavan transienttien kohteiden havaintomäärää $10^5 - 10^7$ havaintoon yössä nykyiseen $10 - 10^2$ havaintoon verrattuna [12]. Kuvassa 1 esitetään, miltä transientit kohteet saattavat näyttää havaintokuvissa. Hetkellisen muutoksen vuoksi on mielekästä tunnistaa transientit kohteet nopeasti, jotta jatkotutkimusta ehditään tekemään. Data-analytiikan kannalta olennaisimmat haasteet ovat suuren datamäärän nopea ja reaaliaikainen käsittely, sekä tunnistusprosessilta tai -algoritmilta vaadittava tarkkuus. Tietokonepohjaiset data-analyysitekniikat havaintojen tunnistamiseen tai suodattamiseen ovat olennaisia. [12, 13]



Kuva 1: Erilaisia transientteja kohteita, joita survey-tutkimuksen aikasarjoja tarkastellessa saatetaan havaita. Esitettynä on kolme erilaista kohdetta: vasemmalta oikealle flare-tähti, kääpiönova ja blazari (eräänlainen aktiivinen galaksiydin). Kuvissa alhaalla on aikasarjassa edeltävä havaintokuva, ylhäällä uudempi havaintokuva, sekä keskellä taiteellinen esitys (ei oikea havaintokuva) havaitusta ilmiöstä. Olennaista on, että täysin erilaiset ilmiöt voivat näyttäytyä lähes identtisenä hetkellisenä kohteen kirkastumisena aikasarjassa, ja tarkka luokittelu usein vaatii lisähavaintoja. [13]

Tähtitieteellisille kohteille on runsaasti tietokanta- tai katalogimuotoista dataa, kuten avoin SIMBAD-tietokanta [14], SDSS-datajulkaisut [15] ja ESA:n GAIA-datajulkaisut [16]. Katalogitietoa on runsaasti, mutta se on usein hajautunutta tai standardisoimatonta. Myös olemassa oleviin datakokoelmiin kohdistuu uutta tutkimusta, ja tehokkaat analyysimenetelmät ovat tarpeen. Tästä prosessista käytetään käsitettä datanlouhinta (eng. Data mining), ja se on kasvavasti tarpeellista olemassa olevan datan tehokasta hyödyntämistä varten. Informaatioteknologian menetelmien sovelluksia käytettäessä tähtitieteen dataan puhutaan myös tieteenalana astroinformatiikasta (eng. Astroinformatics). [1; 3, p. 436]

2.4 Datan moniulotteisuus, piirreirroitus ja esikäsittely

Klassiset koneoppimismenetelmät eivät voi suoraan käsitellä suurinta osaa alkuperäisestä tähtitieteellisestä datasta, vaan ne käsittelevät alkioita numeerisesti esitettävien ominaisuuksien eli piirteiden kautta. Datan piirre-esitys muodostetaan piirreirroituksella. Piirreirroituksen tavoite on kaapata olennaisin informaatiosisältö datasta niin, että piirre-esitystä käsittelevät koneoppimismenetelmät toimisivat mahdollisimman tehokkaasti ja tarkasti. [13]

Piirteiden joukkoa yksittäisestä kohteesta kutsutaan usein piirrevektoriksi. Tähtitieteelliselle kohteelle kuvadatasta eriteltäviä piirteitä voivat olla esimerkiksi kohteen magnitudi eli

kirkkaus tietyllä aallonpituusalueella, detektorilla havaittu koko ja kirkkauden jakauma kohteen koon suhteen [12]. Piirteitä voidaan eritellä myös esimerkiksi kohteen spektristä, jossa kuvataan kirkkauden jakaumaa eri aallonpituuksien suhteen tai valokäyrästä, joka kuvaa kirkkauden muutosta ajan suhteen [13]. Yksinkertaisimmillaan piirre-esityksenä voidaan käyttää kuvadatan yksittäisten pikseleiden arvoja. Klassisten menetelmien vaatima piirreirroitus lisää tämän datan esikäsittelyvaiheen. Toisaalta hyvä piirreirroitus tehostaa koneoppimismenetelmiä huomattavasti ja voi myös vähentää ylisovituksen riskiä malleja koulutettaessa. Lisäksi piirreirroitus voi normalisoida keskenään eri parametreillä tehtyjä havaintoja yhtenäiseksi datajoukoksi. [12, 13]

Ohjattujen koneoppimismenetelmien käyttöön tarvitaan opetusjoukko, joka koostuu valmiiksi luokitellusta edustusjoukosta, jonka tarkoitus on antaa algoritmille esimerkkejä kuhunkin kohdekategoriaan lajiteltavista kohteista. Esimerkiksi kuvissa esiintyvien transienttien kohteiden erittely artefakteista kuten kosmisista säteistä vaatii luokitellun joukon oikeita transientteja ja artefaktihavaintoja. Opetusjoukkojen muodostaminen on usein työlästä, ja vaatii tutkijatyönä tehtyä datan luokittelua. [17] Luokittelutyö voidaan välttää käyttämällä ohjaamattomia koneoppimismenetelmiä, jotka ryhmittelevät annetun datajoukon määrättyyn määrään luokkia keskinäisten samanlaisuuksien perusteella. Toisaalta luokittelutyötä voidaan yrittää tehostaa, esimerkiksi käyttämällä kansalaistiede- tai joukkoistamisprojekteja. Esimerkkinä tällaisesta on Galaxy Zoo -projekti [18], jossa joukkoistettiin galaksien luokittelu avoimella internet-alustalla.

Uudessa tutkimuksessa on usein mielekästä yhdistää hajautettua tietoa yhdestä kohteesta, esimerkiksi spektridatan liittäminen kohteen kuvadataan. Datan epätäydellisyys kuitenkin korostuu liitoksia tehdessä. Tietokannoissa on usein puuttuvia tietueita joillekin kohteille, tai tietueet sisältävät käsittelykelvottomia arvoja: esimerkiksi ”NaN”, ”0” tai ”-9999” kuvaamassa puuttuvaa numeerista tietoa. Tietueen arvo voi myös olla epäfyysinen: mahdoton joko havaintolaitteistolle tai kohteen tyypille. Kaikki koneoppimismenetelmät eivät voi käsitellä puuttuvia arvoja ollenkaan, tai huonot arvot voivat olennaisesti heikentää mallin tehokkuutta. Nämä poikkeamat on havaittava ja normalisoitava tai poistettava ennen koneoppimismalleilla käsittelyä. [19]

3 Koneoppimismenetelmistä

Koneoppimismenetelmiä on lukuisia, ja niitä voidaan jaotella tai kategorisoida usealla tavalla. Ensinnäkin menetelmät jakautuvat karkeasti klassisiin menetelmiin ja syväoppimismenetelmiin. Kuten Luvussa 2 mainittiin, klassiset menetelmät eivät käsittele alkuperäistä dataa, vaan datasta muodostettua piirreirroitusta. Syväoppimismenetelmät taas pohjautuvat neuroverkkoarkkitehtuureihin, jotka voivat käyttää syötteenään myös täysin alkuperäistä dataa kuten kuvatiedostoja tai -sekvenssejä. [13, 20]

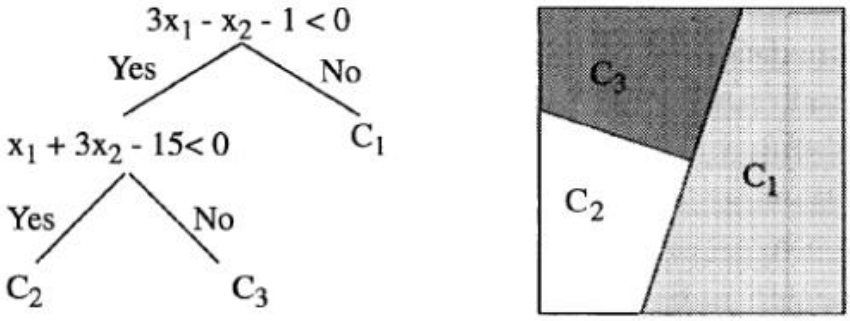
Koulutusmenetelmät jakautuvat ohjattuun ja ohjaamattomaan oppimiseen, sekä näiden yhdistelmänä osittaisohjattuun oppimiseen. Ohjattua oppimista käytettäessä opetusjoukko on valmiiksi luokiteltu, eli kaikki kohdeluokat ovat tiedossa. Ohjaamaton koneoppimismenetelmä taas muodostaa kohdeluokat itse käytetyn koulutusdatan perusteella. Osittaisohjatut menetelmät eivät välttämättä tarvitse valmiiksi luokiteltua dataa, mutta voivat hyödyntää valmiita luokitteluja niiden olemassa ollessa. Lisäksi on olemassa vahvistusoppimismenetelmiä, jossa algoritmille annetaan jatkuvaa palautetta tehtävässä suoriutumisesta. Suurin osa tähtitieteessä sovelletuista menetelmistä ovat ohjattuja tai ohjaamattomia. [1]

Tässä luvussa käsiteltävät koneoppimismenetelmät on valikoitu Lukuun 4 valikoitujen yksittäisten tutkimusten sekä tähtitieteen alan yleisessä kirjallisuudessa esiintyvien menetelmien mukaan [19]. Keskeisesti käsitellään siis klassisista menetelmistä päätöspuita ja metsiä, tukivektorikoneita (eng. support vector machine, SVM), lähimmän naapurin algoritmeja, sekä erilaisia neuroverkko- ja syväoppimismenetelmiä. Algoritmeista käsitellään yleisen menetelmän kuvauksen lisäksi olennaisimmat ominaisuudet sekä vahvuuksia tai haasteita.

3.1 Klassiset koneoppimismenetelmät

3.1.1 Päätöspuut ja metsät

Päätöspuu on binääripuumuotoinen, pääasiassa luokitteluun käytettävä algoritmi, jossa aloitetaan koko koulutusjoukon sisältävällä juurisolmulla, ja päädytään johonkin määrään päätesolmuja yksisuuntaisten kaarien kautta. Juurisolmu jakautuu lapsisolmuihin, joihin juurisolmun sisältämä joukko jaetaan solmun kynnysarvon tai funktion perusteella. Solmut jakautuvat edelleen ja edelleen, kunnes päästään lopetuskriteeriin, esimerkiksi maksimisyvyyteen tai minimipopulaatioon yhdessä solmussa. Kuvassa 2 on esitettynä yksinkertainen esimerkki päätöspuusta, jonka tehtävänä on kaksiulotteisen avaruuden jako kolmeen luokkaan avaruuden pisteen sijainnin perusteella. [19]



Kuva 2: Yksinkertainen päätöspuu, jossa puulle annetaan kaksi syötettä x_1, ja x_2, jotka esittävät kuvassa oikealla kuvatun avaruuden pisteitä. Päätöspuu jakaa avaruuden kolmeen luokkaan: c1, c2 ja c3. [19]

Päätöspuualgoritmia koulutettaessa keskeisin tavoite on muodostaa solmuille päätöskriteeri, joka jakaa solmun populaation mahdollisimman tehokkaasti tavoiteluokkia kohti. Päätöskriteerit voivat olla jaottelu yhden piirteen perusteella, tai kuten kuvan 2 esimerkissä sisältää useita piirteitä. Keskeisenä etuna päätöspuumenetelmässä on yksinkertaisuus ja tulkittavuus: solmujen päätöskriteereistä on suoraan pääteltävissä tiettyä luokitusta varten olennaiset piirteet. Suurimpia ongelmia päätöspuumenetelmissä taas ovat säädettävien parametrien määrä ja herkkyys ylisovitukselle. Päätöspuita voidaan yhdistää muodostaen esimerkiksi satunnaismetsämalleja. Satunnaismetsä (RF) koulutetaan rakentamalla piirteiden määrän verran päätöspuita, jotka koulutetaan samoilla parametreillä satunnaistaen alkion tietty piirre jokaisessa puussa. Satunnaismetsä käsittelee data-alkion antamalla tämän syötteeksi näille puille yhdistäen tulokset esimerkiksi keskiarvon tai moodiluokituksen perusteella. [3, 19]

3.1.2 Tukivektorikoneet

Tukivektorikoneet eli SVM:t ovat tähtitieteen alalla huomattavan suosittuja koneoppimismenetelmiä. Tukivektorikonetta käsitellessä on luontevaa ajatella datajoukon piirre-esitystä n-ulotteisena vektorina, jossa n on piirteiden määrä. Tukivektorikone muodostaa tähän vektoriavaruuteen hypertason, joka jakaa sen kahteen kohdeluokkaan. Keskeisyys kahtiajakoon rajoittaa hieman tukivektorikoneen sovelluksia, mutta yleistyksiä useampiin luokkiin jakavista koneista on kehitetty. Tukivektorikone on yksinkertainen malli soveltaa, sillä säädettävien parametrien määrä on pieni. [1, 19]

Tukivektorikoneen etuja ovat kohinansietokyky, yksinkertaisuus soveltaa, sekä deterministisyys. Deterministisyydellä tarkoitetaan tässä sitä, että SVM löytää aina saman ratkaisun samalle syötedatalle, eikä voi jäädä jumiin parhainta ratkaisua löytämättä ylisovituksen tai virhefunktion lokaalin ääriarvokohdan vuoksi. Toisaalta mainittu keskeisyys kahteen luokkaan jaottelussa, herkkyys epärelevanteille piirteille ja huono suorituskyky ja tulkittavuus ovat merkittäviä heikkouksia. Tukivektorikone ei myöskään muodosta varsinaista mallia, vaan se jakaa ainoastaan syötteenä annetun datajoukon. [19]

3.1.3 Lähimmän naapurin menetelmät

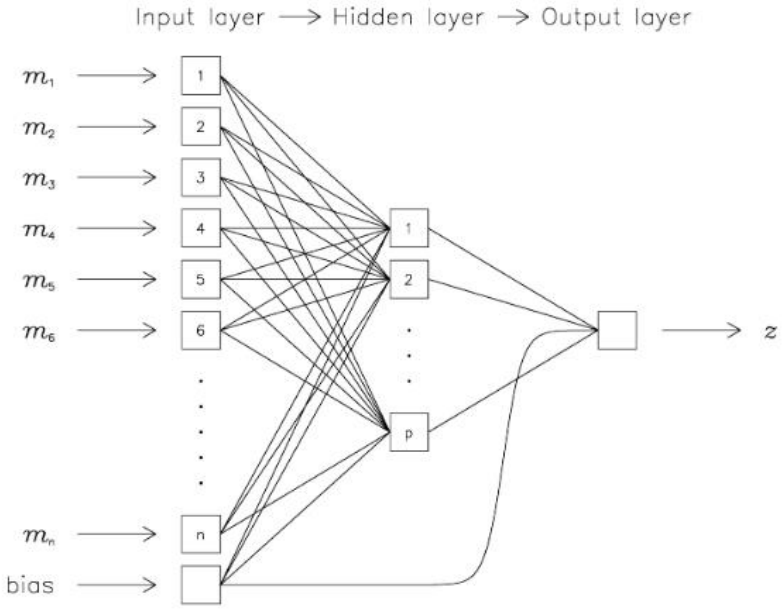
Lähimmän naapurin menetelmistä yksinkertaisin on k-nearest-neighbors -menetelmä (kNN). Menetelmä on ohjattu, eli vaatii luokitellun opetusjoukon. Malli muodostetaan yksinkertaisesti asettamalla koulutusjoukon alkiot vektoriavaruuteen piirrevektoreidensa mukaisesti. Testialkio asetetaan samaan vektoriavaruuteen, ja etsitään sille määrätty määrä geometrisesti lähimpiä naapureita. Näiden naapurien luokitus opetusjoukossa määrää testialkion luokan, johon algoritmi testialkion sijoittaa. Algoritmi on verrattain nopea, sillä se ei vaadi erillistä koulutusta, vaan opetusjoukko itsessään on algoritmin käyttämä ”malli”. Suoritusaika voi kuitenkin olla merkittävä, erityisesti datan sisältäessä hyvin monta ulottuvuutta. [1]

Ohjaamaton vastine kNN -menetelmälle on k-rypästelymenetelmä (eng. k-means-clustering). Tätä menetelmää käytettäessä asetetaan pisteet avaruuteen samoin kuin kNN-menetelmässä. Algoritmin aloitusparametrinä on tavoitejoukkojen määrä, sekä aloituspisteet joukkojen keskipisteelle. Algoritmi määrittää iteratiivisesti keskipisteiden perusteella ensin joukkoon kuuluvat alkiot, sitten uuden keskipisteen joukolle, ja tätä toistetaan edelleen, kunnes joukot ovat stabiileja eli eivät enää muutu suoritettaessa lisää iteraatioita. Tämä algoritmi ei

muodosta varsinaista mallia, jota voitaisiin käyttää uudemman datan käsittelyyn, vaan jaottelee ainoastaan syötteenä annetut alkiot tiettyyn määrään kohdejoukkoja. [19]

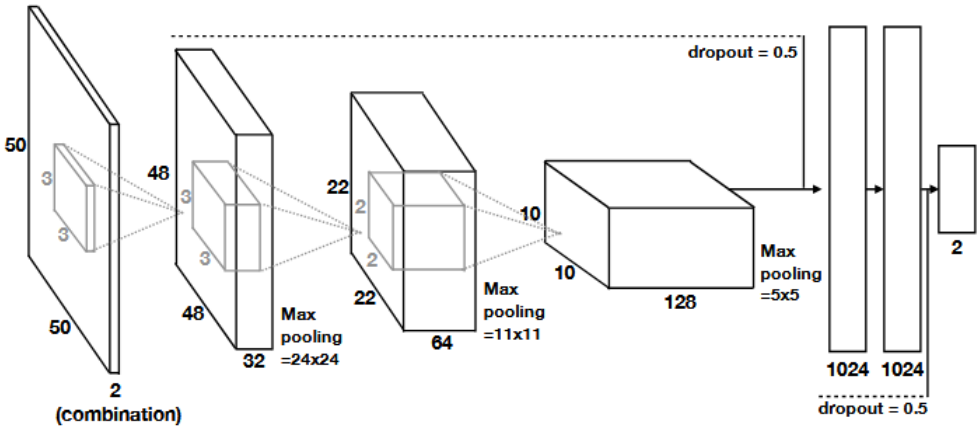
3.2 Neuroverkko- ja syväoppimismenetelmät

Neuroverkkomenetelmät ovat verkkoja, jotka koostuvat neuroneista. Perinteinen neuroverkko (ANN) koostuu neuroneista, jotka sisältävät jonkin määrän syötteitä. Koulutettaessa neuroni muodostaa sisäisen painoarvon jokaiselle syötearvolle, ja prosessoi painotetut syötteet aktivaatiofunktiolla, kuten kynnysarvo- tai käyräfunktiolla, ja välittää tämän tuloksen eteenpäin. Neuronit ovat kytkettyinä kerroksittaisessa rakenteessa, jossa neuroni saa syötteensä edeltävältä kerrokselta ja antaa tuloksensa seuraavalle kerrokselle. Ensimmäinen kerros saa syötteen lähdedatasta, ja viimeinen kerros muodostaa loppuarvon neuroverkolle: esimerkiksi luotettavuustason sille, että syötedata kuuluu tiettyyn kohdeluokkaan. Välikerroksia kutsutaan piilokerroksiksi. Neuroverkon suorituskykyä arvioidaan virhefunktiolla, joka kuvaa mallin epätarkkuutta tai virhettä suoritettavassa tehtävässä. Koulutus perustuu virhefunktion minimoimiseen, ja voidaan toteuttaa usealla eri algoritmilla. Tähtitieteessä suosituin on tunnettu vastavirta-algoritmi, jossa painoja ja aktivaatiofunktiota muutetaan yksitellen seuraten neuroverkkoa käänteisessä suunnassa. Kuvassa 3 on esitettynä yksinkertainen neuroverkkoarkkitehtuuri, jota käytettiin galaksien punasiirtymän estimointiin tutkimuksessa [21], jossa kohdejoukko oli lineaarinen parametri z eli kohteen punasiirtymä. Esimerkiksi luokittelutehtävää suorittava neuroverkko sisältäisi useamman ulostuloneuronin, yhden jokaiselle kohdeluokalle. [19]

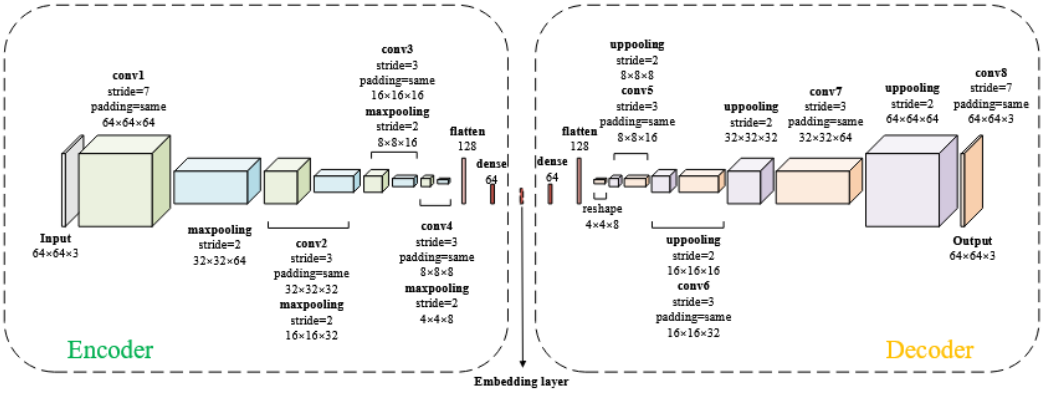


Kuva 3: Esitettynä on yksinkertainen neuroverkkoarkkitehtuuri. Kyseessä on monitasoinen perseptroni, jossa vasemmalta oikealle on syötekerros, yksi piilokerros, sekä ulostulokerros, ja yksi bias-arvo. Monimutkaisemat neuroverkot voivat sisältää lisää piilokerroksia ja/tai useamman neuronin ulostulokerroksessa. [21]

Syväoppimismenetelmiksi kutsutaan sellaisia neuroverkkoarkkitehtuureja, jotka eivät tarvitse piirreirroitusta, vaan neuroverkko itse muodostaa olennaiset ominaisuudet raakadatasta. Olennaisimpia tällaisia arkkitehtuureja ovat konvoluutioneuroverkot (CNN), esitettynä kuvassa 4, sekä autoenkooderit (AE tai CAE), joka on esitettynä kuvassa 5. Molemmat näistä ovat pääasiassa kuvadatan käsittelyyn optimoituja menetelmiä. CNN- ja CAE-verkot muodostavat implisiittisen piirreirroituksen, ja rakentuvat konvoluutio- pooling- ja yhdistelmäneuroverkkokerroksista. Konvoluutiokerroksen neuronit koostuvat konvoluutiomatriisista, jolla syötekerroksen data prosessoidaan. Pooling-kerrokset yksinkertaistavat piirrekarttaa. Yhdistelmäkerrokset taas ovat perinteisten neuroverkkojen piilokerrosten kaltaisia. Nämä kerrokset muodostavat lopullisen luokituksen tai piirre-esityksen, ja ovat CNN:n viimeinen kerros, kun taas AE:t sisältävät toisen, samanlaisen mutta peilatun, verkkoarkkitehuurin kytkettynä edeltävän ulostuloneuroneihin. [1, 20]



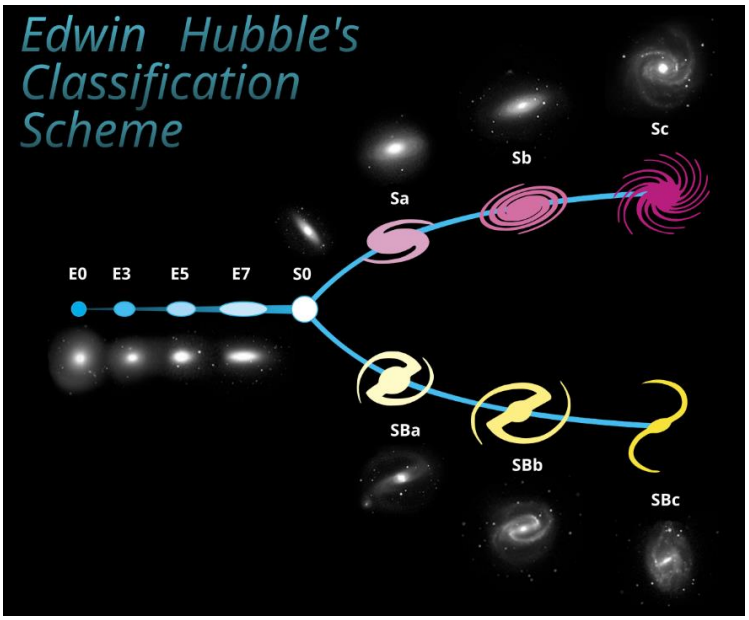
Kuva 4: Konvoluutioneuroverkko, jossa neuroverkko saa syötteenään 50x50px-resoluution kuvadataa, ja sisältää kolme konvoluutiokerrosta, kaksi piilokerrosta ja kaksi kohdeluokkaa. Konvoluutiokerros ottaa syötteenään tietynkokoisen (tässä 3x3px, 3x3px, ja 2x2px) otteen syötedatasta, ja muodostaa vektorituloa käyttäen uuden arvon seuraavaa kerrosta varten. [22, p. 8]



Kuva 5: Autoenkooderi, joka muodostuu kahdesta vastakkain asetetusta symmetrisestä konvoluutioverkosta, joista ensimmäinen on enkooderikerros ja toinen dekooderikerros. Kuvassa vihreät ja oranssit värit kuvaavat konvoluutiokerroksia ja siniset ja violetit pooling-kerroksia. Enkooderi muodostaa piirre-esityksen syötedatasta, ja dekooderi muodostaa rekonstruktion syötedatasta piirre-esityksen perusteella. [23]

3.3 Kohteiden luokittelu ja ryhmittely

Luokittelu ja ryhmittely ovat tähtitieteessä yksi olennaisimmista tehtävistä informaation organisoinnin kannalta. Esimerkkejä tällaisista tehtävistä ovat tähtien ja galaksien erittely toisistaan, galaksityyppien luokittelu muodon perusteella ja tähtien luokittelu spektroskooppisiin luokkiin. Olennaista on huomata, että luokittelu ja ryhmittely ovat kuitenkin olennaisesti erilaisia: luokittelussa nojataan ihmistutkijoiden määrittelemiin luokkiin, kun taas ryhmittelyssä luokat muodostuvat käsiteltävän datan pohjalta. Luokittelu on näin ollen luontaisesti ohjatun menetelmän vaativa tehtävä, kun taas ryhmittely on luontaisesti ohjaamatonta. Kuvassa 6 esitetään galaksien Hubble-luokitus. Galaksien luokittelu Hubble-luokkiin on yksi olennaisimmista luokittelutehtävistä tähtitieteen datassa. [3, 18]



Kuva 6. [24] Galaksien luokittelu morfologisesti eli muodon mukaan Hubble-luokkiin. Galaksit jakautuvat elliptisiin (E0-E6), spiraaleihin (Sa-Sc), sauvaspiraaleihin (SBa-SBc) ja linssigalakseihin S0.

3.4 Poikkeamanhavainto

Poikkeamanhavainnolla tai anomaliteettien tunnistuksella tarkoitetaan tehtävää, jossa datajoukosta etsitään olennaisia tai kiinnostavia poikkeama-alkioita. Tähtitieteessä poikkeamien havainta on hankalaa, sillä data on harvoin puhdasta. Kuten Luvussa 2 käsiteltiin, tähtitieteellinen data sisältää usein kohinaa sekä havaintolaitteiston tai ympäristön tuottamia artefakteja, esimerkiksi radiohäiriöitä tai kosmisia säteitä, jotka esiintyvät havaintokuvissa kirkkaina viivoina. Keskeinen tavoite on löytää datasta kohteita, jotka ovat ominaisuuksiltaan normista poikkeavia, ja eritellä ne havaintolaitteiston tuottamasta kohinasta, artefakteista, sekä tilastollisesta vaihtelusta. [25]

3.5 Aikasarja-analytiikka

Aikasarjojen analytiikka on modernissa tähtitieteessä keskeistä: esimerkiksi LSST-tutkimuksen yhtenä päätehtävästä on aikasarjojen muodostus havaintokohteista. Aikasarjojen metodisella hyödyntämisellä voidaan määrittää kohteiden aikariippuvaisia muutoksia tai tunnistaa uusia ilmiöitä ja kohteita. Toisaalta aikasarjat ovat huomattavasti staattista tietoa hankalampia käsitellä, erityisesti kun dataa tuotetaan reaaliaikaisesti. Keskeisiä haasteita ovat esimerkiksi eri havaintolaitteilla tai vaihtelevin väliajoin otetut datapisteet, havaintokohteen lähellä olevat kohteet ja vaatimus havaintojen nopeaan käsittelyyn. Lisäksi aikasarjoista voidaan etsiä uusia ilmiöitä, joita ei ole ennemmin havaittu, eikä siten tarvittavaa koulutusdataa ole välttämättä olemassa. [26]

4 Sovellutuksien tarkastelua

Tässä luvussa tarkastellaan tutkimusesimerkkejä tähtitieteessä esiintyvistä tehtävistä, joissa on sovellettu koneoppimismenetelmiä, sekä näillä menetelmillä saatuja tuloksia. Tutkimukset on jaoteltu karkeasti suoritetun tehtävän mukaan. Selkeitä klusterointiongelmia on tähtitieteessä verrattain vähän, joten klusterointia hyödyntäneitä tutkimuksia käsitellään jonkin edeltävän kolmen luokan osana. Lisäksi monet poikkeamanhavainto- ja aikasarja-analyysitehtävät liittyvät jossain määrin toisiinsa, esimerkiksi tietynlaisen muutoksen etsintä kuvasarjasta on sekä poikkeamanhavainto- että aikasarjaongelma. Näitä tutkimuksia käsitellään asiayhteyden mukaan parhaiten sopivassa luokassa. Algoritmien tehokkuutta arvioidaan yleisesti kolmella parametrillä: täydellisyys (eng. completeness tai true positive rate) $TPR = \frac{TP}{TP+FN}$, puhtaus (eng. purity) $PRT = \frac{TP}{TP+FP}$. Lisäksi voidaan puhua virhepositiivisista, kun käsitellään yksittäiseen luokkaan kuuluvia havaintoja (eng. false positive rate), $FPR = \frac{FP}{FP+TN}$. Kaavoissa TP viittaa oikeisiin positiivisiin, eli oikeaan luokkaan luokitellut alkiot ja FP virheellisesti käsiteltävään luokkaan luokitetut alkiot. FN viittaa käsiteltävään luokkaan kuuluneita alkioita, joiden luokitus oli virheellinen ja TN alkioita, jotka luokiteltiin oikein eri luokkaan kuin tarkasteltavaan.

4.1 Luokittelu

Jiménez et al. [27] käsittelee galaksien luokitteluongelmaa kuvadatasta. Tutkimus käsitteli useaa erilaista koneoppimispohjaista lähestymistapaa ongelmaa varten. Piirreirroitusta käyttävien klassisten menetelmien (kNN, RF, ja SVM) suoriutumista verrattiin eri piirreirroitusmenetelmillä: WND-CHARM-luokittelijaa [28] ja CAE-pohjaista menetelmää käyttäen. Näitä menetelmiä verrattiin edelleen CNN-luokittelijaan, jolle alkuperäinen data voitiin syöttää suoraan ilman piirreirroitusta. Menetelmien keskinäisen vertaamisen lisäksi tutkittiin, onko joukkoistamalla kerätyn opetusdatan vaikutusta menetelmien suoriutumiseen käyttäen Galaxy Zoo -joukkoistamisprojektin [18] dataa. Sama algoritmi koulutettiin erikseen joukkoistetulla ja tutkijoiden muodostamalla opetusjoukolla keskinäistä vertailua varten. Keskeisinä päätelminä esitetään konvoluutioverkkojen olevan yleisesti ottaen hyvä kompromissi tarkkuuden ja suorituskyvyn välillä. Neuroverkkoarkkitehtuuria syvennettäessä huomattiin vähenevää tuottavuutta tarkkuuden suhteen. Piirreirroitusta käsittelevistä klassisista menetelmistä satunnaismetsä oli tarkin, mutta ero kNN:ään ja SVM:ään oli verrattain pieni. Autoenkooderien todetaan olevan lupaavia piirreirroitusmenetelmiä

erityisesti siksi, että niitä käyttämällä voidaan erottaa piirreirroitusprosessi ja luokittelijan koulutus toisistaan. CAE oli WND-CHARM-irroitusta huomattavasti tehokkaampi suorituskyvyltään, mutta hieman epätarkempi. Joukkoistamisen todetaan myös olevan lupaava tehostusmenetelmä opetusjoukkojen muodostamista varten. [27]

Cheng et al. [22] käsittelee erilaisten galaksityyppien morfologista luokittelua elliptisiin ja spiraaligalakseihin. Tutkimuksessa verrataan konvoluutioverkkoja, K-naapurin menetelmää, SVM:ää, satunnaismetsiä sekä neuroverkkoja. Keskeisenä tutkimuskysymyksenä on luokitteluun parhaiten soveltuvan menetelmän selvittäminen, sekä yhdistelmät analyysimenetelmistä, joilla tarkkuutta voitaisiin parantaa. Merkittävä löydös oli konvoluutioverkon paljastama puute lähdemateriaalissa. Materiaali ei sisältänyt luokitusta linssimäiselle (S0) galaksille, ainoastaan spiraali- ja ellipsigalakseille. Ohjatuillakin menetelmillä lisäluokka huomattiin, sillä sen piirteitä ilmentävät galaksit tasaisesti antoivat hyvin heikon luottamuksen tunnettuihin kohdekategorioihin. Virheluokitusten huomioinnin jälkeen päästiin yli 0.99 TPR-tarkkuuteen galaksien tyyppiluokittelussa konvoluutioverkolla, joka oli menetelmistä selkeästi tarkin. Satunnaismetsillä saavutettiin n. 0.95 TPR-tarkkuus, ja SVM:llä hieman yli 0.75 TPR. [22]

4.2 Poikkeamanhavainto

Han et al. [29] käsittelee poikkemanhavaintoa Galaxy Zoo -julkaisun [18] galaksien kuvadatasta ohjaamattomalla oppimisella, käyttäen kNN-menetelmää itsenäisesti ja autoenkooderin avustamana. Verrattavana olivat: kNN, kNN+CAE ja kNN+CAE+Huomiomekanismi. Aineistodata sisälsi kuvadataa galakseista, lajiteltuna 5 eri luokitukseen: elliptinen, osittaiselliptinen, sikarimainen elliptinen, linssimäinen, ja spiraali. Poikkeamanhavaintoa varten muodostettiin uudet datajoukot, jotka koostuivat pääasiassa tyypin 0 eli elliptisistä galakseista, sisältäen pienen määrän jonkin muun luokan galakseja. Löydöksinä esitetään kNN:n soveltuvan tähtitieteelliseen aineistoon, mutta suorituskyvyn ja tarkkuuden: saavutettu TPR enintään 0.25 olevan puutteellisia, erityisesti käsin rakennetun piirre-esityksen heikkouden vuoksi. KNN-menetelmään liitettiin autoenkooderimalli piirreirroituksen automatisoinniksi, mikä selkeästi paransi algoritmin suorituskykyä sekä tarkkuutta: näin saavutettiin maksimi TPR 0.56. Huomiomekanismin lisäys paransi tutkimuksessa mallin tarkkuutta edelleen, maksimi TPR 0.78, suorituskykyä heikentämättä. Huomattavaa tässä tutkimuksessa oli datajoukon yksinkertaisuus ja synteettisyys realistiseen tutkimukseen verrattuna. Tarkoituksena oli ennemmin yleisen kNN:n ja ohjaamattoman

oppimisen soveltuvuuden kartoitus. Tästä todetaan lopputuloksena myönteisesti: erityisesti kehittyneemmillä piirreirroitusmenetelmillä ohjaamaton kNN-menetelmä sopii tähtitieteen aineistolle. [29]

Mutukrishna et al. [25] tarkastelee poikkeamanhavaintoa transienteista valokäyristä, kohdistuen uusiin ennen näkemättömiin havaintoihin esimerkiksi LSST:n kautta. Tunnistettavana ovat harvinaiset transientit kohteet, verrattuna yleisiin. Tarkoitus on siis välttää ongelma, että ohjatut luokittelijat voivat havaita vain sellaisia anomaliteetteja, joilla ne on koulutettu. Verrattavana ovat aika-avaruudellinen konvoluutioverkko (jossa siis lisäulottuvuutena aikasarja, eteen kytkettynä kerroksittain) ja bayesiaaninen matemaattinen malli. Molemmat mallit todetaan lupaaviksi, ja DNN:än mukautuvuutta, nopeutta ja riippumattomuutta tutkijatyöstä korostetaan, mutta itse poikkeamanhavainto-ongelmassa matemaattinen malli on selkeästi tarkempi. Mallit esitetään hyödyllisinä tulevaa tutkimusta varten, lisätutkimusta ehdotetaan osittaisohjattua tai aktiivista oppimista hyödyntäviin malleihin. [25]

Tarkkarajaiset poikkeamanhavainto-ongelmat ovat tähtitieteessä harvinaisempia. Lisäksi tutkimuksista kuvautuu, että näihin ongelmiin soveltuu usein myös matemaattinen malli koneoppimismenetelmien sijaan. Esimerkkinä tästä on tutkimus Han et al. [30], jossa elliptistä rajausta käytettiin esikäsittelymenetelmänä koneoppimismallien syötedatalle, parantaen mallien tarkkuutta huomattavasti.

4.3 Aikasarja-analytiikka

Wright et al (2015) [17] käsittelee transienttien kohteiden havaintoa Pan-STARRS1 -surveyn differenssikuvista. Differenssikuva esittää kahden eri aikaan otetun havaintokuvan eroa, ja muodostetaan pikselittäisellä erotuksella uuden ja vanhan havaintokuvan välillä. Ongelma yksinkertaistetaan differenssikuvien luokitteluun real-bogus -luokkiin. Real-luokka koostuu aidoista transienteista kohteista, ja bogus-luokka virheellisistä eli artefaktihavainnoista, joita muodostavat esimerkiksi havaintolaitteen toimintahäiriöt tai esikäsittelyprosessissa tapahtuneet virheet. Tutkimuksessa käsitellään sekä klassisia menetelmiä: satunnaismetsiä ja SVM:ää, että neuroverkkomenetelmää kolmikerroksista ANN-verkkoa käyttämällä. Piirreirroitus muodostettiin suoraan kuvadatan pikseliarvoista – jokainen pikseli muodostaa yhden piirteen, jonka arvo on pikselin havaittu intensiteetti. Data muodostui 20x20 pikselin differenssikuvista, ikkunoituna havaintokohteeseen keskitetysti. Piirre-esityksenä käytettiin jokaisen pikselin kirkkausarvoa erikseen, eli jokaiselle kohteelle muodostui 400 piirrettä.

Tehokkain luokittelualgoritmi oli satunnaismetsä, toisena tukivektorikone (SVM) ja heikoimpana neuroverkko (ANN). Lopputuloksena näillä menetelmillä esitetään korkein saavutettu TPR 0.938 ja PRT 0.99 real-luokitus. Haasteita ja virheellisiä luokituksia tuottivat erityisesti hyvin kirkkaat kohteet. Lisätutkimusta ehdotetaan koskien kehittyneempiä piirreirroitusmenetelmiä, suurempia ja neuroverkkoarkkitehtuureja, sekä osittaisohjattujen menetelmien sovellutuksien tarkastelua. [17]

Killestein et al. (2021) [31] käsittelee transienttien kohteiden havaitsemista konvoluutioneuroverkon avulla GOTO-observatorion datasta. Ongelma yksinkertaistetaan samoin kuin edeltävässä tutkimuksessa havaintokandidaattien aito–väärä- luokitteluun. Opetusjoukkona käytettiin synteettistä sarjaa, joka muodostui asteroidikohteesta, visuaalisesti hyvin samankaltainen räjähdystyyppisen (esim. supernova) transientin kohteen kanssa. Opetusjoukko muodostettiin synteettisesti, sillä tutkittavana olleella uudella havaintolaitteistolla saatua dataa ei vielä ollut riittävästi. Huomattavasti oikea transientti saattaa esiintyä monimutkaisemmassa ympäristössä sitä ympäröivän galaksin vuoksi, kun taas synteettisesti muodostettu opetusjoukko esitti galaksien ulkopuolisia kohteita. Käytetty konvoluutioverkko perustui 16-kerroksiseen VGG-16 -arkkitehtuuriin [32]. Keskeisinä löydöksinä esitetään neuroverkkopohjaisen datankäsittelymallin soveltuvan käsiteltyyn tehtävään. Neuroverkkomallilla onnistuttiin havaitsemaan kiinteällä FPR 0.01 kaikki paitsi 0.5 % aidoista kohteista, esittäen 30 % parannuksen aikaisemmilla tutkimuksilla saavutettuun tarkkuuteen verrattuna. Tutkimuksessa kehitetyn arkkitehtuurin esitetään soveltuvan GOTO-havaintolaitteistolla tuotetun datan käsittelyyn jatkossakin. [31]

5 Yhteenveto

Tutkielmassa tarkasteltiin tähtitieteellisen datan erityisominaisuuksia, ja haasteita, eritellen näitä klassisiin big data -haasteisiin sekä tähtitieteelle uniikkeihin haasteisiin. Tutkielmassa esiteltiin tähtitieteen dataan sovellettuja koneoppimismenetelmiä sekä tutkimusesimerkkejä koneoppimismenetelmien soveltamisesta tähtitieteen dataan.

Tutkimuskysymyksen TK1 suhteen huomataan tähtitieteen suurien datajoukkojen ilmentävän monia klassisia big data -haasteita. Erityisesti korostuvat datan määrä, varieteetti sekä luotettavuus – tähtitieteen dataa tuotetaan monilla erilaisilla tutkimuksilla, ja se on tallennettuna vaihtelevissa rakenteissa. Lisäksi datan puhtaus aiheuttaa huomattavia haasteita, sillä havaitsevana tieteenalana instrumenttien tuottama data ei koskaan ole täysin eheää, kohinatonta tai virheetöntä. Tähtitieteessä on useita erilaisia ongelmia: tietokantojen datan jalostus, uusien tutkimusten datan reaaliaikainen prosessointi, ja datajoukkojen yhdistys, joihin kaikkiin tarvitaan erilaisia kehittyneitä analyysimenetelmiä, joiden on siedettävä tähtitieteen datalle ominaiset haasteet.

Kysymyksen TK2 suhteen voidaan todeta, että koneoppimismenetelmien sovellukset ovat hyvin lupaavia tähtitieteen datan prosessoinnissa. Klassisista menetelmistä satunnaismetsät sekä SVM:t ovat suosittuja, ja tutkimustulokset näiden soveltuvuudesta ovat melko hyviä. Modernimmat syväoppimismenetelmät ovat erityisesti lupaavia, uudemmissa tutkimuksissa kuten [31] on saatu erinomaisia lopputuloksia syväoppimismalleilla. Toisaalta syväoppiminen ei kuitenkaan ole yleisratkaisu kaikkiin data-analyysiongelmiin. Erityisesti poikkemanhavainto-ongelmissa perinteisemmät tai puhtaasti matemaattiset mallit voivat joissain tapauksissa olla soveltuvampia.

Yleisesti ottaen koneoppimismenetelmät vaikuttavat hyvin lupaavilta tähtitieteen alan modernien data-analyysin ongelmien ratkaisun työkaluna. Modernit kuvadatan käsittelyyn erikoistuneet syväoppivat neuroverkot soveltuvat hyvin myös tähtitieteen datan tutkimukseen ja autoenkooderimallit ovat lupaava menetelmä piirreirroitukseen klassisia menetelmiä varten. Ohjatut koneoppimismenetelmät ovat tähtitieteessä suosituimpia, mutta haastetta aiheuttaa vaaditun opetusdatan muodostaminen, varsinkin uusilla havaintolaiteasetelmilla tai uusia ilmiöitä tutkiessa, missä esimerkkidataa ei ole riittävästi, jos ollenkaan. Tutkimusta osittaisohjattujen mallien soveltamisen suhteen ehdotetaan useasti. Joukkoistus vaikuttaa lupaavalta keinolta tehostaa opetusdatan luokittelua.

Lähteet

[1] S. Sen, S. Agarwal, P. Chakraborty, and K. P. Singh, “Astronomical big data processing using machine learning: A comprehensive review,” *Exp. Astron.*, vol. 53, no. 1, pp. 1–43, Feb. 2022, doi: 10.1007/s10686-021-09827-4.

[2] Y. Zhang and Y. Zhao, “Astronomy in the Big Data Era | Data Science Journal,” May 2015, doi: 10.5334/dsj-2015-011.

[3] T. D. Oswalt and H. E. Bond, *Planets, Stars, and Stellar Systems volume 2: astronomical techniques, software and data*. doi: 10.1007/978-94-007-5618-2.

[4] S. G. Djorgovski *et al.*, “The Palomar Digital Sky Survey (DPOSS),” Sep. 15, 1998, *arXiv*: arXiv:astro-ph/9809187. doi: 10.48550/arXiv.astro-ph/9809187.

[5] M. F. Skrutskie *et al.*, “The Two Micron All Sky Survey (2MASS),” *Astron. J.*, vol. 131, no. 2, pp. 1163–1183, Feb. 2006, doi: 10.1086/498708.

[6] D. G. York *et al.*, “The Sloan Digital Sky Survey: Technical Summary,” *Astron. J.*, vol. 120, no. 3, pp. 1579–1587, Sep. 2000, doi: 10.1086/301513.

[7] “Science Results - SDSS.” Accessed: Nov. 08, 2025. [Online]. Available: https://www.sdss.org/science/

[8] Ž. Ivezić *et al.*, “LSST: From Science Drivers to Reference Design and Anticipated Data Products,” *Astrophys. J.*, vol. 873, no. 2, p. 111, Mar. 2019, doi: 10.3847/1538-4357/ab042c.

[9] M. J. Dyer *et al.*, “The Gravitational-wave Optical Transient Observer (GOTO)”.

[10] B. Furht and F. Villanustre, *Big Data Technologies and Applications*.

[11] L. S. S. Telescope, “Data Management,” Rubin Observatory. Accessed: Oct. 21, 2025. [Online]. Available: https://www.lsst.org/about/dm

[12] A. D’Isanto *et al.*, “An analysis of feature relevance in the classification of astronomical transients with machine learning methods,” *Mon. Not. R. Astron. Soc.*, vol. 457, no. 3, pp. 3119–3132, Apr. 2016, doi: 10.1093/mnras/stw157.

[13] C. Donalek, S. G. Djorgovski, A. A. Mahabal, M. J. Graham, A. J. Drake, and A. A. Kumar, “Feature selection strategies for classifying high dimensional astronomical data sets,” doi: 10.1109/BigData.2013.6691731.

[14] M. Wenger *et al.*, “The SIMBAD astronomical database. The CDS reference database for astronomical objects,” *Astron. Astrophys. Suppl. Ser.*, vol. 143, pp. 9–22, Apr. 2000, doi: 10.1051/aas:2000332.

[15] S. Collaboration *et al.*, “The Nineteenth Data Release of the Sloan Digital Sky Survey,” Jul. 09, 2025, *arXiv*: arXiv:2507.07093. doi: 10.48550/arXiv.2507.07093.

[16] G. Collaboration, “The Gaia mission,” *Astron. Astrophys.*, vol. 595, p. A1, Nov. 2016, doi: 10.1051/0004-6361/201629272.

[17] D. E. Wright *et al.*, “Machine learning for transient discovery in Pan-STARRS1 difference imaging,” *Mon. Not. R. Astron. Soc.*, vol. 449, no. 1, pp. 451–466, May 2015, doi: 10.1093/mnras/stv292.

[18] C. J. Lintott *et al.*, “Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey,” *Mon. Not. R. Astron. Soc.*, vol. 389, no. 3, pp. 1179–1189, Sep. 2008, doi: 10.1111/j.1365-2966.2008.13689.x.

[19] N. M. Ball and R. J. Brunner, “Data Mining and Machine Learning in Astronomy,” *Int. J. Mod. Phys. D*, vol. 19, no. 07, pp. 1049–1106, Jul. 2010, doi: 10.1142/S0218271810017160.

[20] Z. He, B. Qiu, A.-L. Luo, J. Shi, X. Kong, and X. Jiang, “Deep learning applications based on SDSS photometric data: detection and classification of sources,” *Mon. Not. R. Astron. Soc.*, vol. 508, no. 2, pp. 2039–2052, Dec. 2021, doi: 10.1093/mnras/stab2243.

[21] A. A. Collister and O. Lahav, “ANN *z* : Estimating Photometric Redshifts Using Artificial Neural Networks,” *Publ. Astron. Soc. Pac.*, vol. 116, no. 818, pp. 345–351, Apr. 2004, doi: 10.1086/383254.

[22] T.-Y. Cheng *et al.*, “Optimising Automatic Morphological Classification of Galaxies with Machine Learning and Deep Learning using Dark Energy Survey Imaging,” *Mon. Not. R. Astron. Soc.*, vol. 493, no. 3, pp. 4209–4228, Apr. 2020, doi: 10.1093/mnras/staa501.

[23] Y. Han, Z. Zou, N. Li, and Y. Chen, “Identifying Outliers in Astronomical Images with Unsupervised Machine Learning,” *Res. Astron. Astrophys.*, vol. 22, no. 8, p. 085006, Aug. 2022, doi: 10.1088/1674-4527/ac7386.

[24] Cosmogoblin, *Hubble Tuning Fork Diagram*. 2022. Accessed: Mar. 23, 2026. [Online]. Available: https://commons.wikimedia.org/wiki/File:Hubble_Tuning_Fork_diagram.svg

[25] D. Muthukrishna, K. S. Mandel, M. Lochner, S. Webb, and G. Narayan, “Real-time detection of anomalies in large-scale transient surveys,” *Mon. Not. R. Astron. Soc.*, vol. 517, no. 1, pp. 393–419, Nov. 2022, doi: 10.1093/mnras/stac2582.

[26] M. Crispim Romão, D. Croon, and D. Godines, “Anomaly detection to identify transients in LSST time series data,” *Mon. Not. R. Astron. Soc.*, vol. 543, no. 1, pp. 351–357, Oct. 2025, doi: 10.1093/mnras/staf1477.

[27] M. Jimenez, M. Torres Torres, R. John, and I. Triguero, “Galaxy Image Classification Based on Citizen Science Data: A Comparative Study,” *IEEE Access*, vol. 8, pp. 47232–47246, 2020, doi: 10.1109/ACCESS.2020.2978804.

[28] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, “WND-CHARM: Multi-purpose image classification using compound image transforms,” *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1684–1693, Aug. 2008, doi: 10.1016/j.patrec.2008.04.013.

[29] Y. Han, Z. Zou, N. Li, and Y. Chen, “Identifying outliers in astronomical images with unsupervised machine learning,” *Res. Astron. Astrophys.*, vol. 22, no. 8, p. 085006, Aug. 2022, doi: 10.1088/1674-4527/ac7386.

[30] B. Hoyle, M. M. Rau, K. Paech, C. Bonnett, S. Seitz, and J. Weller, “Anomaly detection for machine learning redshifts applied to SDSS galaxies,” *Mon. Not. R. Astron. Soc.*, vol. 452, no. 4, pp. 4183–4194, Oct. 2015, doi: 10.1093/mnras/stv1551.

[31] T. L. Killestein *et al.*, “Transient-optimized real-bogus classification with Bayesian convolutional neural networks – sifting the GOTO candidate stream,” *Mon. Not. R. Astron. Soc.*, vol. 503, no. 4, pp. 4838–4854, Jun. 2021, doi: 10.1093/mnras/stab633.

[32] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 2015, doi: 10.48550/arXiv.1409.1556.

Liitteet

Liite 1. Tähtitieteen survey-tutkimuksia ja niiden tuottamia datamääriä

Survey-tutkimus	Lyhyt kuvaus tutkimuksesta ja tutkimustehtävästä	Kokonais-datamäärä [2]
Palomar Digital Sky Survey (DPOSS)	Digitoitu versio POSS-II pohjoisen taivaan kartoitustutkimuksesta.	3 TB
Two Micron All-Sky Survey (2MASS)	Koko taivaan kartoitus infrapuna-alueella. Toteutettu kahdella teleskoopilla Arizonassa ja Chilessä. 1997–2001	10 TB
Galaxy Evolution Explorer (GALEX)	Ultraviolettialueen tutkimus, keskittyen galaksien UV-ominaisuuksiin. Avaruusteleskooppi, NASA 2003–13.	30 TB
Sloan Digital Sky Survey (SDSS)	Koko taivaan optinen ja infrapuna-alueen spektroskopinen kartoitus. 2000–jatkuu	40 TB
SkyMapper Southern Sky Survey (SMSS)	Eteläisen taivaan laaja-alainen monivärikartoitus, Australian National University 2014–jatkuu	500 TB
Panoramic Survey Telescope and Rapid Response System (PanSTARRS)	Havaijilla sijaitseva havaintojärjestelmä, tehtävänä jatkuva transienttien kohteiden etsintä, sekä tunnettujen kohteiden astrometria (paikannus) ja fotometria (kirkkausmittaus). PS1 2006-14; PS2 2014–jatkuu	~40 PB Odotettavissa
Legacy Survey of Space and Time (LSST)	Vera C. Rubin observatorion eteläisen taivaan optinen kartoitus; tarkoitus muodostaa vuosikymmenen mittainen aikasarja. Aloittamassa lähiaikoina.	~200PB Odotettavissa
Square Kilometer Array (SKA)	Neliökilometrin kokoinen radioteleskooppiryhmä, rakenteilla Australiaan. Tarkoituksena kartoittaa miljardi galaksia havaittavan universumin reunoilla, galaksikehityksen ja historian tutkiminen.	~4.6EB Odotettavissa