# Koneoppimisen validaatio radiologisessa kuvantamisessa

**Jussi Tohka**

**A.I Virtanen Instituutti, UEF**

**Jussi.tohka@uef.fi**

KYS

7.2.2020

Programme for Sustainable Growth and Jobs

Leverage from the EU 2014–2020

European Union
European Social Fund

UNIVERSITY OF EASTERN FINLAND

Elinkeino-, liikenne- ja ympäristökeskus

# Sisältö

- KUBIAC hanke
- Koneoppiminen radiologisessa kuvantamisessa
- Koneoppimismentelmien validaatio
  - Mittarit
  - Kuinka laskea mittarit rajallisesta datasta

# KUBIAC hanke

- Innovaatiopotentiaalin kasvattaminen kuva-analyysiosaamista kehittämällä - kuva-analyysikeskus KUBIAC (ESR S21770)
- Tekoäly-pohjaisten kuva-analyysimenetelmien luomat mahdollisuudet, ja toisaalta niiden haasteet.
- Kuva-analyysimenetelmien kehittämisen huippuosaajien tarve on kasvanut voimakkaasti.
- Kehittäjien ja hyödyntäjien kumppanuudesta syntyisi uudenlaiseen osaamiseen pohjautuvaa liiketoimintaa, ja tämän avulla voitaisiin paremmin ennakoida muuttuvia osaamistarpeita.

# KUBIAC hanke

Pilotti 3: Tekoälyyn pohjautuva aneurysman detektio magneettiresonanssiangiografia (MRA)-kuvista yhteistyössä KYS:n kuvantamiskeskuksen kanssa

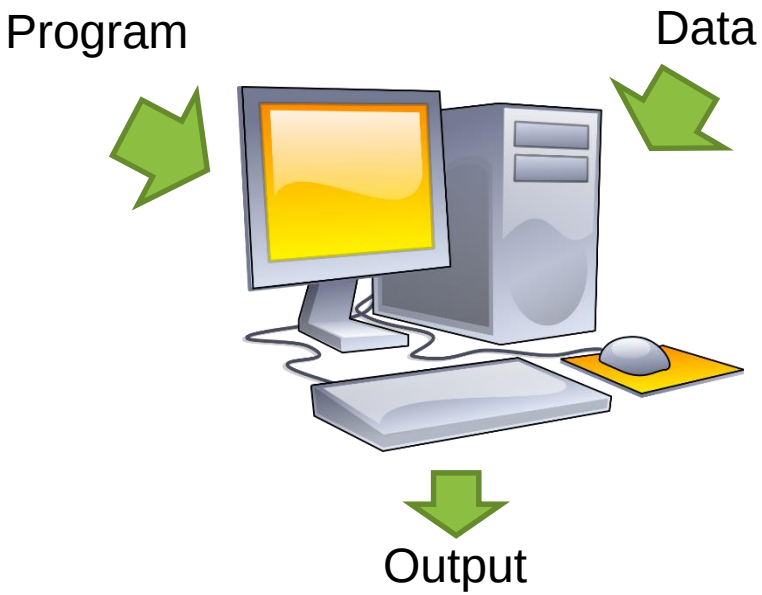
www.uef.fi/kubiac

https://blogs.uef.fi/kubiac/

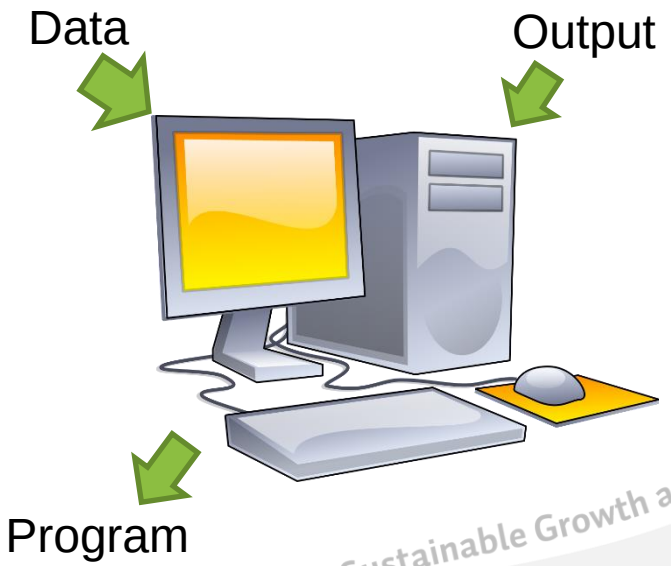
# Koneoppiminen

- Koneoppimisalgoritmi rakentaa matemaattisen mallin näytedataan perustuen (opetusdata), jonka avulla se voi tehdä ennusteita näytedatan ulkopuolisesta datasta

Perinteinen ohjelmointi




Koneoppiminen



# Bayes-luokitin

- Optimaalinen luokitin kun ongelman tilastolliset ominaisuudet tunnetaan
- Kaikki käytännön koneoppimisalgoritmit approksimoivat Bayes-luokitinta
- Piirrevektorille X valitaan luokka jonka posterior-todenäköisyys on suurin kun havaitaan X
- Posterior todennäköisyys lasketaan luokan prioritodennäköisyyden ja luokan jakauman X:ssä tulona

# Esimerkki: miksi priori-todennäköisyydet ovat tärkeitä

- Disclaimer: prosentit esimerkissä eivät varmastikaan pidä kutiansa
- 1% naisista jotka osallistuvat rintasyöpäseulontaan on rintasyöpä; 99% ei ole rintasyöpää
- 80% seurantatutkimuksista löytää olemassa olevan rintasyövän
- 9.6% tutkimuksista löytää rintasyövän vaikkei sitä olisi
- Tutkittavan tulos on positiivinen. Millä todennäköisyydellä hänellä on rintasyöpä?

# Esimerkki: miksi priori-todennäköisyydet ovat tärkeitä

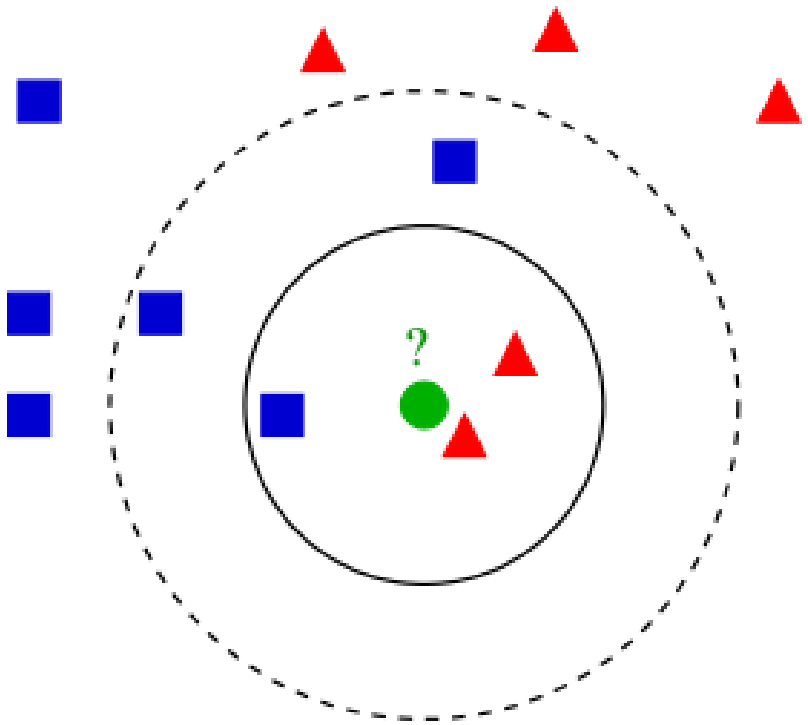
- 1% on rintasyöpä; 99% ei ole rintasyöpää
- 80% seurantatutkimuksista löytää rintasyövän
- 9.6% tutkimuksista löytää rintasyövän vaikkei sitä olisi
- Tutkittavan tulos on positiivinen. Millä todennäköisyydellä hänellä on rintasyöpä?
- $P(S+|Test+) = P(Test+|S+)P(S+)/P(Test+)$

$= 0.8*0.01/0.1026 = 7.8\%$

$P(Test+) = P(Test+|S-)P(S-) + P(Test+|S+)P(S+)$

$= 0.096*0.99 + 0.8*0.01 = 0.1026$

# Lähimmän naapurin luokitin




- Antti Ajanki AnAj [CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)]

# Koneoppiminen radiologisessa kuvantamisessa

- Diagnoosin avustaminen
- Ennusteiden tekeminen (aikainen diagnoosi, prognoosi)
- Kvantitaatio
- Kuvan laadun parantaminen
- https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

# Koneoppimisalgoritmin hyvyyden mittaaminen

- Mitä mitataan?
- Miten mitataan?
- Mitä varten mitataan?
- Aiheesta kirjoitettu paljon viime aikoina
- Tulossa: Tohka, van Gils: ” Evaluation of machine learning algorithms for Health and Wellness applications: a tutorial”

# Koneoppimisalgoritmien hyvyyden mittaaminen

**Key Considerations for Authors, Reviewers, and Readers of AI/ML Manuscripts in *Radiology***

| Key Considerations |
|---|
| Are all three image sets (training, validation, and test sets) defined? |
| Is an *external* test set used for final statistical reporting? |
| Have multivendor images been used to evaluate the AI algorithm? |
| Are the sizes of the training, validation, and test sets justified? |
| Was the AI algorithm trained using a standard of reference that is widely accepted in our field? |
| Was preparation of images for the AI algorithm adequately described? |
| Were the results of the AI algorithm compared with radiology experts and/or pathology? |
| Was the manner in which the AI algorithm makes decisions demonstrated? |
| Is the AI algorithm publicly available? |

Note.—AI = artificial intelligence, ML = machine learning.

D. A. Bluemke et al, Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers; from the radiology editorial board, 2019

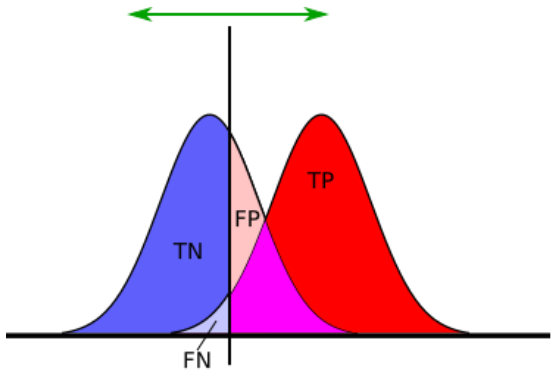
# Luokitustuloksen mittaaminen: mitä mitataan

- Luokitustarkkuus (accuracy)
- Sensitiivisyys, spesifisyys
- ”Balanced accuracy”
- ROC-käyrä ja AUC
- Sekaannusmatriisi

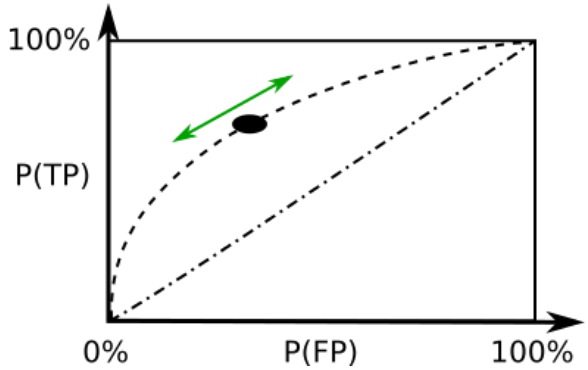
# Sekaannusmatriisiesimerkki: Wikipedia

| | | Patients with bowel cancer (as confirmed on endoscopy) | | |
|---|---|---|---|---|
| | | Condition positive | Condition negative | |
| **Fecal occult blood screen test outcome** | Test outcome positive | **True positive** (TP) = 20 | **False positive** (FP) = 180 | Positive predictive value (PPV) = TP / (TP + FP) = 20 / (20 + 180) = **10%** |
| | Test outcome negative | **False negative** (FN) = 10 | **True negative** (TN) = 1820 | Negative predictive value (NPV) = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%** |
| | | **Sensitivity** = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%** | **Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%** | |

# ROC-kayrä




| | |
|---|---|
| TP | FP |
| FN | TN |




By Sharpr - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=44059691

**ROC-käyrä**:
X-koordinaatti:
1 – Specificity
Y-koordinaatti:
Sensitivity
**AUC:**
Area under the curve
From 0 to 1
0.5 chance level
Reference: Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction.

# Segmentaatiotarkkuden mittaaminen

- Yleensä Dice-kerroin manuaalisen ja automaattisen segmentaation välillä


Figure credit: Miguel Valverde

Dice = $2|X \cap Y|/(|X| + |Y|)$
Välillä [0,1]
Mikä on hyvä arvo?
Zijdenbos et al, IEEE-TMI 1994
Dice ei kerro kaikkea

# Kuinka arvioidaan algoritmeja?

- Eri tapauksissa eri vaatimukset
- Tärkeää: opetusdataa EI saa käyttää testaukseen
- Jo opetetun koneoppimismallin testaus kliiniseen käyttöön: -> Erillinen testijoukko
- Uuden sovelluksen potentiaalin testaus ja validaatio -> ristiinvalidointi (tai bootstrap). Huomaa, jos vain yksi datasetti ristiinvalidointi aina parempi kuin holdout
- Sekä ristiinvalidointi että holdout: Testi ja opetusjoukot riippumattomia (kaikki koehenkilö A:n näytteet aina joko testi tai opetusjoukossa)

# Ristiinvalidointi

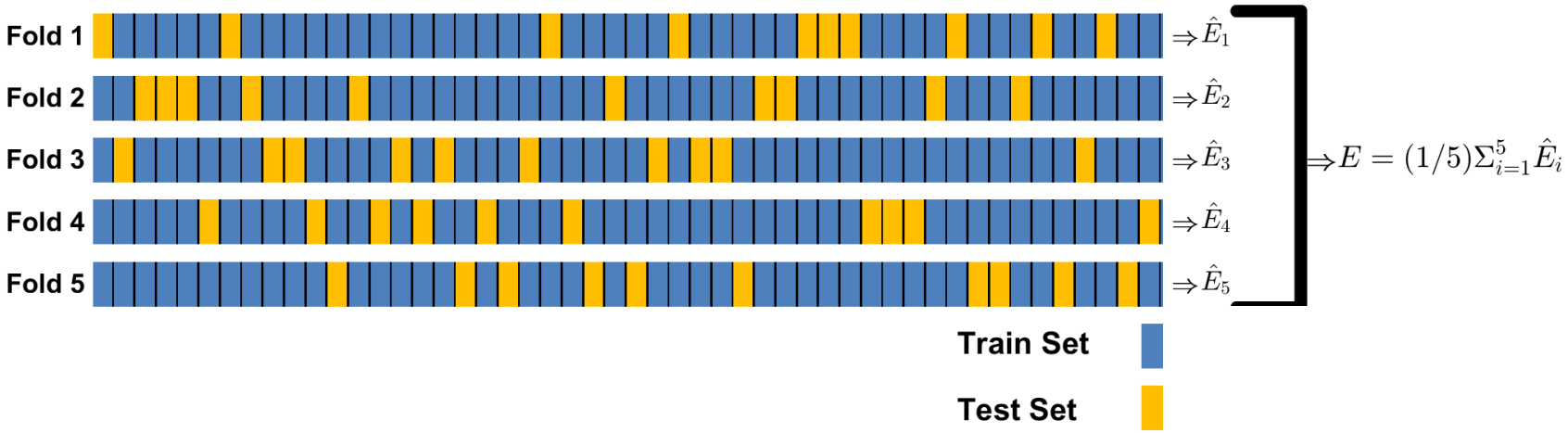
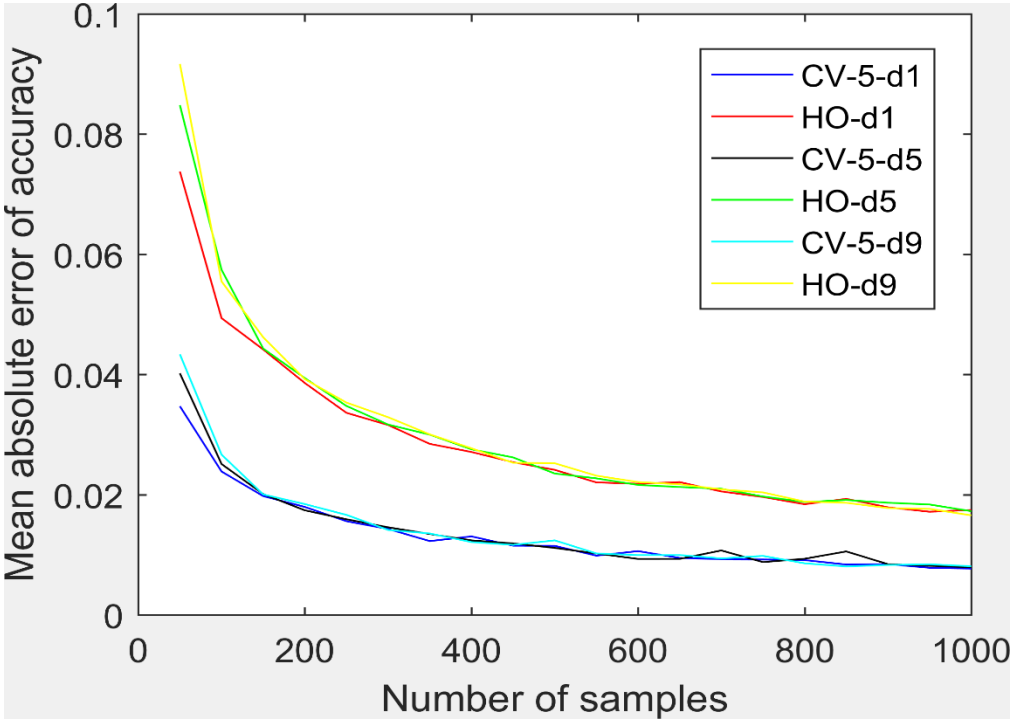



Figure credit: Vandad Imani

# Ristiinvalidointi vs. holdout

- Jos pelkästään tietty opetusjoukko ja halutaan estimaatti koneoppimisalgoritmin performanssille, ristiinvalidointi




Bayes-error fixed at 10%

# Muita tapoja

- Toistettu ristiinvalidointi
- Bootstrap (Random Forests)

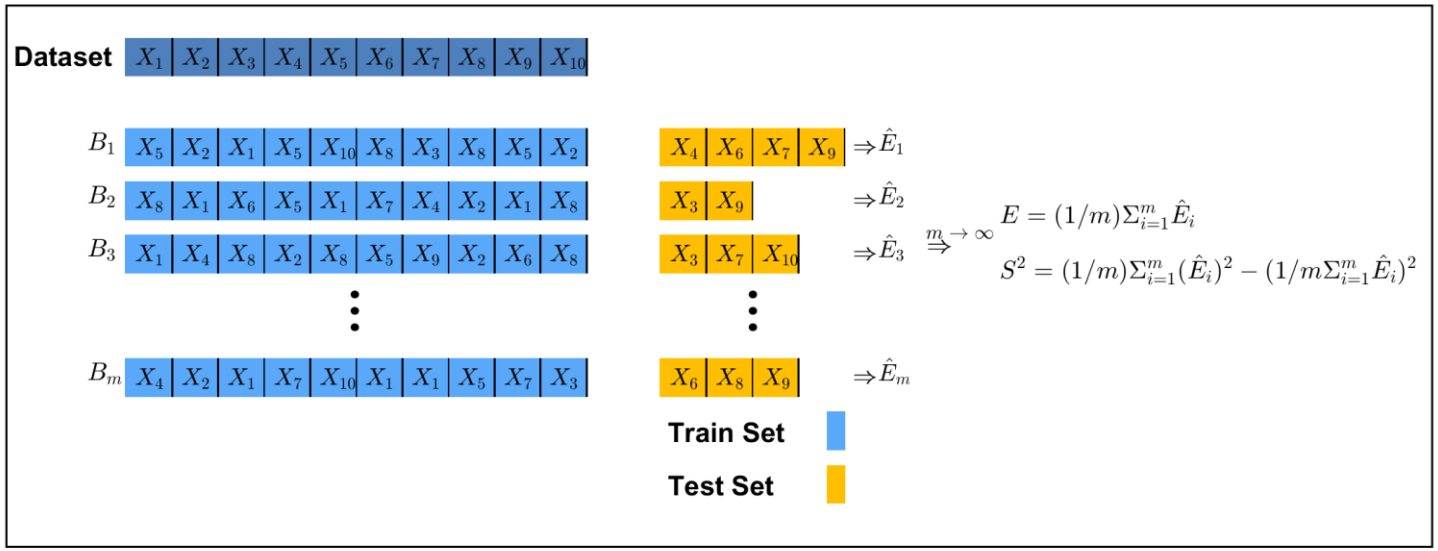



Figure credit: Vandad Imani

# Algoritmien vertailu

- Suositeltu testi: Corrected resampled t-test statistic (Nadeau, Bengio 2003, Bouckaert,Frank 2004)
- Huomaa että toistetun CV:n eri toistot EIVÄT ole riippumattomia (tyypillinen virhe aivokuvantamispapereissa)
- Parempi kuin ”chance-level” -> permutaatiotesti (lue kuitenkin Ojala 2010 ennen käyttöä)

# Tyypillisiä virheitä ja muuta huomattavaa

- Muuttujien valinta käyttäen koko dataa (ei vain opetusjoukkoa, ks. Huttunen, Manninen, Tohka, FCSE 2012)
- Parametrien valinta testijoukon tuloksia käyttäen
- Ristiinvalidoidun virhe-estimaatin varianssi on huomattava, jos näytemäärä pieni (ks. Edward Doughertyn tuotanto asiasta)
- Jos näytemäärä pieni, parametriset mallit hyödyllisiä

# Koneoppimisalgoritmien hyvyyden mittaaminen

**Key Considerations for Authors, Reviewers, and Readers of AI/ML Manuscripts in *Radiology***

| Key Considerations |
|---|
| Are all three image sets (training, validation, and test sets) defined? |
| Is an *external* test set used for final statistical reporting? |
| Have multivendor images been used to evaluate the AI algorithm? |
| Are the sizes of the training, validation, and test sets justified? |
| Was the AI algorithm trained using a standard of reference that is widely accepted in our field? |
| Was preparation of images for the AI algorithm adequately described? |
| Were the results of the AI algorithm compared with radiology experts and/or pathology? |
| Was the manner in which the AI algorithm makes decisions demonstrated? |
| Is the AI algorithm publicly available? |

Note.—AI = artificial intelligence, ML = machine learning.

D. A. Bluemke et al, Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers; from the radiology editorial board, 2019