Optimering av maskinlæring for økt bærekraft

Hvem er jeg?

- Adela Nedisan Videsjorden
- Bachelor i Språkteknologi (UiO)
- Mastergrad i Programmering og systemarkitektur (UiO)
- I dag: Master of Science (SINTEF – Trustworthy Green IoT)

Dagens tema:

- Veien til bærekraftig maskinlæring

SINTEF

UiO : Universitetet i Oslo

Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.

Industri 4.0

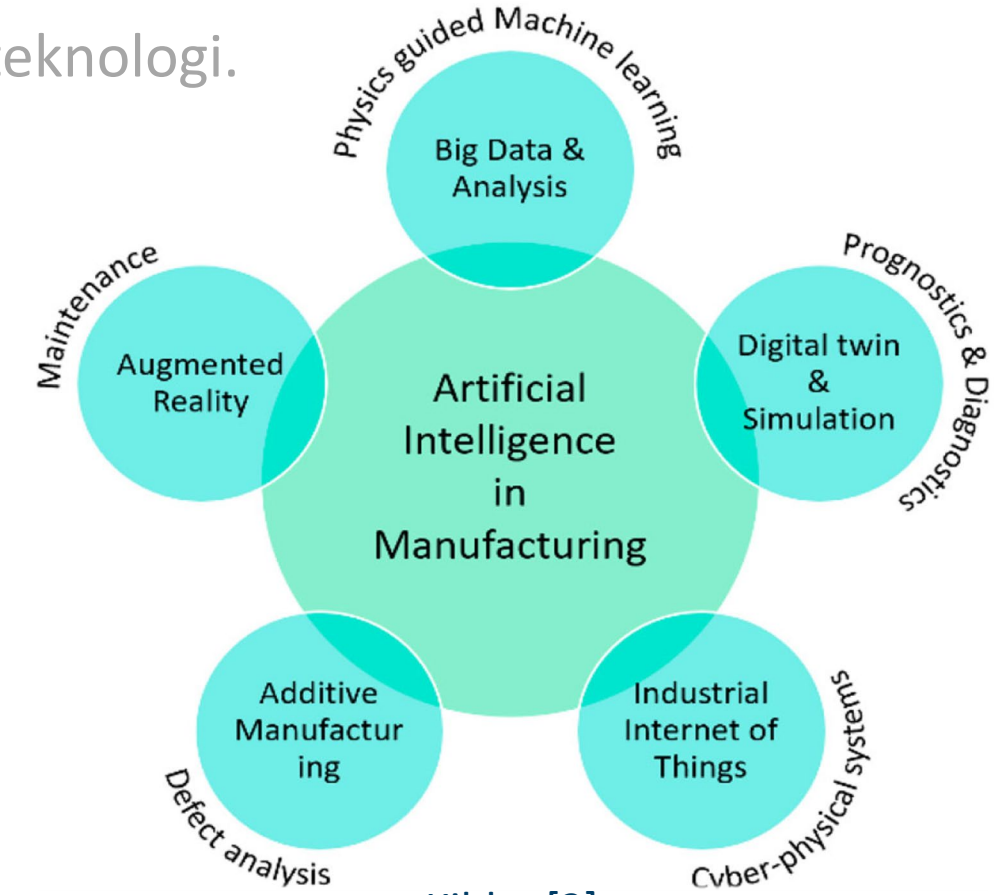
- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.
- **Manufacturing** (produksjon) er spesielt preget:

"The area of greatest impact by I4.0 is manufacturing"

Kilde: [1]

KI er en hovedingrediens som mulliggjør I4.0!

KI-bruk innen produksjon:



Kilde: [2]

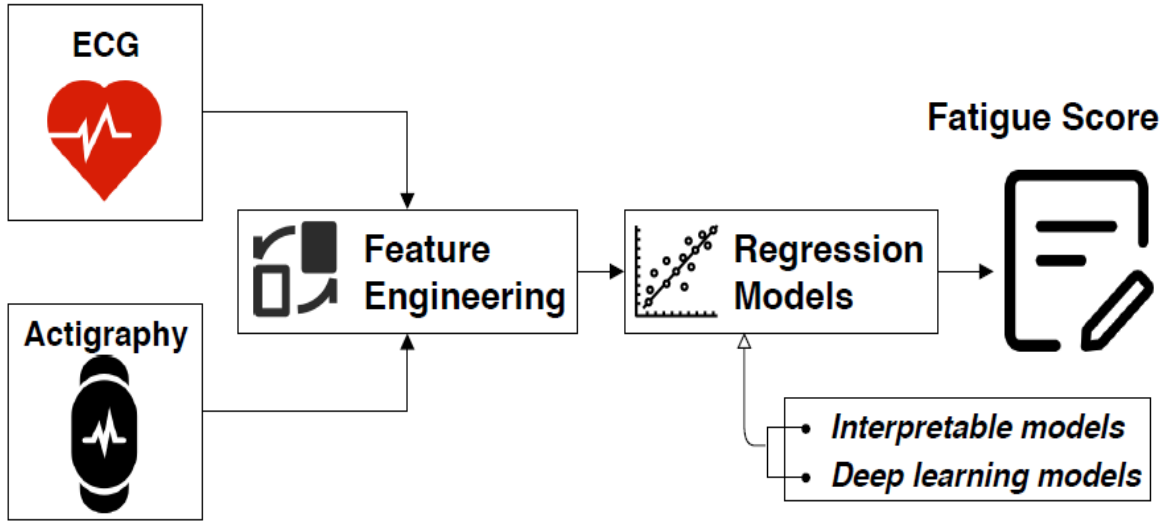
Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.
- **Manufacturing** (produksjon) er spesielt preget.
- Bruk av KI innen **helsesektoren**:

Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.
- **Manufacturing** (produksjon) er spesielt preget.
- Bruk av KI innen **helsesektoren**:

- Støtter spesialister i beslutningsprosessen [3]
- Legemiddelutvikling [4]
- Deteksjon av tretthet, stress og depresjon [5,6,7]



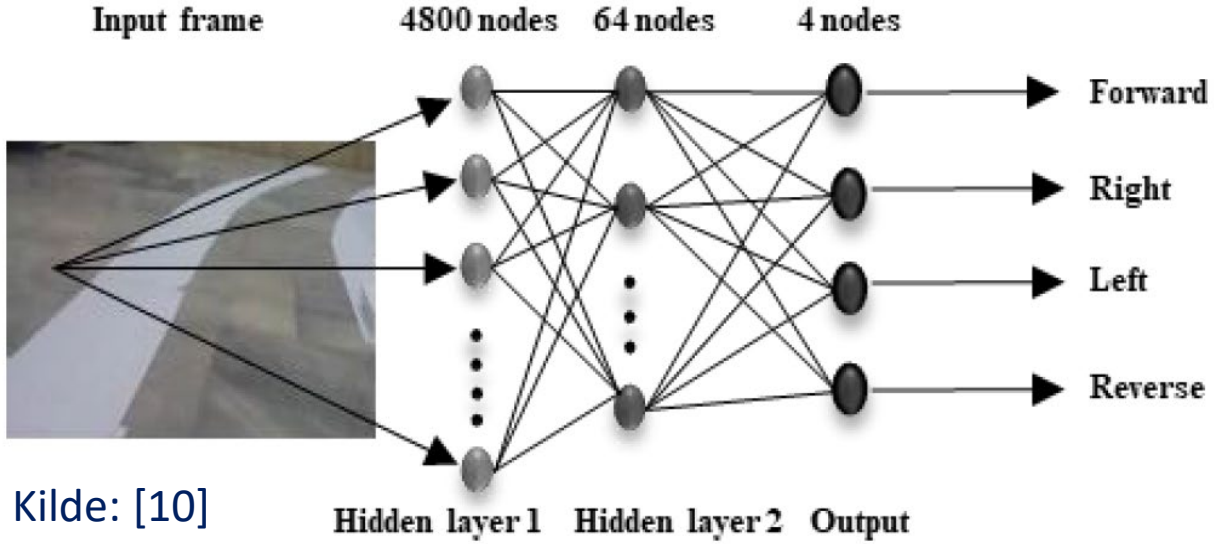
Kilde: [5]

Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.
- **Manufacturing** (produksjon) er spesielt preget.
- Bruk av KI innen **helsesektoren** og **logistikk**:

Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologi.
- **Manufacturing** (produksjon) er spesielt preget.
- Bruk av KI innen **helsesektoren** og **logistikk**:
- Ifølge [8], KI har følgende fordeler:
  - ✓ **Raskere leveringshastighet**
  - ✓ **Høyere pålitelighet**
  - ✓ **Lavere driftskostnader**
  - ✓ **Forbedret driftseffektivitet**



Kilde: [10]

- Eksempler: pakkelevering via droner [9], selvkjørende biler [10]

Industri 4.0

- Økning i antall **digitale løsninger** innenfor ulike industriområder.
- Økning i bruk av **sensorer** og **kunstig intelligens** (KI) teknologier.
- **Manufacturing** (produksjon) er spesielt preget.
- Bruk av KI innen **helsesektoren** og **logistikk**.

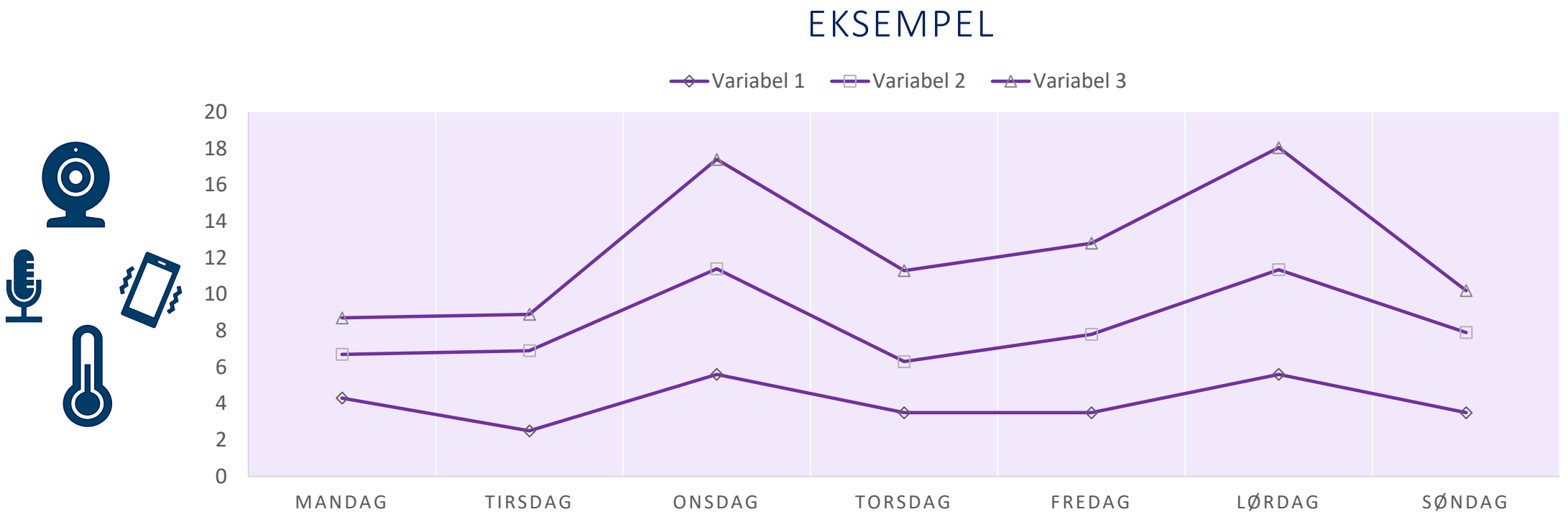
- **Og så videre …**

Tidsseriedata (eng: Time-series)

- En **sensor** måler variabler ved bestemte tidsintervaller

Tidsseriedata (eng: Time-series)

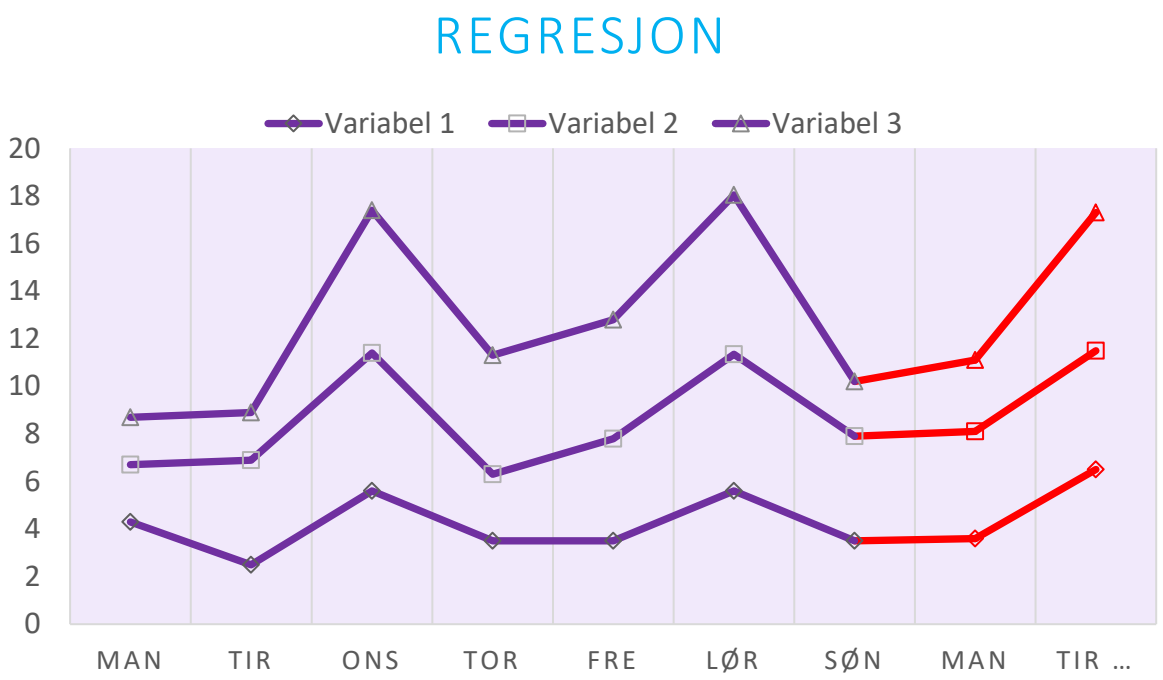
- En **sensor** måler variabler ved bestemte tidsintervaller → tidsseriedata



Illustrasjon: sensor som måler 3 ulike variabler en gang om dagen over en periode på en uke.

Tidsseriedata (eng: Time-series)

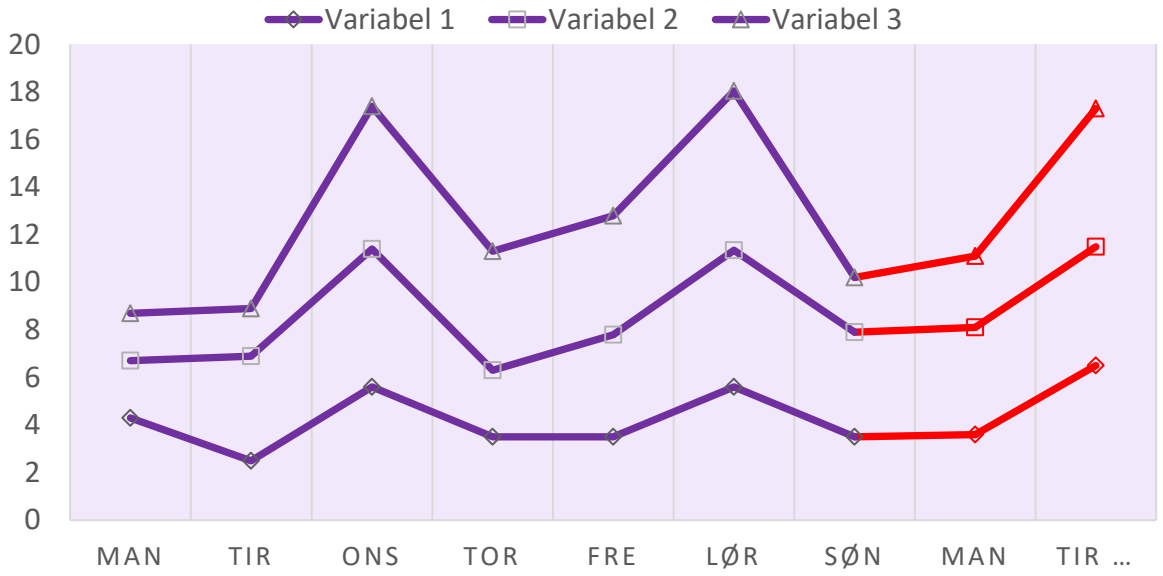
- En **sensor** måler variabler ved bestemte tidsintervaller → tidsseriedata
- **Maskinlæringsmodeller** (ML) trent på disse dataene kan brukes til **å predikere fremtidige hendelser**



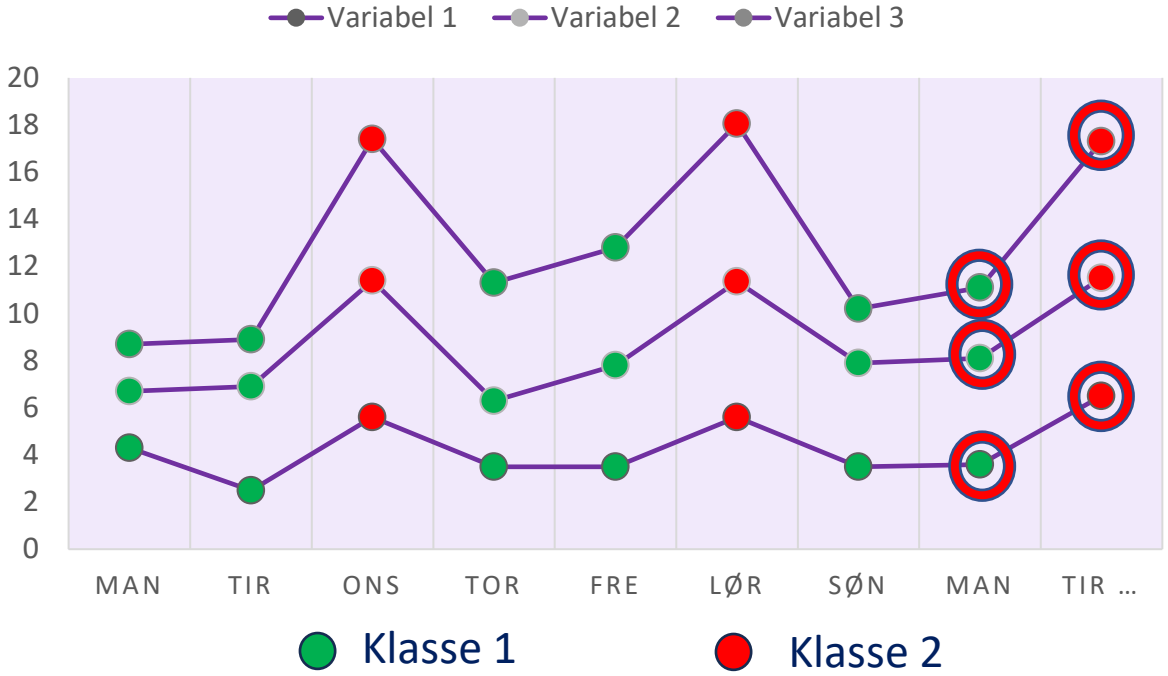
Tidsseriedata (eng: Time-series)

- En **sensor** måler variabler ved bestemte tidsintervaller → tidsseriedata
- **Maskinlæringsmodeller** (ML) trent på disse dataene kan brukes til **å predikere fremtidige hendelser** eller **klassifisere hendelser**.

REGRESJON



KLASSIFIKASJON



Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv.

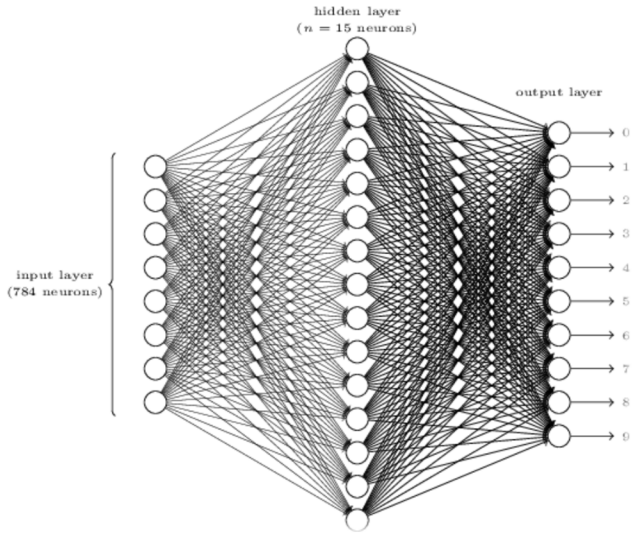
Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

Dyplæring (eng: Deep Learning)

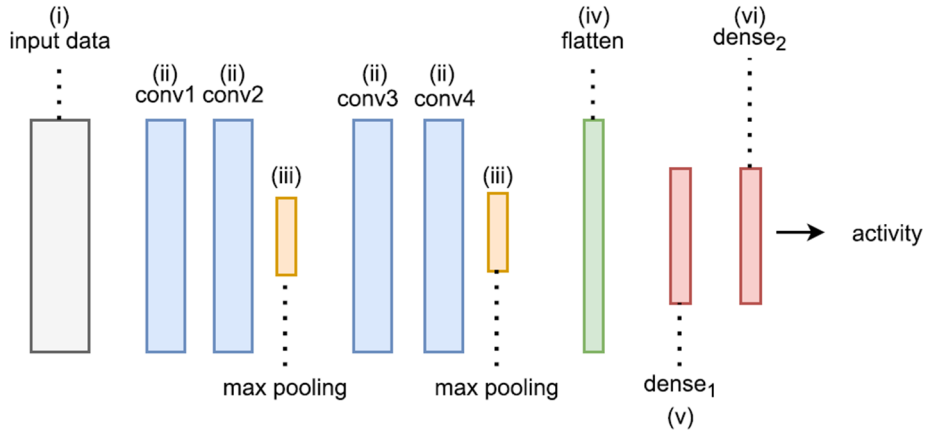
- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

Multi-Layer Perceptron (MLP)



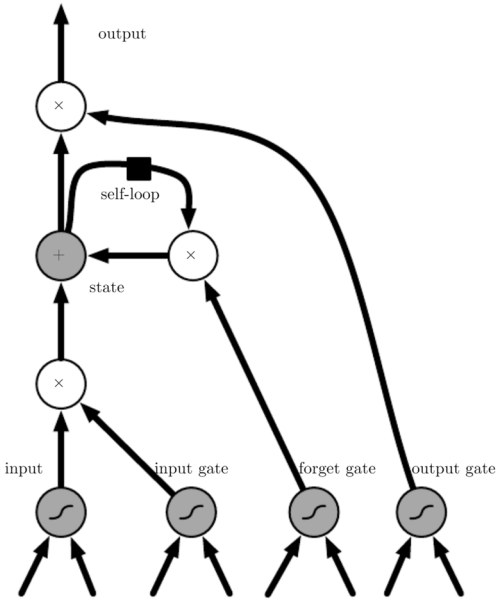
Kilde: [11]

1D Convolutional Neural Network (CNN)



Kilde: [12]

Long Short-Term Memory (LSTM) – celle:



Kilde: [13]

Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

Skytjenester (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14].

Dyplæring (eng: Deep Learning)

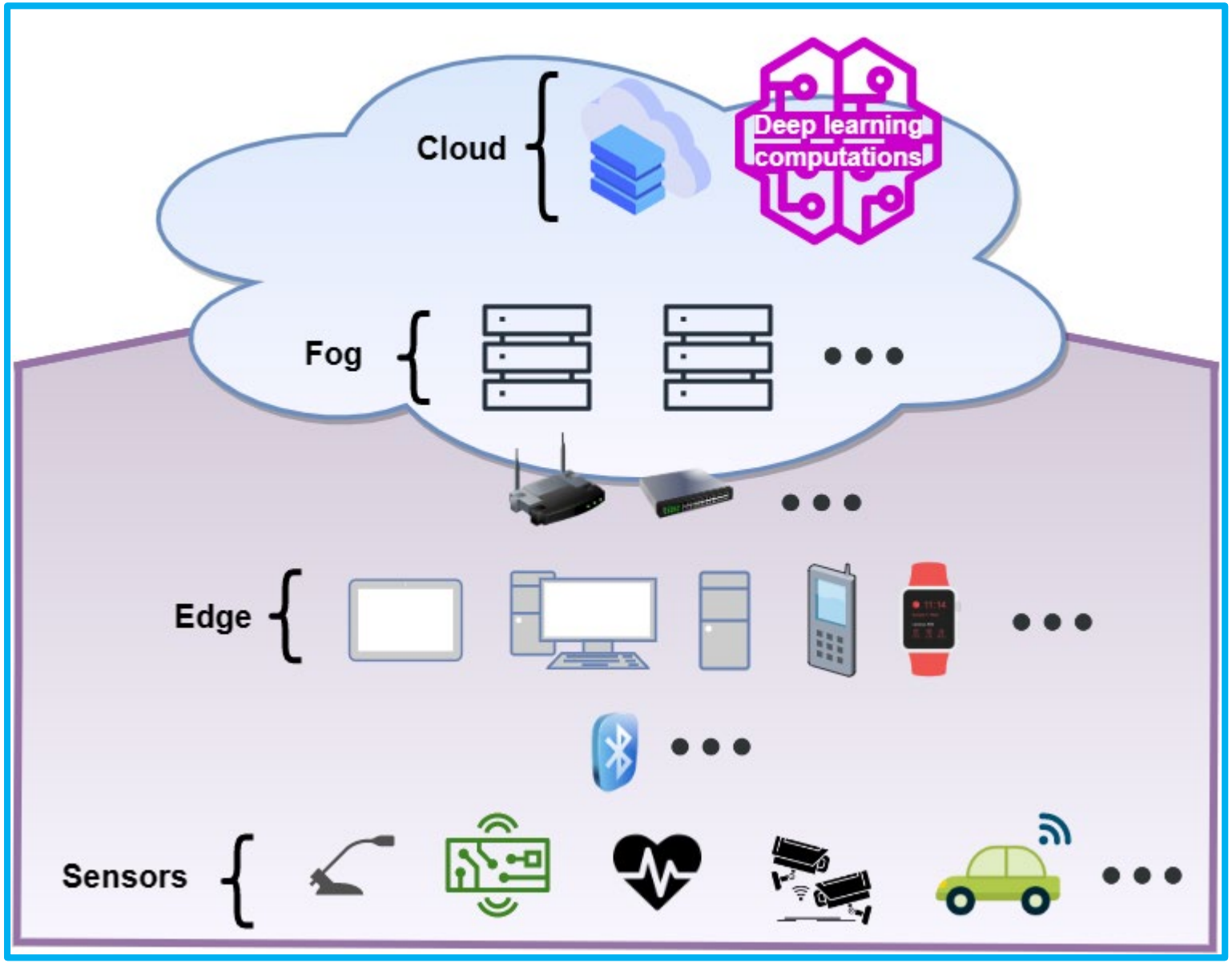
- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!
- **Skytjenester** (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14]. **MEN!** Dette er ikke alltid gjennomførbart pga begrenset nettverksbåndbredde eller av personvernshensyn [15,16].

Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

  Skytjenester (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14]. **MEN!** Dette er ikke alltid gjennomførbart pga begrenset nettverksbåndbredde eller av personvernshensyn [15,16].
- **Løsning**: flytt tunge beregninger nærmere **edgen**! [16]

Fra skytjeneste til edge



Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

  Skytjenester (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14]. **MEN!** Dette er ikke alltid gjennomførbart pga begrenset nettverksbåndbredde eller av personvernshensyn [15,16].

- **Løsning**: flytt tunge beregninger nærmere **edgen**! [16]

  - ✓ Behandler data nærmere kilden
  - ✓ Reduserer miljøpåvirkningen knyttet til store datasentre

Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

  Skytjenester (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14]. **MEN!** Dette er ikke alltid gjennomførbart pga begrenset nettverksbåndbredde eller av personvernshensyn [15,16].
- **Løsning**: flytt tunge beregninger nærmere **edgen**! [16]
- **MEN!** Utfordringer på grunn av begrensede enhetsressurser: beregnings- og minnekrav, strøm- og energiforbruk.

Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk **arkitekturer** som kan behandle tidsseriedata!

  Skytjenester (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurselastisitet** [14]. **MEN!** Dette er ikke alltid gjennomførbart pga begrenset nettverksbåndbredde eller av personvernshensyn [15,16].
- **Løsning**: flytt tunge beregninger nærmere **edgen**! [16]
- **MEN!** Utfordringer på grunn av begrensede enhetsressurser: beregnings- og minnekrav, strøm- og energiforbruk.
- **Optimaliseringer** blir derfor anvendt både på **maskinvare-** og **programvarenivå** [17].

Dyplæring (eng: Deep Learning)

- Populært valg: ikke nødvendig å manuelt finne viktige trekk i data, sterk ikke-lineær læringsevne, osv. Flere nettverk-arkitekturer som kan behandle tidsseriedata!
- **Skytjenester** (eng: Cloud computing) muliggjør **trening** og **inferens** av dyp læring-modeller ved å tilby **ressurs**[illegible] ikke alltid gjennomførbart pga begrenset nettverksbåndbredde[illegible]shopsvp [15,16].
- **Løsning**: flytt tunge beregninger nærmere [illegible] [16]
- **MEN!** Utfordringer på grunn av [illegible]ssurser: beregnings- og minnekrav, strøm- og energiforbruk.
- **Optimaliseringer** blir derfor anvendt både på **maskinvare-** og **programvarenivå** [17].

«En-passer-alle»
optimalisering
på tvers av nettverk
arkitektur?

Modell optimisering

- **Hvorfor?**
- Eliminere unødvendig optimaliseringsinnsats for modeller som håndterer sensor data.

Modell optimisering

- **Hvorfor?**
- Eliminere unødvendig optimaliseringsinnsats for modeller som håndterer sensor data.
- **Hvordan?**
- Finn minimalt sett med **optimaliseringsteknikker** som oppfyller følgende to kriterier:
  - ✓ Redusere ressursutnyttelsen.
  - ✓ Bevare nøyaktigheten.

Modell optimisering

- **Hvorfor?**
- Eliminere unødvendig optimaliseringsinnsats for modeller som håndterer sensor data.
- **Hvordan?**
- Finn minimalt sett med **optimaliseringsteknikker** som oppfyller følgende to kriterier:

  ✓ Redusere ressursutnyttelsen.

  ✓ Bevare nøyaktigheten.

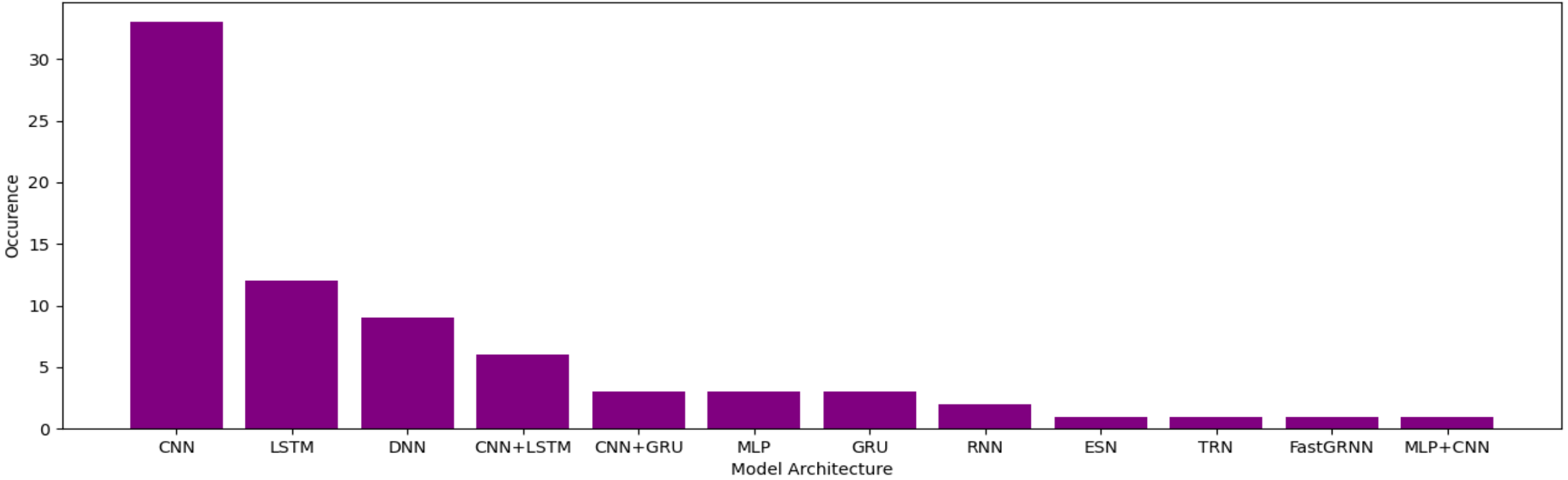
- Vi bruker **tidligere forskningsbidrag** til å designe et **rammeverk.** Vi søker etter potensielle **optimaliseringer**, men også **arkitekturer**, **enheter** og måter å **måle ressursbruk**.

Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.

Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.



Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.

Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.
- **Målinger**: det finnes felles indikatorer for ressursbruk/prediksjonsytelse.

ID	Name	Description
KPI_1	Precision / Macro average precision	Correctness of the proportion of identifications in a model
KPI_2	Recall / Macro average recall	Correctness of the proportion of actual positives being correctly identified by the model
KPI_3	Harmonic Precision-Recall Mean (F1) / Macro average F1	Combines precision and recall
KPI_4	Inference Latency	Measure speed during one complete inference request
KPI_5	Parameters	Estimates memory requirements and model complexity
KPI_6	Compressed file size	Estimates necessary memory requirements for a compressed model.
KPI_7	Non-volatile memory	Portion of space allocated for non-volatile storage of model
KPI_8	CPU usage	Estimates system-wide CPU utilization as a percentage
KPI_9	Power consumption	Quantifies the power needed to perform a complete inference request
KPI_{10}	Energy consumption	Quantifies the energy needed to perform a complete inference request

Nøyaktighet

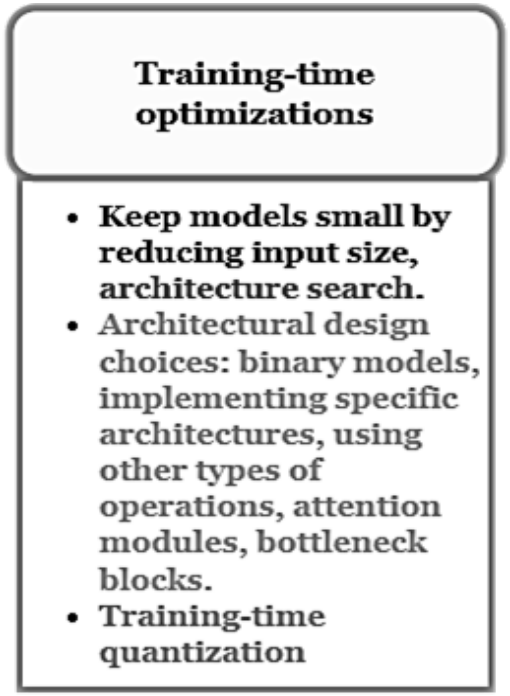
Ressursbruk

Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.
- **Målinger**: det finnes felles indikatorer for ressursbruk/prediksjonsytelse.
- **Optimiseringer**: kan fordeles i tre kategorier:
- 1) Maskinvare og distribusjons optimiseringer.

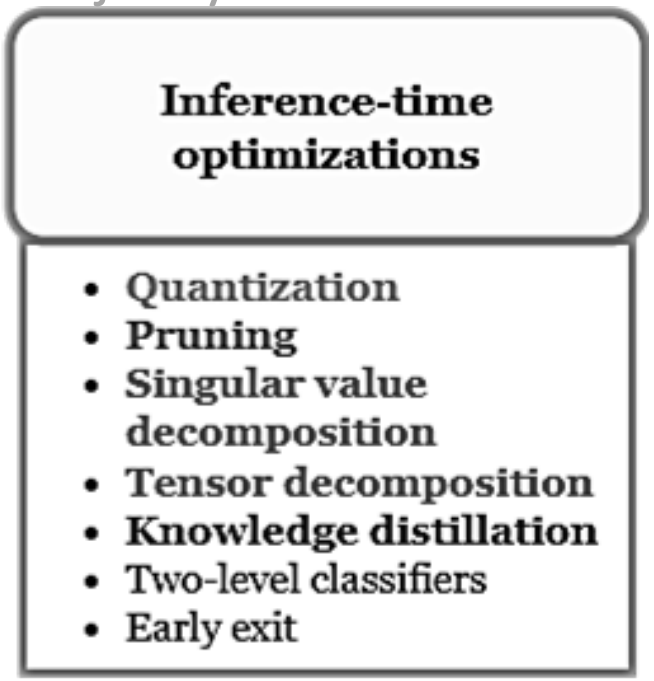
Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.
- **Målinger**: det finnes felles indikatorer for ressursbruk/prediksjonsytelse.
- **Optimiseringer**: kan fordeles i tre kategorier:
- 1) Maskinvare og distribusjons optimiseringer.
- 2) Optimisering før og mens modellen trener.



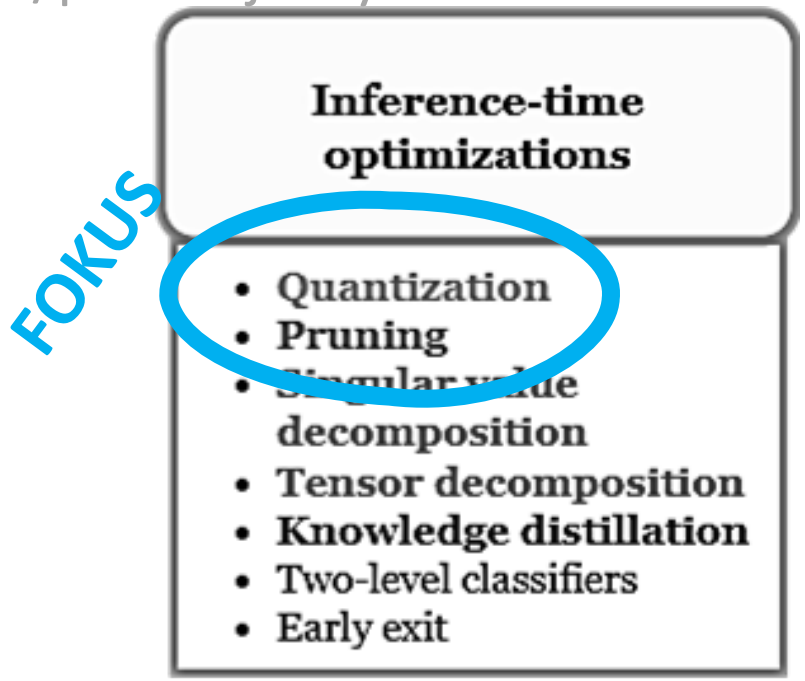
Kort om funn rundt tidsserieanalyse

- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.
- **Målinger**: det finnes felles indikatorer for ressursbruk/prediksjonsytelse.
- **Optimiseringer**: kan fordeles i tre kategorier:
- 1) Maskinvare og distribusjons optimiseringer.
- 2) Optimisering før og mens modellen trener.
- 3) Optimisering etter at modellen er trent.

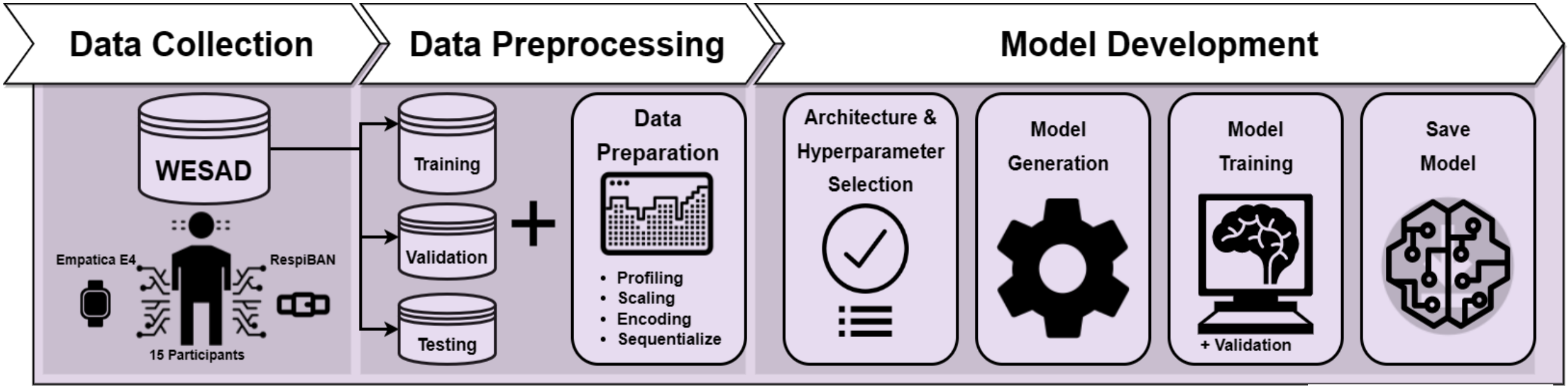


Kort om funn rundt tidsserieanalyse

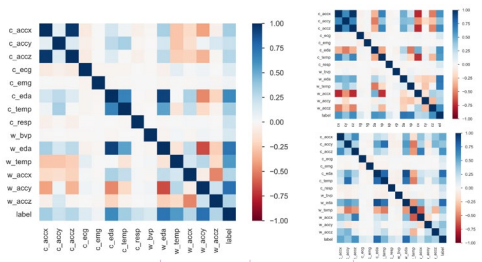
- **Formål**: helseapplikasjoner utgjør det mest utbredte bruksområdet.
- **Arkitektur**: de mest populære modellarkiteturene er CNN, LSTM, MLP.
- **Enhet**: mange eksperimenter kjører på Raspberry Pi eller lignende enheter.
- **Målinger**: det finnes felles indikatorer for ressursbruk/prediksjonsytelse.
- **Optimiseringer**: kan fordeles i tre kategorier:
- 1) Maskinvare og distribusjons optimiseringer.
- 2) Optimisering før og mens modellen trener.
- 3) Optimisering etter at modellen er trent.

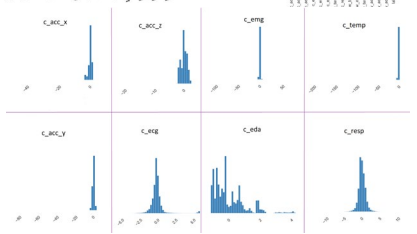


Eksperiment oppsett (I)



	Number Participants	Collection Methods	Dataset Size	Labeled / unlabeled	Problem Types	Measurements / Observations	Data Type	Data Format
Empatica	EEG: 14 participants E4: 11 participants	8-channel OpenBCI headset Empatica E4 wristband	86.06 MB	Labeled by FAS scores (fatigue state score)	Fatigue detection	Electroencephalographic (EEG) signals Photoplethysmography (PPG) HR: Heart rate IBI: Inter-beat interval TEMP: Skin temperature EDA: Electrodermal Activity BVP: Blood Volume Pulse ACC: Accelerometer data SRates: Sampling Rates of the ACC, BVP, EDA, HR, IBI and TEMP	Time-series	.zip (included .mat) and matlab file
Toadstool	10 participants (5 female, 5 male)	Empatica E4 wristband Camera:a 1.3-MP webcam of a Samsung Series 9 Notebook NP900X4C	7.5 GB	Unlabeled	Facial expression studies Sentiment analysis Game studies Reinforcement learning Emotionally-aware ML	Controller input Video at 30 frames per second with a resolution of 640 × 480. HR: Heart rate IBI: Inter-beat interval TEMP: Skin temperature EDA: Electrodermal Activity BVP: Blood Volume Pulse ACC: Accelerometer data	Video files Time-series	.avi, JSON, .csv, .text
WESAD	15 participants (12 male, 3 female)	Chest-worn device (RespiBAN) Empatica E4 wristband.	Instances: 63000000 2.1 GB	Labeled by self-reports on emotional states (PANAS, STAI, SAM, SSSQ approaches: fill-in questions)	Stress and affect detection Classification Regression Affective computing	Electrocardiogram (ECG) Electrodermal activity (EDA) Electromyogram (EMG) Respiration Body temperature Three-axis acceleration BVP: Blood Volume Pulse EDA: Electrodermal Activity TEMP: Skin temperature ACC: Accelerometer data	Multivariate Time-series	
GNOMON	15 participants (all female)	Fitbit Charge 5		Labeled by self-reports (fatigue score)	Fatigue and stress detection	Stress Score Physical Activity Menstrual Health Metrics Fitness fatigue component HR: Heart rate Sleep Score Others (i.e. blood glucose)	Time-series	.csv, JSON





Size	Parameter range	MLP	CNN	LSTM
1	50,000-100,000	58,805	61,601	73,345
2	200,000-300,000	235,217	277,761	235,713
3	400,000-500,000	470,561	497,921	474,753
4	600,000-700,000	647,479	678273	606,337
5	800,000-900,000	884,111	862,721	885,057
6	1,000,000-1,100,000	1,061,765	1,049,409	1,003,905

Table 3: MLP architectural layouts.

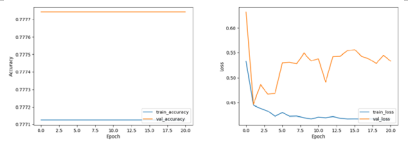
Size	Number layers	Number units	Layer followed by dropout
1	1	2	1
2	1	8	1
3	2	16-8	2
4	3	22-16-16	1-3
5	3	30-32-32	1-3
6	4	36-32-32-32	2-4

Table 4: LSTM architectural layouts.

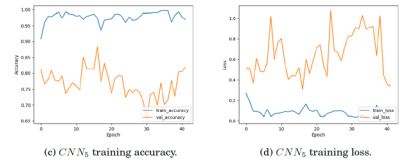
Size	Number layers	Number units	Layer followed by dropout
1	1	128	1
2	1	256	1
3	2	256-128	2
4	3	256-128-128	1-3
5	3	256-256-64	1-3
6	4	256-256-128-64	2-4

Table 5: CNN architectural layouts.

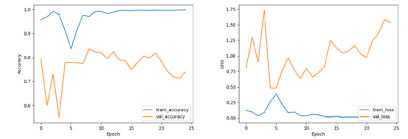
Size	Number Conv layers	Number filters	Kernel size	Pooling type	Layer followed by pooling	Pooling size
1	1	32	128	Global max	1	-
2	1	64	256	Global max	1	-
3	2	64-64	256-64	Global max	2	-
4	3	128-64-64	64-64-8	Global max	3	-
5	3	128-128-128	256-16-8	Max, global max	1-2-3	64-1
6	4	128-32-64-32	256-128-4-4	Max, global max	2-4	64



(a) MLP_1 training accuracy. (b) MLP_1 training loss.



(c) CNN_5 training accuracy. (d) CNN_5 training loss.



(e) $LSTM_5$ accuracy plot. (f) $LSTM_5$ training loss.

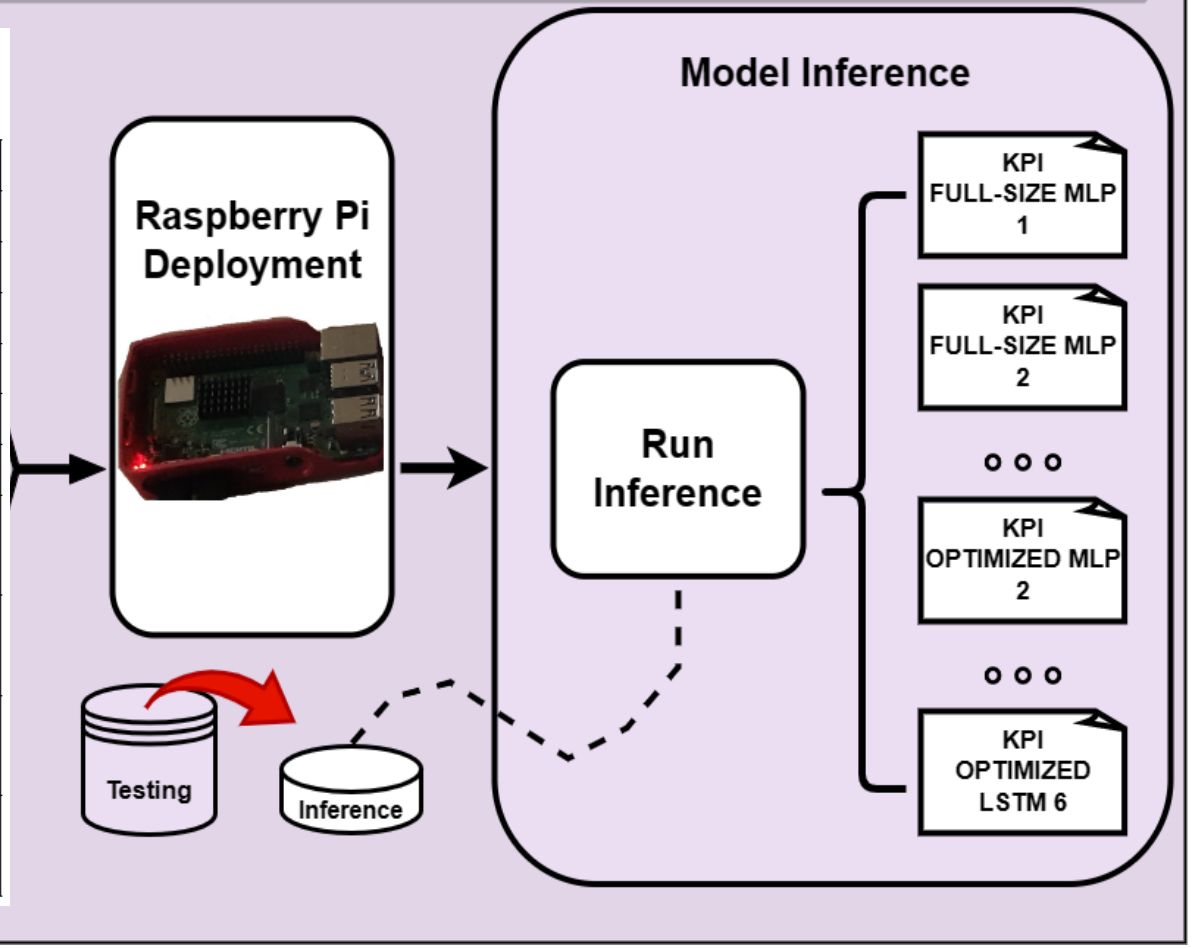
Eksperiment oppsett (II)

Model Optimization

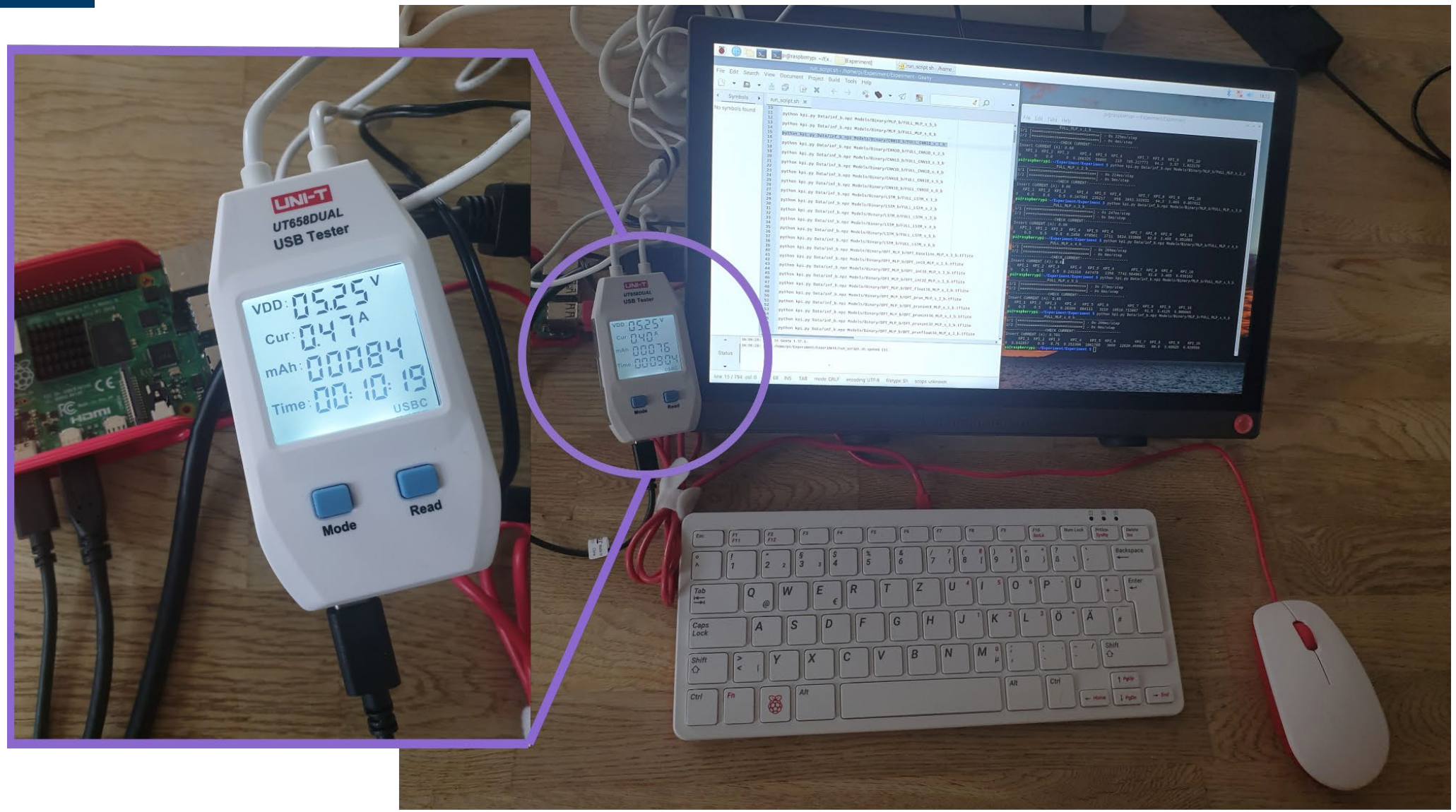
Table 6: Model optimization descriptions.

Optimization ID	Optimization Short Description
baseline	Dynamic range quantization
int8	Quantization integer only (8-bit)
int16	Quantization integer with float fallback (16-bit)
int32	Quantization integer with float fallback (32-bit)
float16	Quantization float (16-bit)
prun	Magnitude-based weight pruning
prunint8	Magnitude-based weight pruning + Quantization integer only (8-bit)
prunint16	Magnitude-based weight pruning + Quantization integer with float fallback (16-bit)
prunint32	Magnitude-based weight pruning + Quantization integer with float fallback (32-bit)
prunfloat16	Magnitude-based weight pruning + Quantization float (16-bit)

Model Deployment and Inference

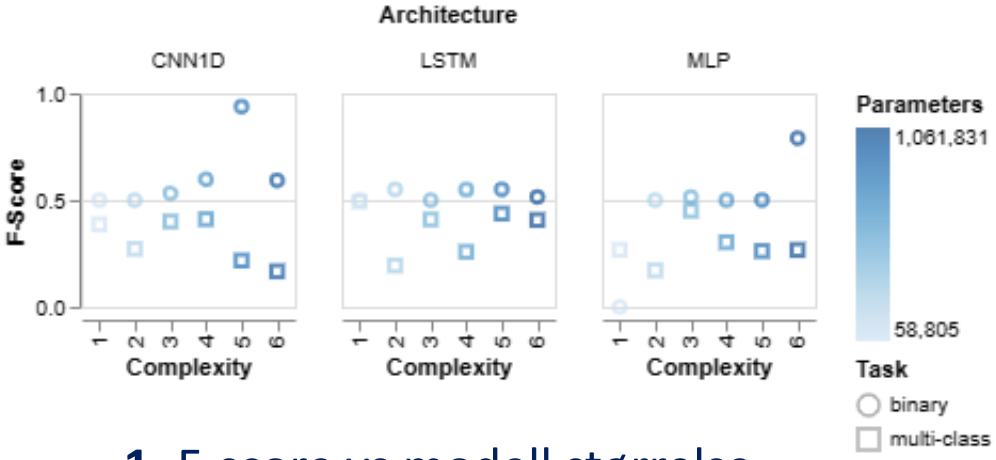


Eksperiment oppsett (III)

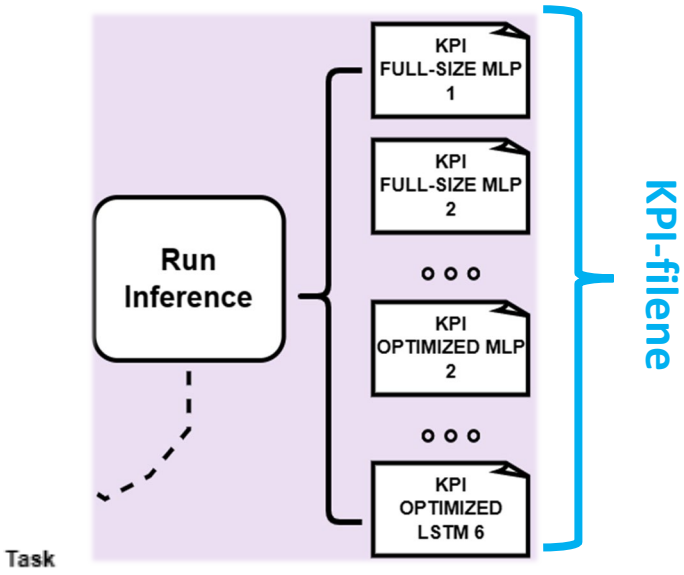


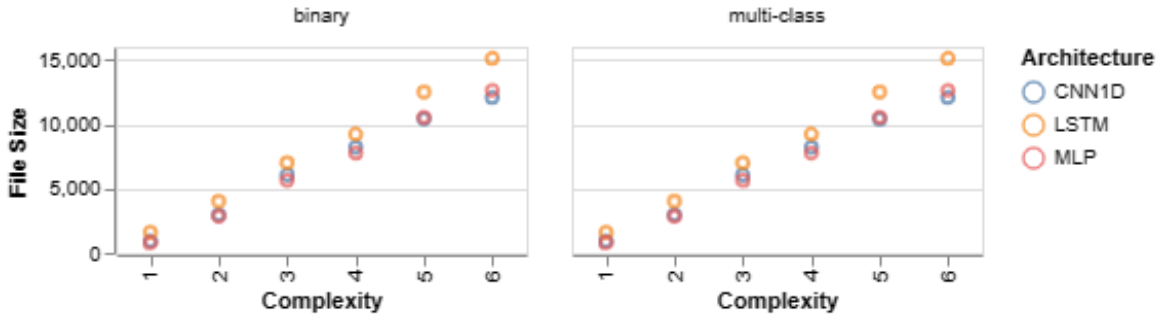
Analyse

- Vi bruker **KPI-filene** til å **analysere**:
- (1) Ressurskravene til de ikke-optimiserte modellene.

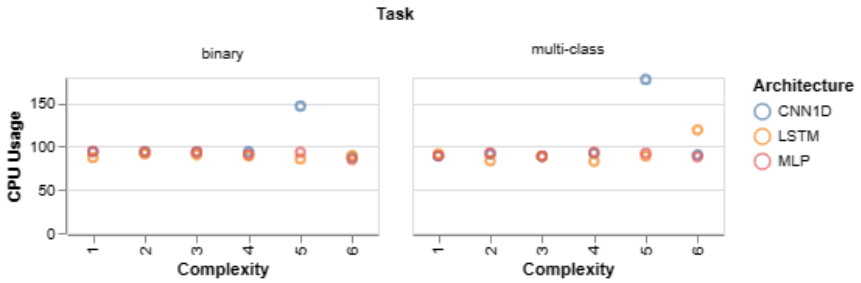


1. F-score vs modell størrelse.

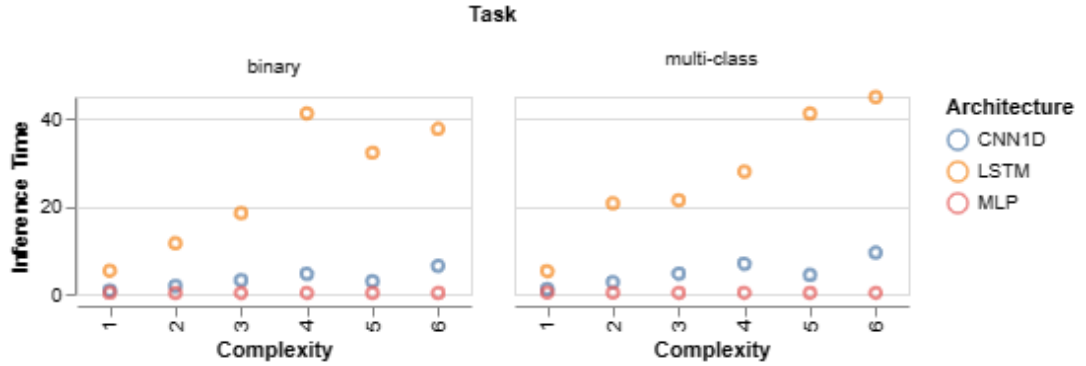




2. Filstørrelse vs modellstørrelse.



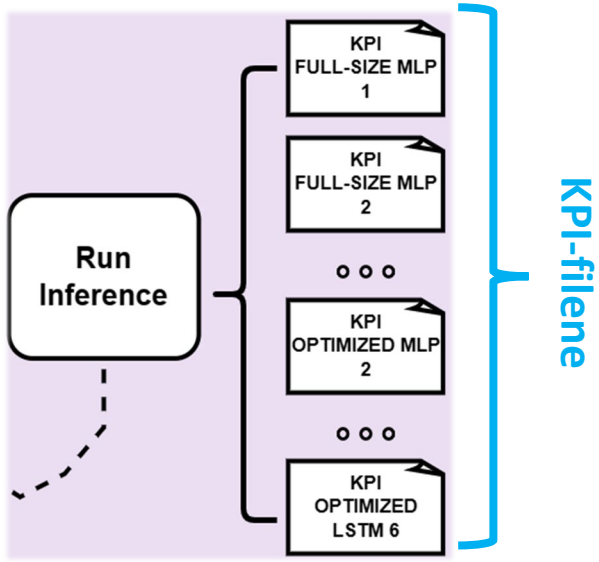
3. CPU-bruk vs modellstørrelse.

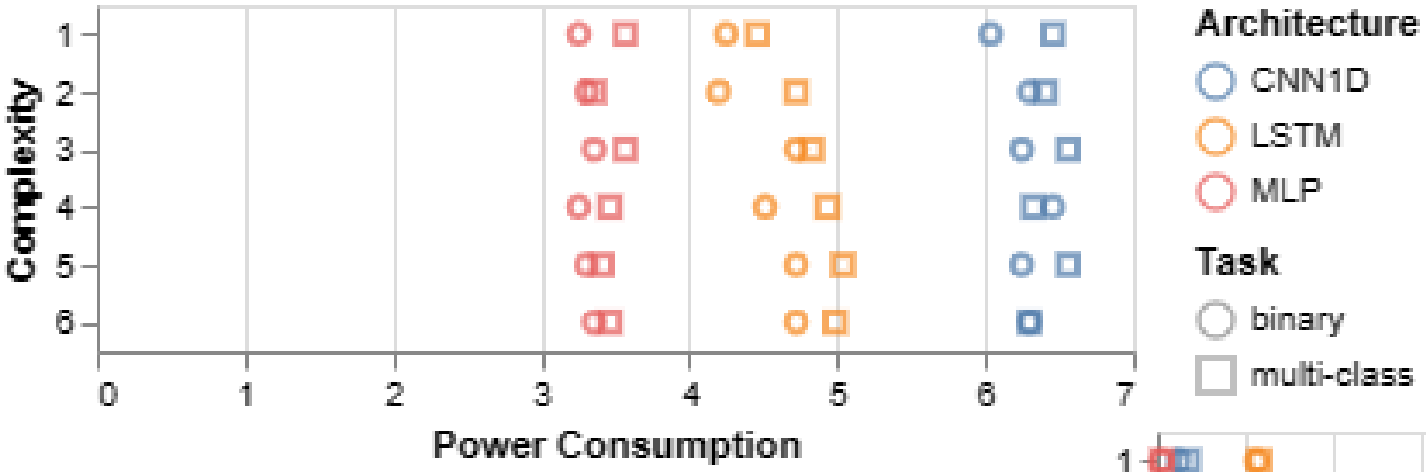


4. Inferenstid vs modellstørrelse.

Analyse

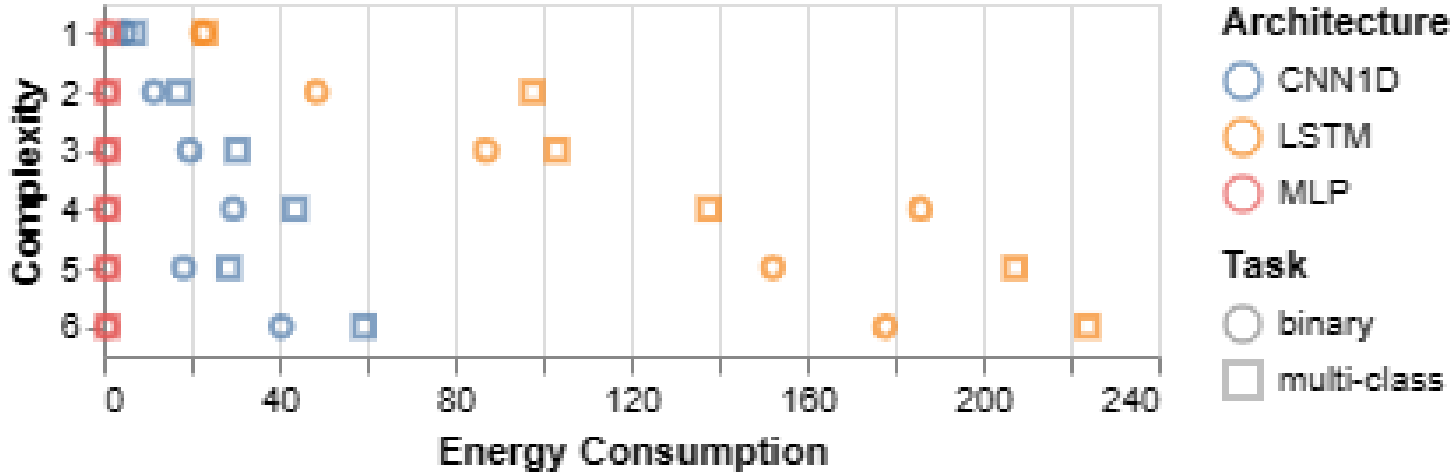
- Vi bruker **KPI-filene** til å **analysere**:
- (1) Ressurskravene til de ikke-optimiserte modellene.





5. Strømforbruk vs modellstørrelse.

6. Energiforbruk vs modellstørrelse.



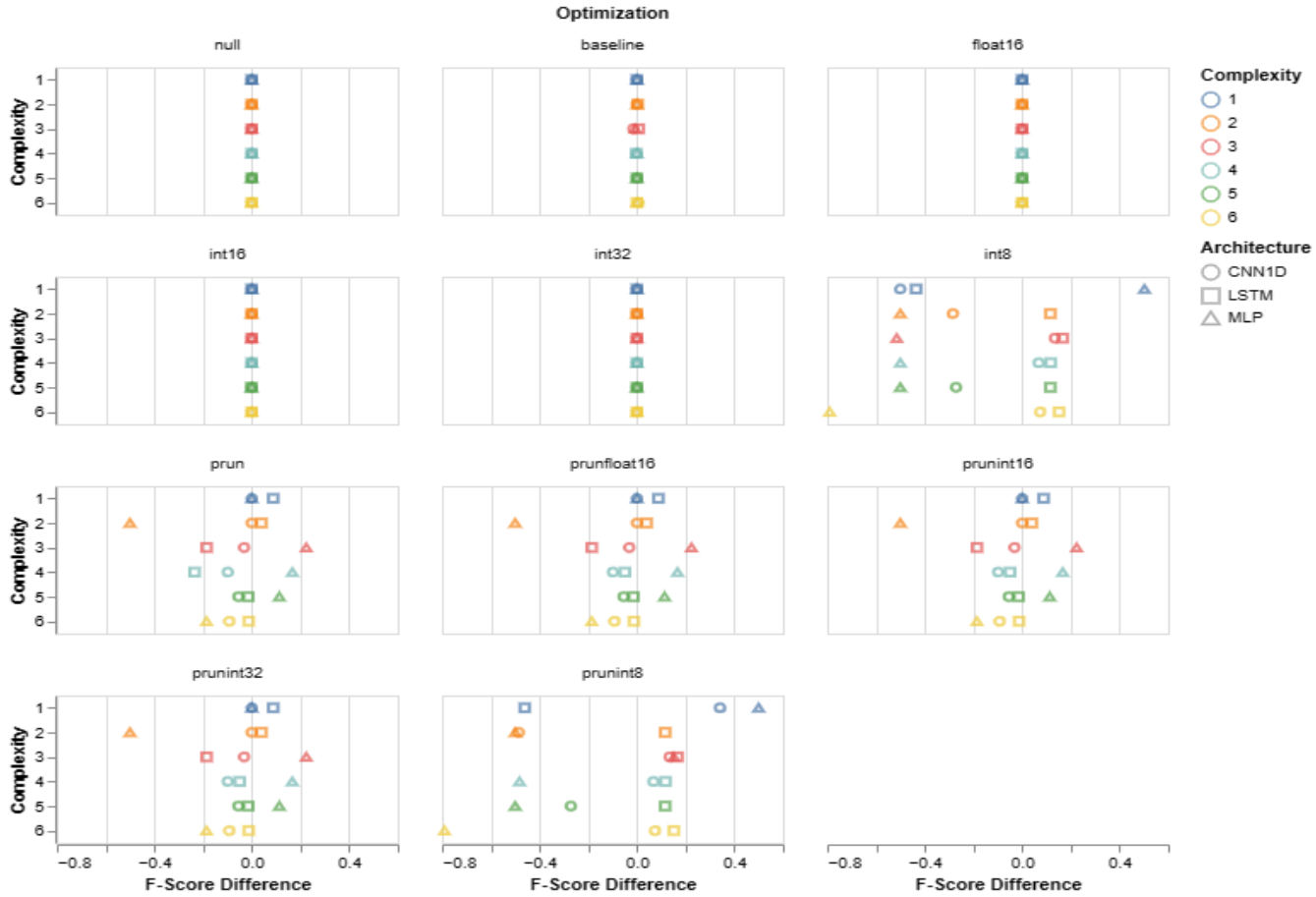
Analyse

- Vi bruker **KPI-filene** til å **analysere**:
- (2) Om de optimiserte modellene oppfyller **suksesskriteriene**.

✓ Redusere ressursutnyttelsen.

✓ Bevare nøyaktigheten.

For å sjekke om nøyaktigheten er bevart beregner man differansen i F-score mellom optimiserte og ikke-optimiserte modeller.



Analyse

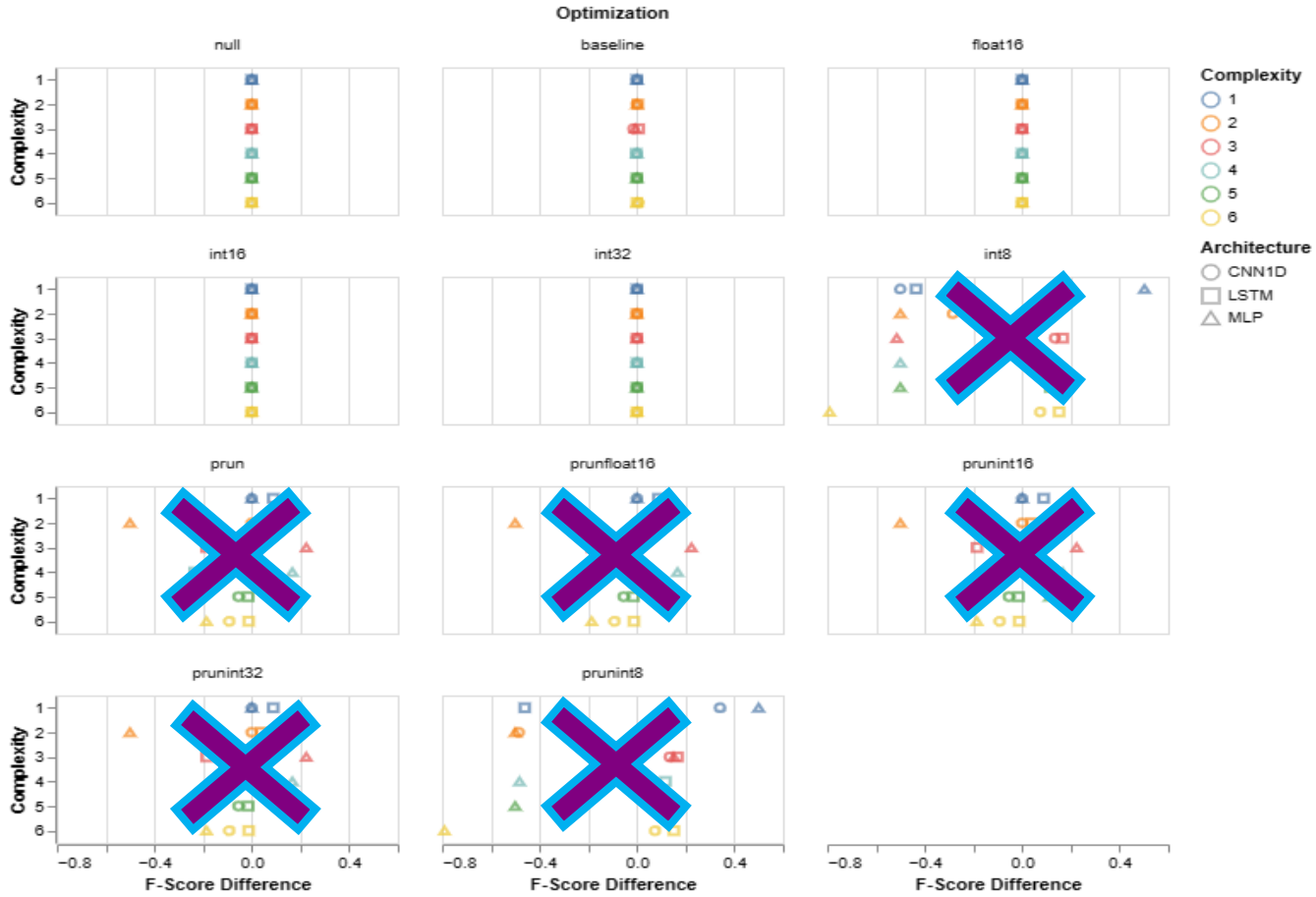
- Vi bruker **KPI-filene** til å **analysere**:
- (2) Om de optimiserte modellene oppfyller **suksesskriteriene**.

✓ Redusere ressursutnyttelsen.

✓ Bevare nøyaktigheten.

For å sjekke om nøyaktigheten er bevart beregner man differansen i F-score mellom optimiserte og ikke-optimiserte modeller.

Kvantiseringsoptimaliseringer (unntatt int8) oppfyller dette kriteriet!



Analyse

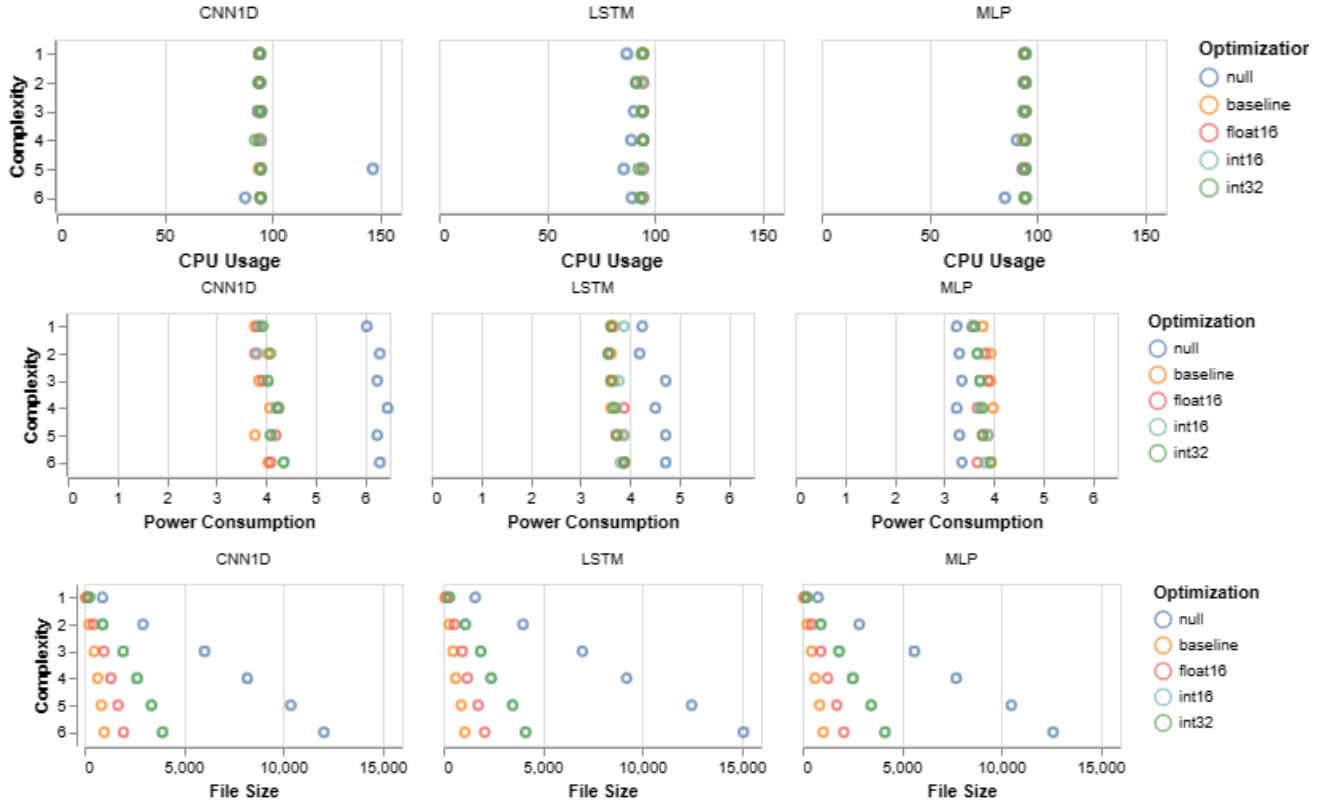
- Vi bruker **KPI-filene** til å **analysere**:
- (2) Om de optimiserte modellene oppfyller **suksesskriteriene**.

✓ Redusere ressursutnyttelsen.

✓ Bevare nøyaktigheten.

For å sjekke om ressursbruken er redusert, sjekker vi alle ressurser:

- CPU-bruk vs modellstørrelse.
- Strømforbruk vs modellstørrelse.
- Minneforbruk vs modellstørrelse.



Analyse

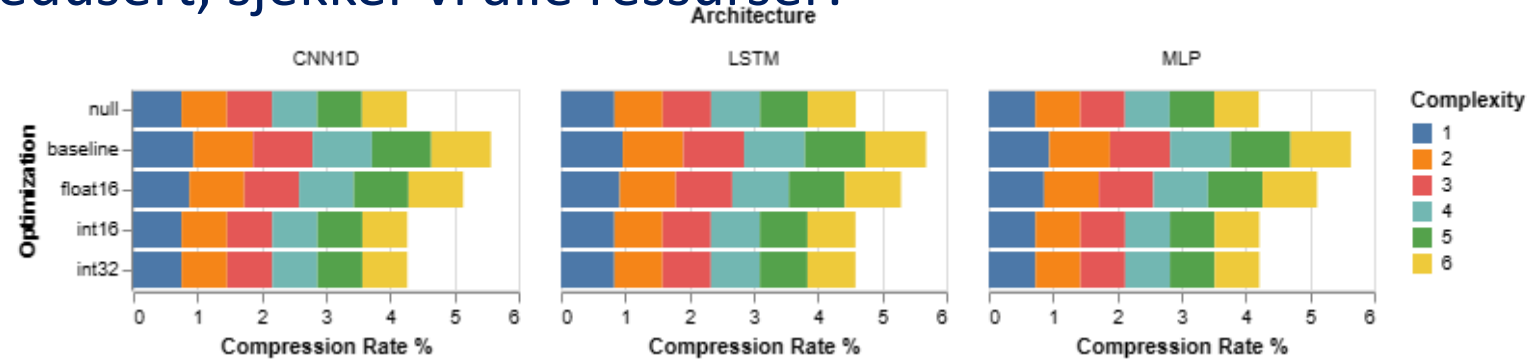
- Vi bruker **KPI-filene** til å **analysere**:
- (2) Om de optimiserte modellene oppfyller **suksesskriteriene**.

✓ Redusere ressursutnyttelsen.

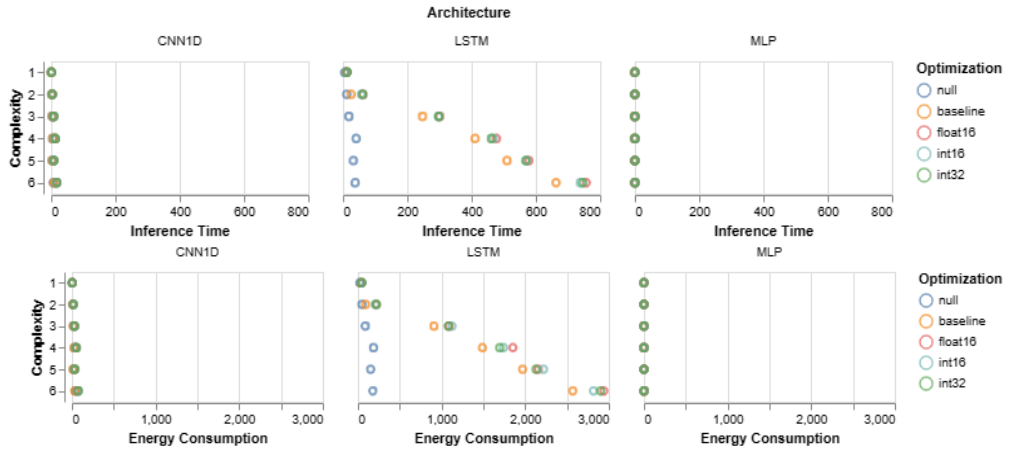
✓ Bevare nøyaktigheten.

For å sjekke om ressursbruken er redusert, sjekker vi alle ressurser:

- Komprimeringsgrad vs modellstørrelse.



- Inferenstid/energiforbruk vs modellstørrelse.



Analyse

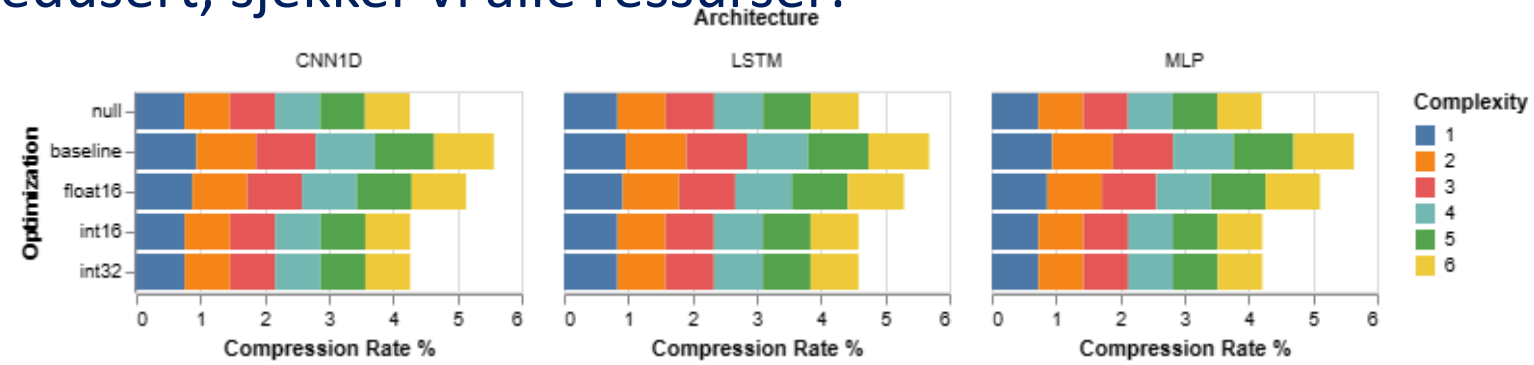
- Vi bruker **KPI-filene** til å **analysere**:
- (2) Om de optimiserte modellene oppfyller **suksesskriteriene**.

✓ Redusere ressursutnyttelsen.

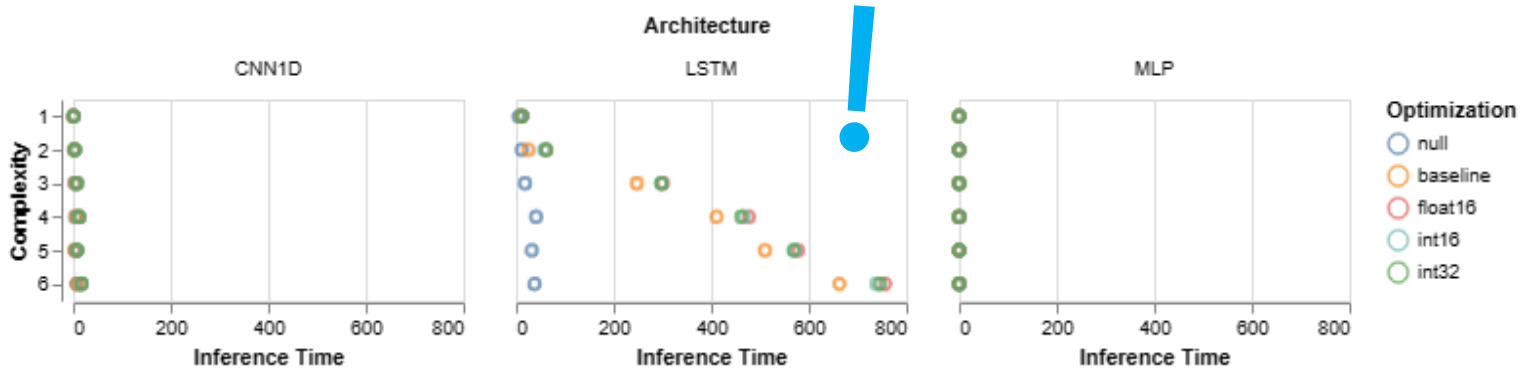
✓ Bevare nøyaktigheten.

For å sjekke om ressursbruken er redusert, sjekker vi alle ressurser:

- Komprimeringsgrad vs modellstørrelse.



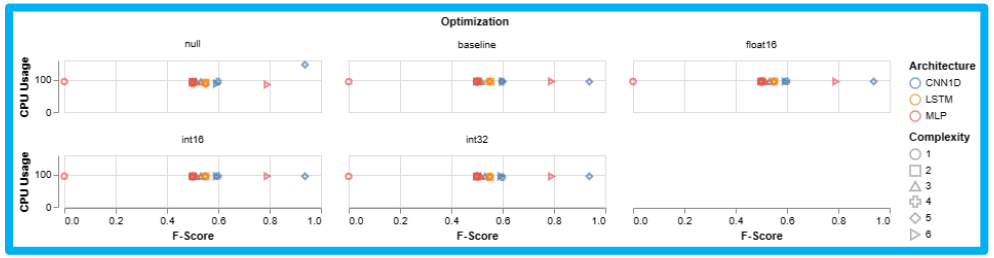
- Inferenstid/energiforbruk vs modellstørrelse.

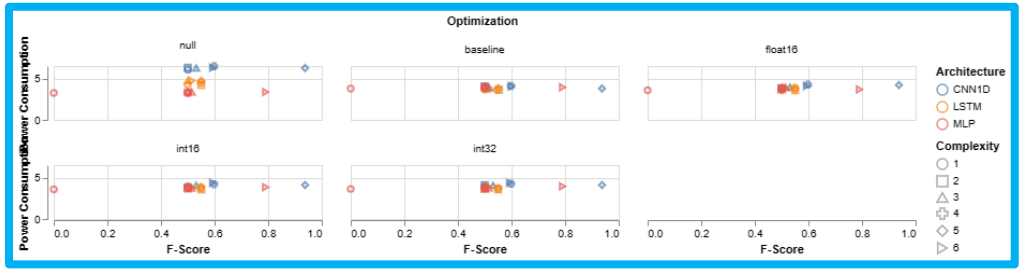


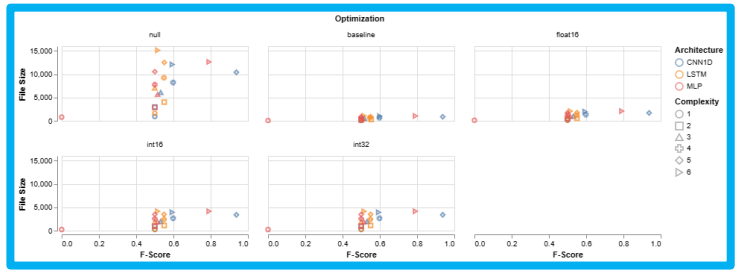
OBS! Ikke-optimiserte LSTM-modeller presterer best!

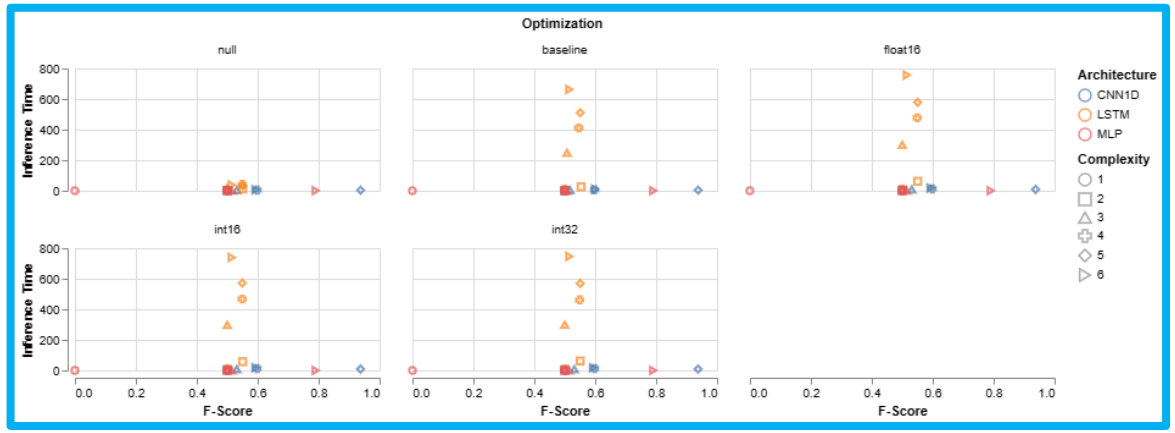
Analyse

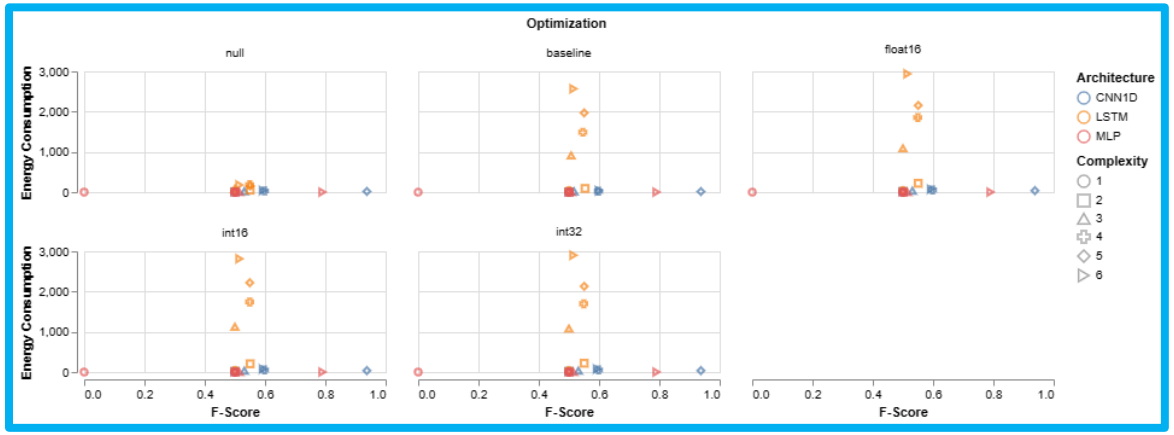
- Vi bruker **KPI-filene** til å **analysere**:
- (3) Ressurs vs nøyaktighet **avveining**: hvilke optimaliseringer oppnår en bedre **balanse**?











Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

(3) → **MEN!**

- baseline og float16 har den beste avveiningen når det gjelder minneforbruk.
- baseline har den beste avveiningen når det gjelder strømforbruk.
- alle optimiseringer viser svært liknende resultater for resten av ressursene.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

(3) → **MEN!**

- baseline og float16 har den beste avveiningen når det gjelder minneforbruk.
- baseline har den beste avveiningen når det gjelder strømforbruk.
- alle optimiseringer viser svært liknende resultater for resten av ressursene.

→ Rammeverket lykkes med å eliminere optimiseringsteknikk som ikke oppfyller suksesskriteriene og fange avveininger mellom nøyaktighet og ressurser.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

(3) → **MEN!**

- baseline og float16 har den beste avveiningen når det gjelder minneforbruk.
- baseline har den beste avveiningen når det gjelder strømforbruk.
- alle optimiseringer viser svært liknende resultater for resten av ressursene.

→ Rammeverket lykkes med å eliminere optimiseringsteknikk som ikke oppfyller suksesskriteriene og fange avveininger mellom nøyaktighet og ressurser.

→ Eksperimentresultater viser at MLP og CNN modeller kan optimaliseres ved hjelp av **baseline**, **int16**, **int32** og **float16**, og samtidig oppfylle begge to kriteriene.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

(3) → **MEN!**

- baseline og float16 har den beste avveiningen når det gjelder minneforbruk.
- baseline har den beste avveiningen når det gjelder strømforbruk.
- alle optimiseringer viser svært liknende resultater for resten av ressursene.

→ Rammeverket lykkes med å eliminere optimiseringsteknikk som ikke oppfyller suksesskriteriene og fange avveininger mellom nøyaktighet og ressurser.

→ Eksperimentresultater viser at MLP og CNN modeller kan optimaliseres ved hjelp av **baseline**, **int16**, **int32** og **float16**, og samtidig oppfylle begge to kriteriene.

→ Pruning kan i noen tilfeller forbedre nøyaktighetsgraden til en modell.

Resultater

(1) → Ulike arkitekturer har ulike behov for optimalisering.

(2) → Vi finner at ingen av optimaliseringene fra kandidatsettet oppfyller kriteriene.

(3) → **MEN!**

- baseline og float16 har den beste avveiningen når det gjelder minneforbruk.
- baseline har den beste avveiningen når det gjelder strømforbruk.
- alle optimiseringer viser svært liknende resultater for resten av ressursene.

→ Rammeverket lykkes med å eliminere optimiseringsteknikk som ikke oppfyller suksesskriteriene og fange avveininger mellom nøyaktighet og ressurser.

→ Eksperimentresultater viser at MLP og CNN modeller kan optimaliseres ved hjelp av **baseline**, **int16**, **int32** og **float16**, og samtidig oppfylle begge to kriteriene.

→ Pruning kan i noen tilfeller forbedre nøyaktighetsgraden til en modell.

→ Rammeverket kan forbedres!

Referanser

Masteroppgave:
Videsjorden, A. N. (2023). *Optimizing deep learning inference for resource-constricted platforms*

[1] Zheng, T., Ardolino, M., Bacchetti, A., & Perona, M. (2021). The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. *International Journal of Production Research*, *59*(6), 1922-1954.
[2] Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. *International Journal of Production Research*, *59*(16), 4773-4778.
[3] Siccoli, A., de Wispelaere, M. P., Schröder, M. L., & Staartjes, V. E. (2019). Machine learning–based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurgical Focus*, *46*(5), E5.
[4] Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., & Allen, C. (2021). Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews*, *175*, 113806.
[5] Bai, Y., Guan, Y., & Ng, W. F. (2020, September). Fatigue assessment using ECG and actigraphy sensors. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers* (pp. 12-16).
[6] Garcia-Ceja, E., Osmani, V., & Mayora, O. (2015). Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics*, *20*(4), 1053-1060.
[7] Little, B., Alshabrawy, O., Stow, D., Ferrier, I. N., McNaney, R., Jackson, D. G., ... & O'Brien, J. T. (2021). Deep learning-based automated speech detection as a marker of social functioning in late-life depression. *Psychological Medicine*, *51*(9), 1441-1450.
[8] Tang, C. S., & Veelenturf, L. P. (2019). The strategic role of logistics in the industry 4.0 era. *Transportation Research Part E: Logistics and Transportation Review*, *129*, 1-11.
[9] Bouman, P., Agatz, N., & Schmidt, M. (2018). Dynamic programming approaches for the traveling salesman problem with drone. *Networks*, *72*(4), 528-542.
[10] Sajjad, M., Irfan, M., Muhammad, K., Del Ser, J., Sanchez-Medina, J., Andreev, S., ... & Lee, J. W. (2020). An efficient and scalable simulation model for autonomous vehicles with economical hardware. *IEEE transactions on intelligent transportation systems*, *22*(3), 1718-1732.
[11] Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25, pp. 15-24). San Francisco, CA, USA: Determination press.
[12] Coelho, Y. L., dos Santos, F. D. A. S., Frizera-Neto, A., & Bastos-Filho, T. F. (2021). A lightweight framework for human activity recognition on wearable devices. *IEEE Sensors Journal*, *21*(21), 24471-24481.
[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press
[14] Buyya, R., Broberg, J., & Goscinski, A. M. (Eds.). (2010). *Cloud computing: Principles and paradigms*. John Wiley & Sons.
[15] Satyanarayanan, M., Bahl, P., Caceres, R., & Davies, N. (2009). The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing*, *8*(4), 14-23.
[16] Parikh, S., Dave, D., Patel, R., & Doshi, N. (2019). Security and privacy issues in cloud, fog and edge computing. *Procedia Computer Science*, *160*, 734-739.
[17] Liu, D., Kong, H., Luo, X., Liu, W., & Subramaniam, R. (2022). Bringing AI to edge: From deep learning's perspective. *Neurocomputing*, *485*, 297-320.

Takk!
Q & A