Statistical Analysis of Sign Language Recognition with CNNs

Final Presentation

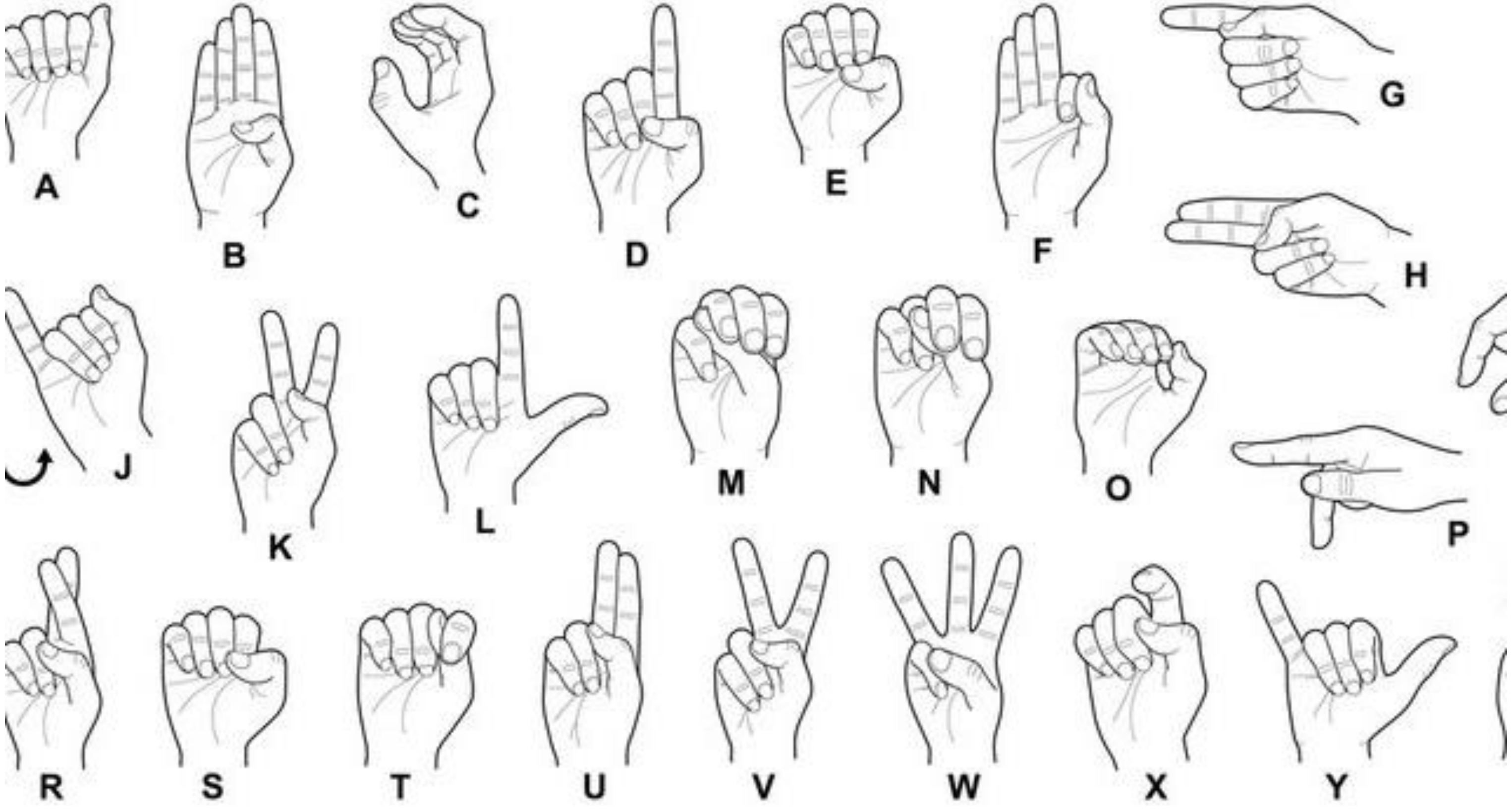
Advanced Topics of AI – P

21.01.2026

Julia Xu

Sign Language Recognition Using Simple CNNs

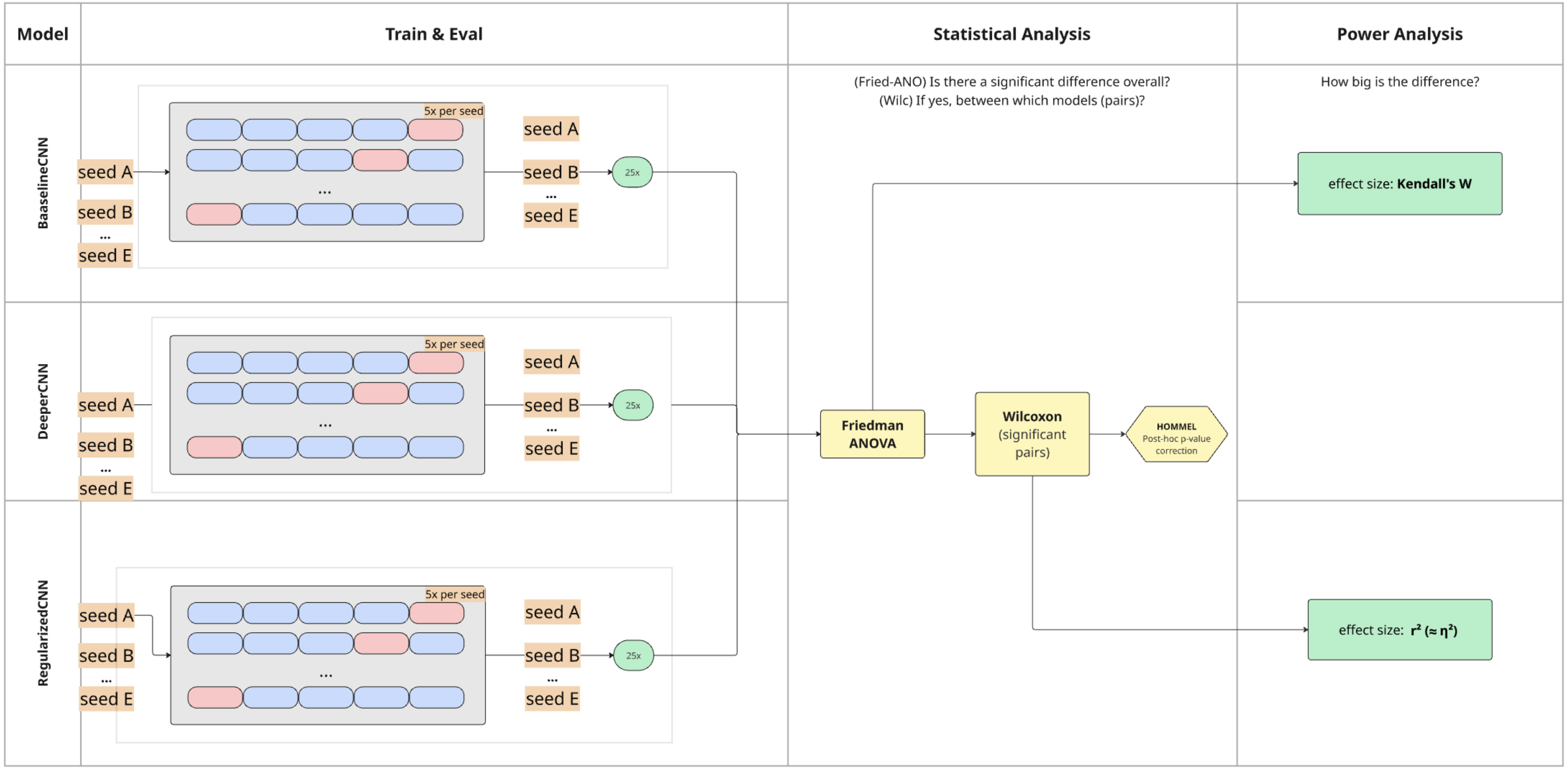
Topic



Agenda

1. Research Plan
2. Model Variants
3. Experiment Setup
4. Results
    1. Training
    2. Statistical Analysis
5. Conclusion

1. Research Plan



Model

Train & Eval

Statistical Analysis

Power Analysis

BaaselineCNN

DeeperCNN

RegularizedCNN

seed A

seed B

...

seed E

5x per seed

25x

(Fried-ANO) Is there a significant difference overall?
(Wilc) If yes, between which models (pairs)?

Friedman ANOVA

Wilcoxon (significant pairs)

HOMMEL Post-hoc p-value correction

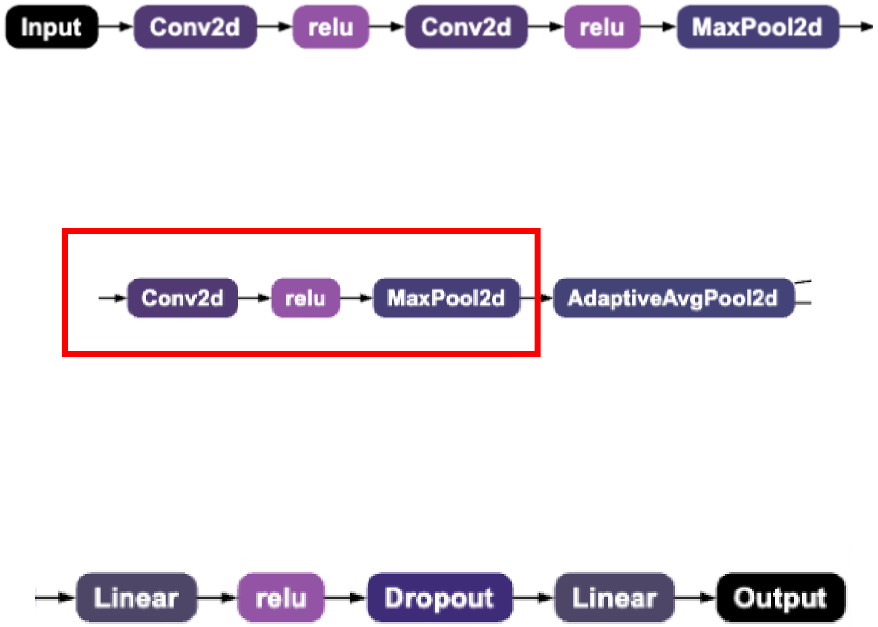
How big is the difference?

effect size: **Kendall's W**

effect size: **r^2 (≈ η^2)**

2. Model Variants/Groups

- CNN model variants:
  - **BaselineCNN**: **2**-layered CNN w/ dropout 0.3
  - **DeeperCNN**: **3**-layered CNN w/ dropout 0.3
  - **RegularizedCNN**: 3-layered CNN w/ dropout **0.5**
- MobileNetV3 small:
  - Pretrained
  - Discarded due to very high accuracies



Hugging Face. Model card for mobilenetv3 small 100.lamb in1k. https://huggingface.co/timm/mobilenetv3_small_100.lamb_in1k. [Accessed 21-11-2025]

3. Experiment Setup

- Preprocessing:
  - Dataset: Kaggle ASL Alphabet (200x200 images)
  - 10% subset, resizing to 224x224:
    - Train samples: 8700
- Training:
    - Macbook Apple M2 chip, 8-core CPU and 8-core GPU
    - device: mps
    - 5 times repeated 5-fold CV per group per seed → 75 measurements

Khan, Anas and Nagaraj, Akash. Asl alphabet. https://www.kaggle.com/datasets/grassknoted/asl-alphabet. [Accessed 03-11-2025]

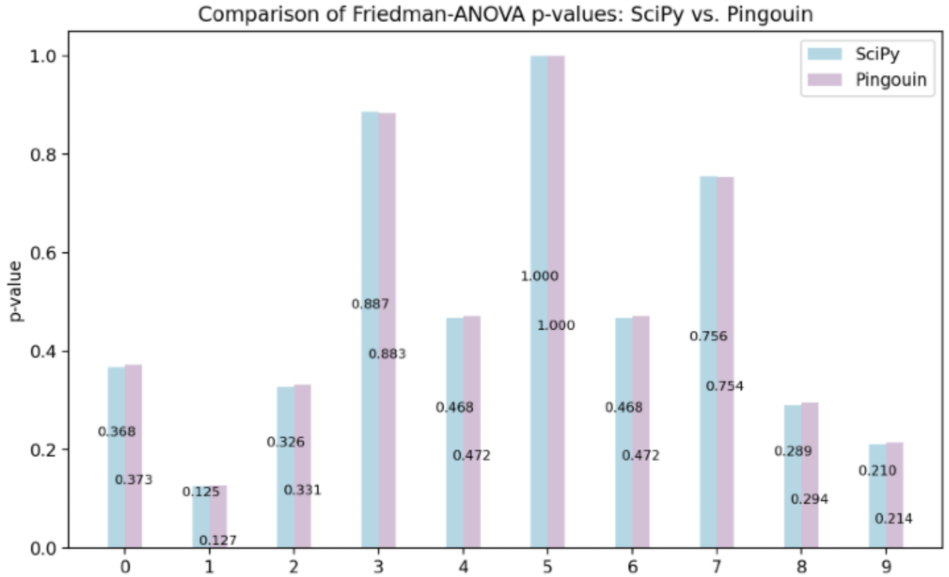
3. Experiment Setup

- Statistical Analysis:
  - Friedman ANOVA
  - Wilcoxon signed-rank test
  - Post-hoc Hommel correction
- Effect Analysis:
  - Friedman ANOVA: effect size Kendall’s W
  - Wilcoxon: effect size r

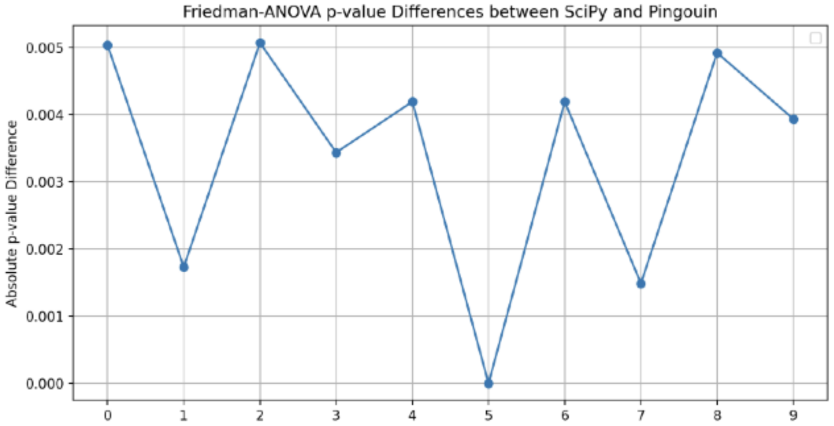
Janez Demšar. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn.Res., 7:1–30, December 2006
G. HOMMEL. A stagewise rejective multiple test procedure based on a modified bonferronitest. Biometrika, 75(2):383–386, 1988

Excursion: scipy vs. pingouin

Results of 10x repeated statistical tests implementation of scipy and pingouin with three example 1x25 vectors → using scipy



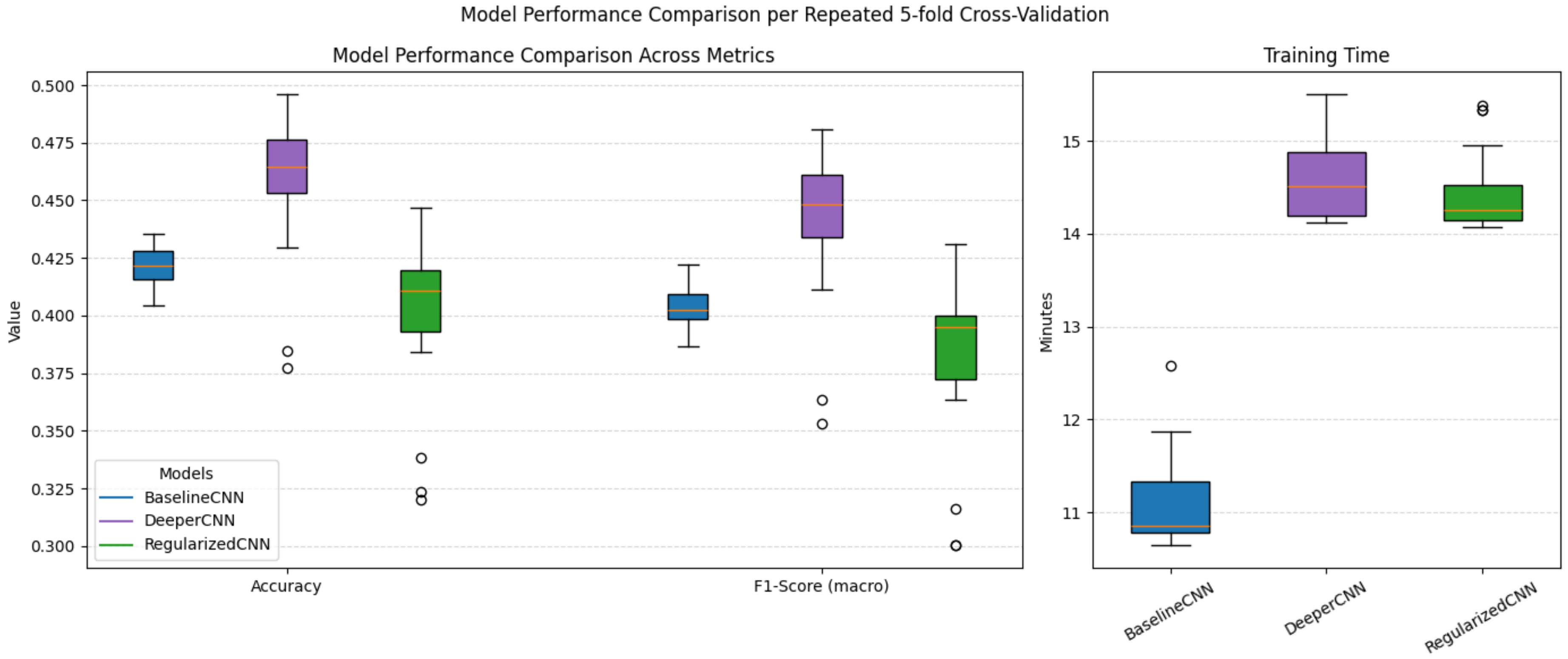
(a) Overview of test results. The x-axis corresponds to the numbered test repetition.



(b) Absolute p-value differences. The x-axis corresponds to the numbered test repetition.

scipy. friedmanchisquare SciPy v1.16.2 Manual. https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.friedmanchisquare.html. [Accessed 27-11-2025].
scipy. wilcoxonSciPy v1.16.2 Manual. https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.wilcoxon.html. [Accessed 27-11-2025]

4. Results: Training Outcome



- ➔ DeeperCNN best accuracy, f1-score values
- ➔ Longest training for DeeperCNN as well

4. Results: Statistical Analysis

Friedman ANOVA

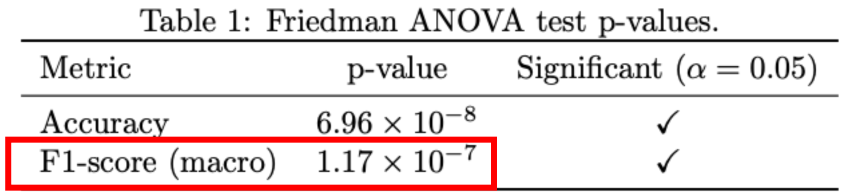


Table 1: Friedman ANOVA test p-values.

Metric	p-value	Significant ($\alpha = 0.05$)
Accuracy	6.96×10^{-8}	✓
F1-score (macro)	1.17×10^{-7}	✓

Post-hoc Hommel corrected Wilcoxon (acc)

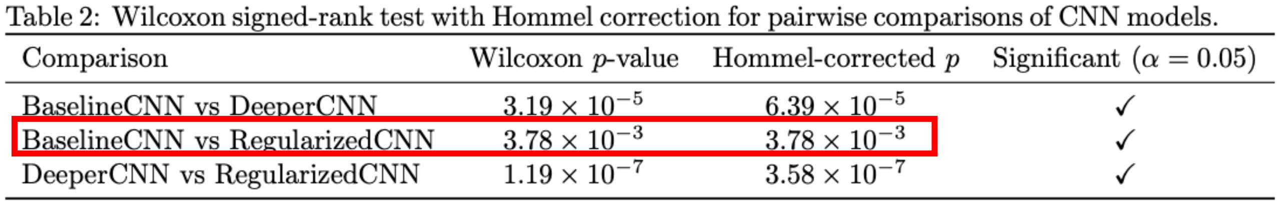


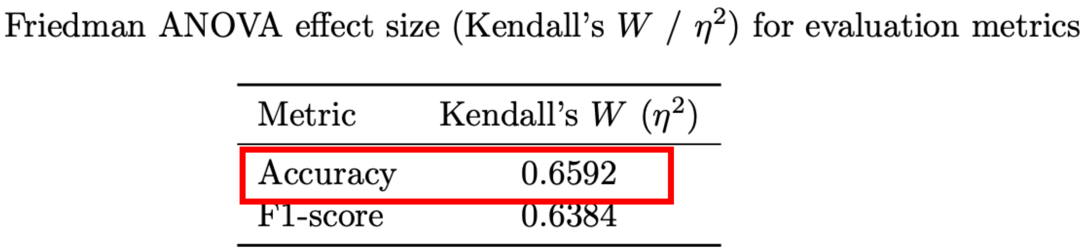
Table 2: Wilcoxon signed-rank test with Hommel correction for pairwise comparisons of CNN models.

Comparison	Wilcoxon p-value	Hommel-corrected p	Significant ($\alpha = 0.05$)
BaselineCNN vs DeeperCNN	3.19×10^{-5}	6.39×10^{-5}	✓
BaselineCNN vs RegularizedCNN	3.78×10^{-3}	3.78×10^{-3}	✓
DeeperCNN vs RegularizedCNN	1.19×10^{-7}	3.58×10^{-7}	✓

Fernanda Fiel Peres. Effect sizes for nonparametric tests. Biochemia Medica, 36(1), December2025

4. Results: Effect Analysis

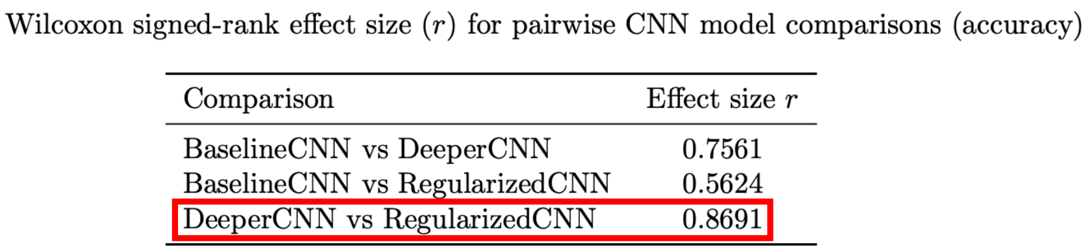
Effect size for Friedman ANOVA: Kendall’s W



Friedman ANOVA effect size (Kendall’s W / η^2) for evaluation metrics

Metric	Kendall’s W (η^2)
Accuracy	0.6592
F1-score	0.6384

Effect size for Wilcoxon Test: r



Wilcoxon signed-rank effect size (r) for pairwise CNN model comparisons (accuracy)

Comparison	Effect size r
BaselineCNN vs DeeperCNN	0.7561
BaselineCNN vs RegularizedCNN	0.5624
DeeperCNN vs RegularizedCNN	0.8691

Fernanda Fiel Peres. Effect sizes for nonparametric tests. Biochemia Medica, 36(1), December2025

Conclusion

- **Friedman ANOVA**: compare all modes together

  → there exists a statistical significance among the CNN variants

- **Wilcoxon signed-rank**: compare model pairs

  → Each model pair is significantly different (most BaselineCNN vs. RegularizedCNN)

- **Effect sizes**:
  - For Friedman ANOVA: very similar, acc slightly higher power
  - For Wilcoxon: highest effect size for DeeperCNN vs RegularizedCNN

Thank you.