Feature-Guided Neighbor Selection for Non-Expert Evaluation of Model Predictions

Courtney Ford[1,2] **and Mark T. Keane**[1,2,3]

[1]Research Ireland CRT in Machine Learning, University College Dublin
[2]School of Computer Science, University College Dublin
[3]Insight Centre for Data Analytics, University College Dublin

The Challenge

- Non-experts struggle to interpret AI decisions in unfamiliar domains, especially for misclassifications
- Traditional k-NN explanations select neighbors based on feature-space proximity

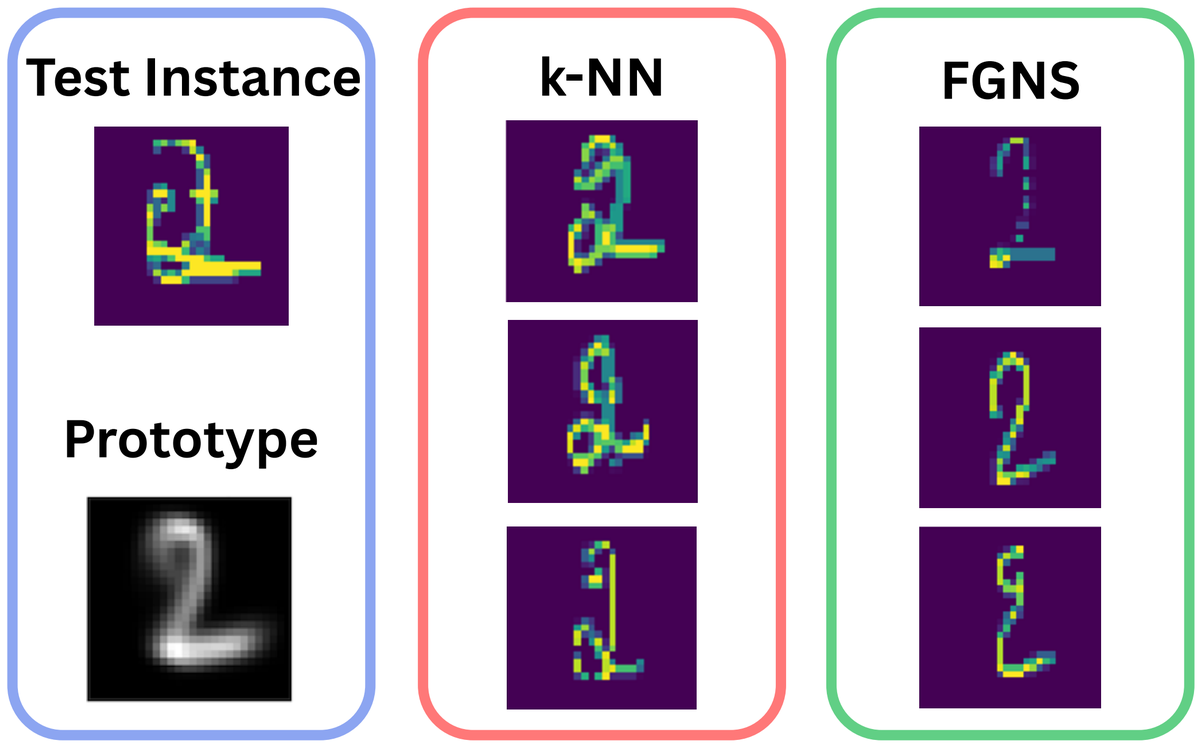
Key Problem:
How can we help non-experts identify model errors when they lack domain knowledge?

Our Solution: FGNS

- Selects neighbors that represent key class characteristics
- Prioritizes semantic alignment over spatial proximity

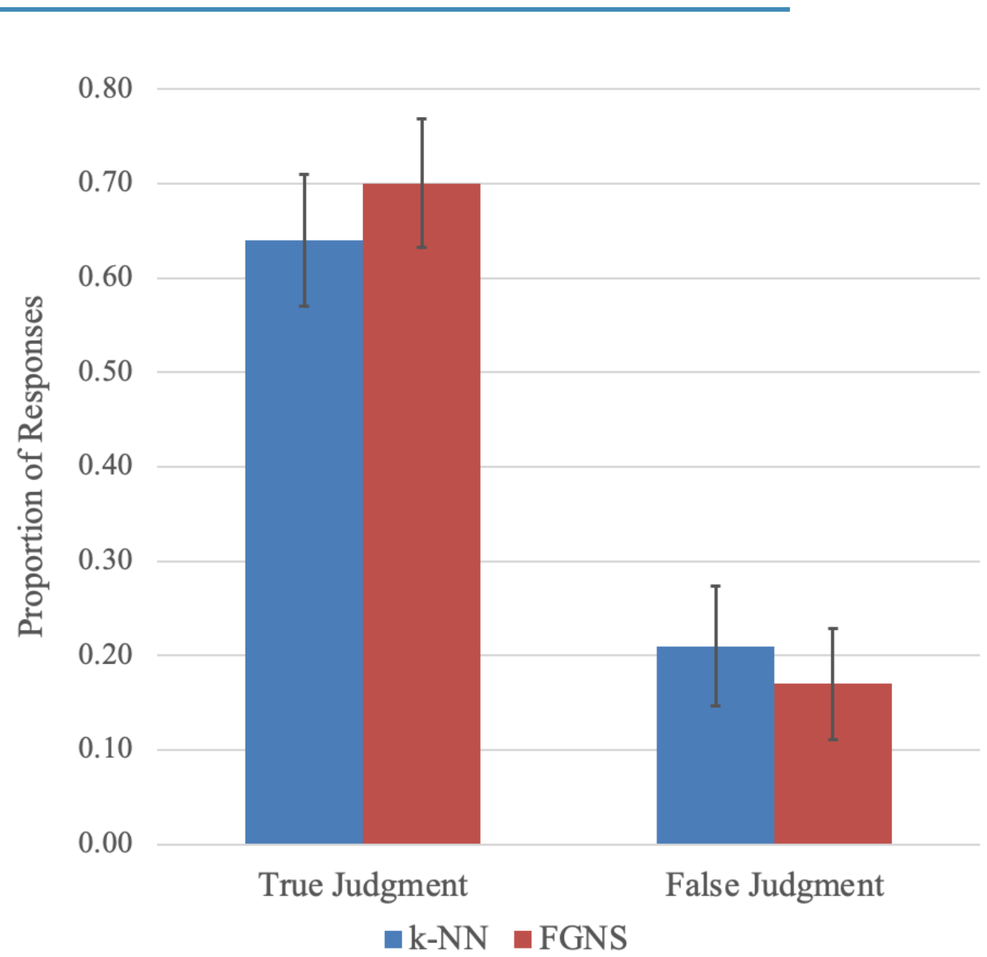
Correct Classifications: Shows typical class examples that confirm the prediction.

Misclassifications: Shows what the predicted class *should* look like, revealing a greater disparity between the test instance and example cases.



FGNS helps users spot errors by showing what the predicted class should look like, rather than just the closest matches.

Results



- FGNS improved True Judgment* and False Judgment Accuracy over k-NN
- FGNS explanations resulted in faster decision making (12.9s vs 15.7s for k-NN)
- FGNS neighbors are closer to, and cluster more tightly around, ground truth prototypes.

Science Foundation Ireland
Centre for Research Training
in Machine Learning