P.T. van der Knaap, p.t.vanderknaap@student.tudelft.nl Supervisors: David M.J. Tax, d.m.j.tax@tudelft.nl & Caspar C.J. van Engelenburg, C.C.J.vanEngelenburg@tudelft.nl

Establishing the Parallel Diagonal Dataset, a CNN problem archetype

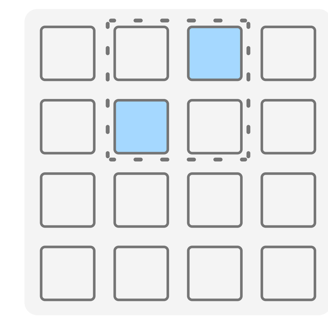
1. Background & Problem

Problem
- No universal 'best' deep-learning model exists
- No systematic overview for mapping problems to models

CNN Inductive Biases
- Spatial locality: nearby pixels are highly related
- Translation invariance: same pattern detected anywhere on the grid

Approach
- Identify dataset characteristics that inherently suit CNNs
- Let models compete, maximise the CNN binary classification gap
- Competing models: MLPs, GCNs, RNNs, Transformers

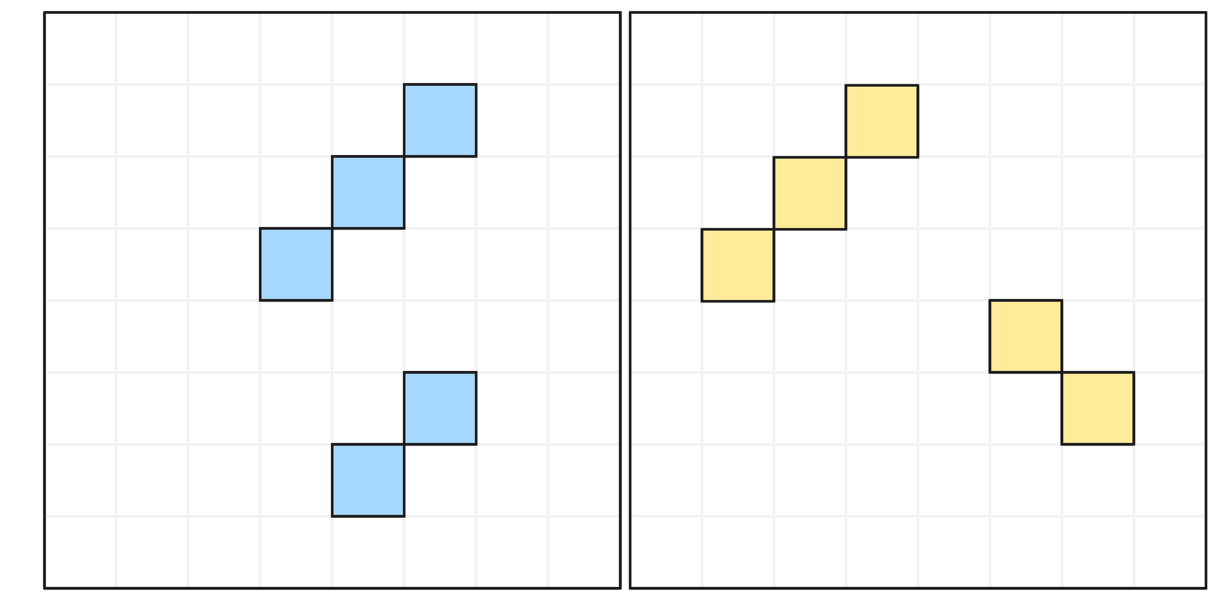
CNNs detect patterns anywhere on a grid by sliding the same filters across the entire grid



Research Question

What synthetic binary classification problem enables CNNs to consistently outperform competing deep learning architectures?

Parallel Diagonal Dataset Samples
Class 0: parallel diagonals (left), class 1: opposing diagonals (right)



3. Problem Archetype: Parallel Diagonal

What synthetic characteristics exploit CNN inductive biases?

Binary Classification Dataset
- Two diagonals
- Class determined by whether orientations match

CNNs inherently suited for the dataset
- Spatial locality:
  - Filters detect diagonals
  - Deeper layers compose orientations
- Translation invariance:
  - Shared filter weights detect the same pattern anywhere on the grid

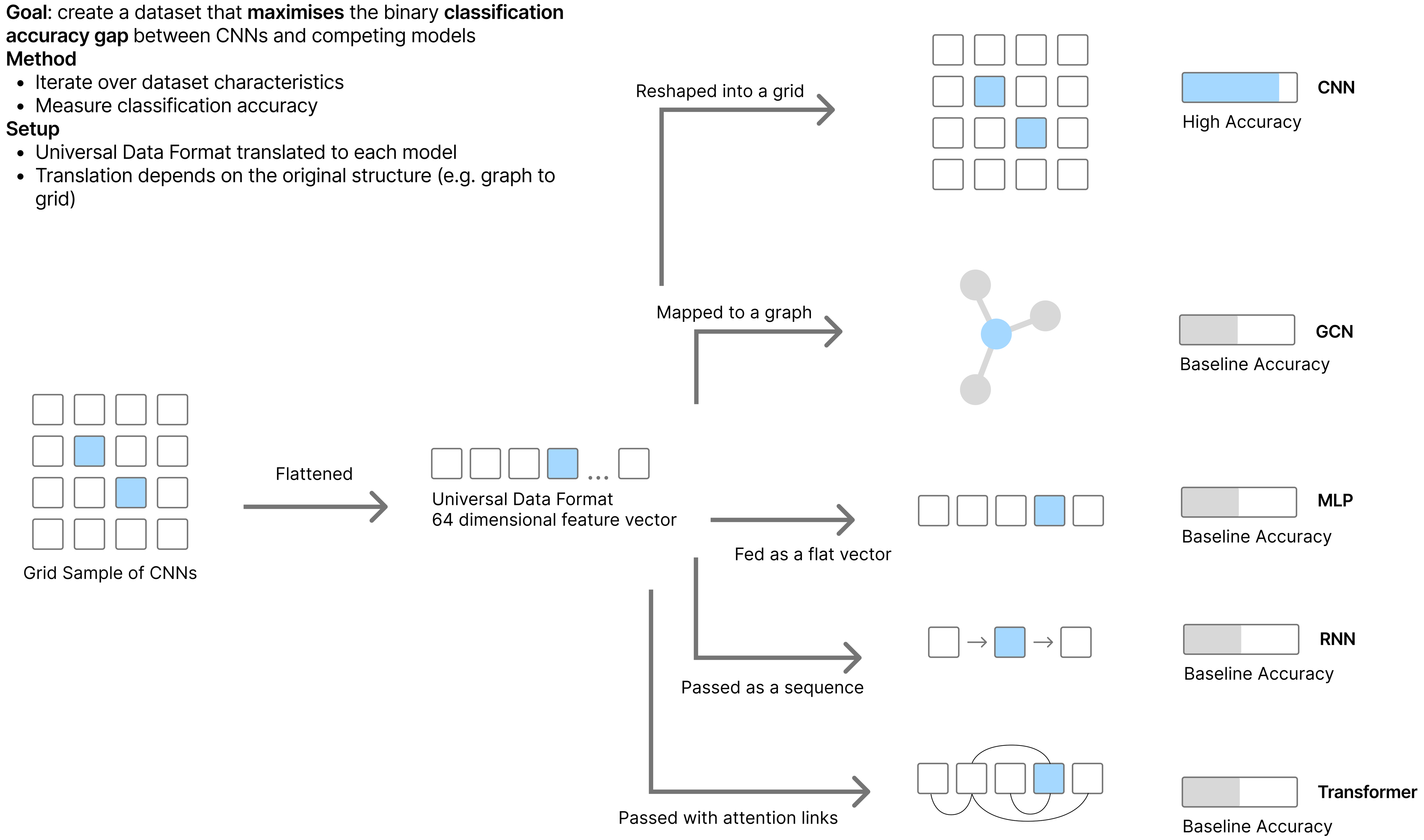
Why competing models struggle
- No spatial awareness: must detect patterns through other signals
- No translation invariance: must generalise across 3456 distinct combinations

Arbitrary placement and diagonal orientations inherently suit CNN inductive biases

2. Experimental Setup: Competition between Models

How can we verify the Problem Archetype of CNNs?

- **Goal**: create a dataset that **maximises** the binary **classification accuracy gap** between CNNs and competing models
- **Method**
  - Iterate over dataset characteristics
  - Measure classification accuracy
- **Setup**
  - Universal Data Format translated to each model
  - Translation depends on the original structure (e.g. graph to grid)



4. Experimental Results

How large is the classification accuracy difference, and does it hold under increased dataset size and noise?

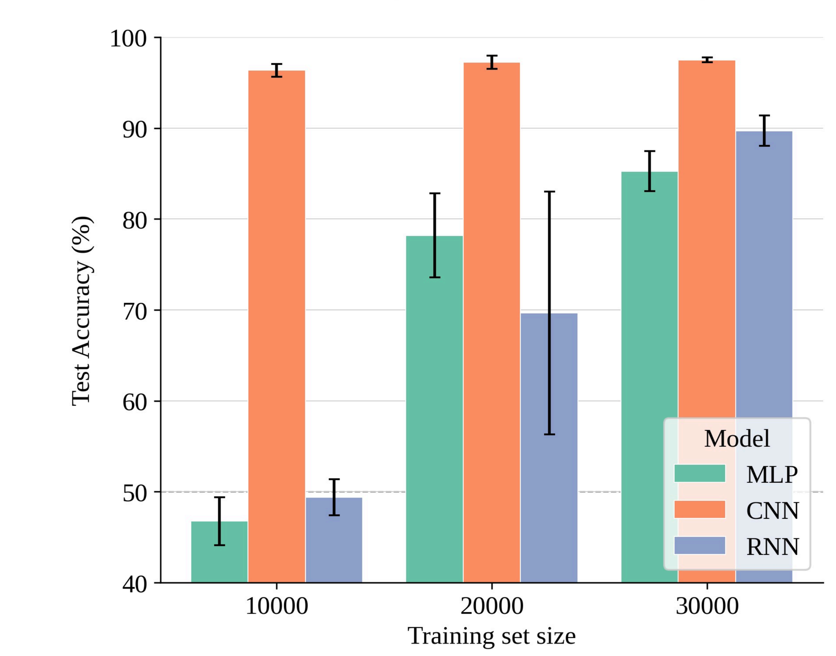
CNNs achieve **90%** accuracy at 1,000 samples, competitors remain **near random chance**

RNNs and **MLPS** require **30×** more samples to close the accuracy gap

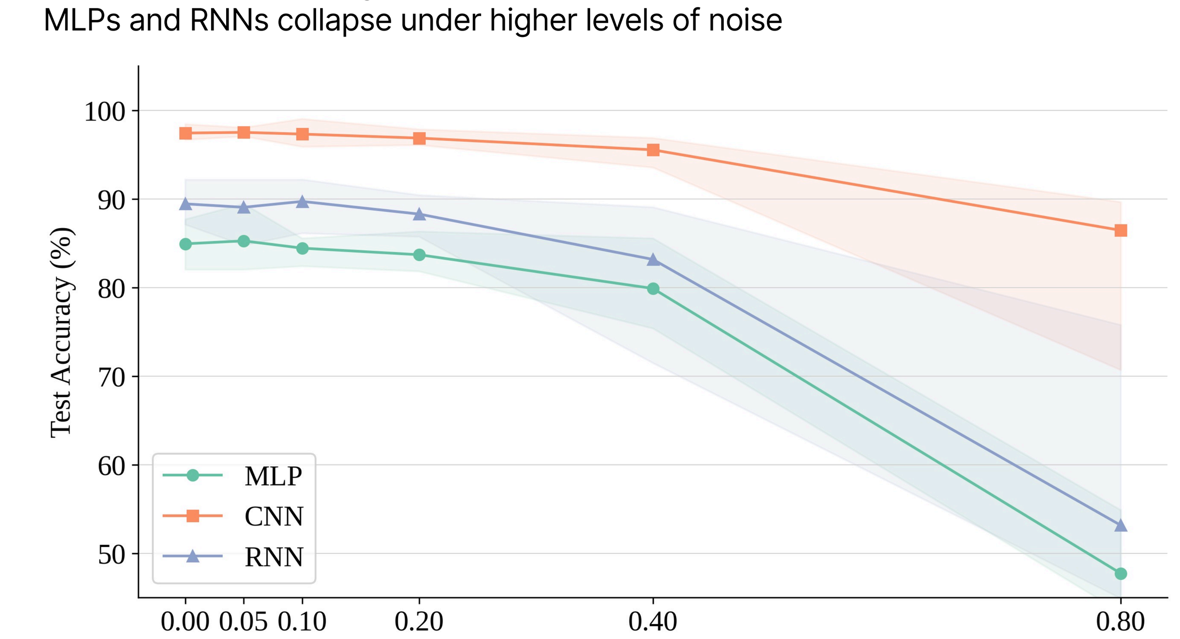
RNNs and **MLPs collapse** under higher levels of **noise**

Consistent classification gap is present, Parallel Diagonal Dataset represents a CNN problem archetype

Classification Accuracy across *sizes*
MLPs and RNNs converge with 30 000 training samples



Classification Accuracy across *noise*
MLPs and RNNs collapse under higher levels of noise



5. Conclusion, Limitations & Future Work

Conclusions
- Parallel Diagonal Dataset establishes a CNN problem archetype empirically
- Model competition is an effective method for constructing archetypes

Limitations & Future Work
- Residual signals allow MLPs & RNNs to converge with enough samples → refine archetype to weaken residual signals
- Unclear if real dataset characteristics map to problem archetypes → validate with real datasets
- Fair translation between domains remains unclear → scope to a single domain or determine translation fairness