

# Semi-Supervised Siamese Network for Identifying Bad Data in Medical Imaging Datasets

Niamh Belton<sup>1,2</sup>, Aonghus Lawlor<sup>3,4</sup> and Kathleen M. Curran<sup>1,2</sup>

<sup>1</sup>Science Foundation Ireland Centre for Research Training in Machine Learning

<sup>2</sup>School of Medicine, UCD

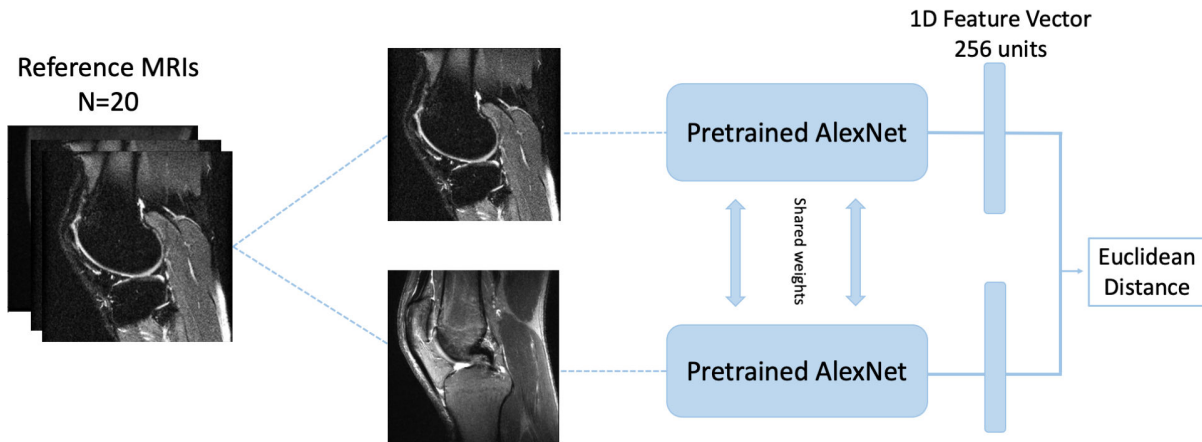
<sup>3</sup>School of Computer Science, UCD

<sup>4</sup>Insight Centre for Data Analytics, UCD, Dublin, Ireland

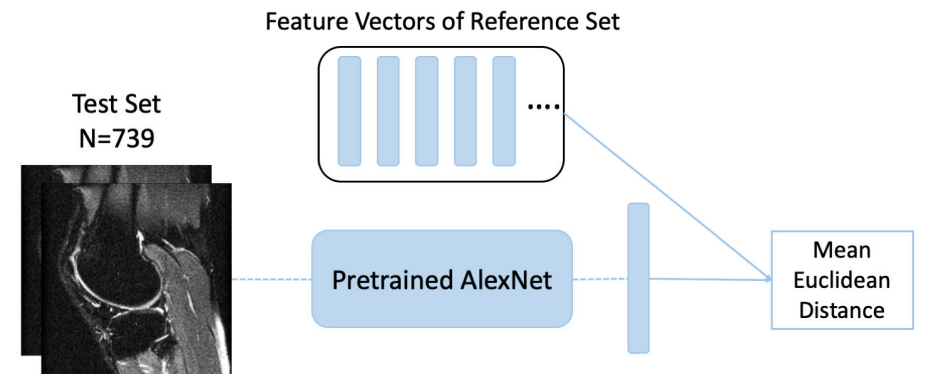
# Proposed Method

**Objective:** Develop a pre-processing technique to identify bad data that could harm the model's training performance in future analysis.

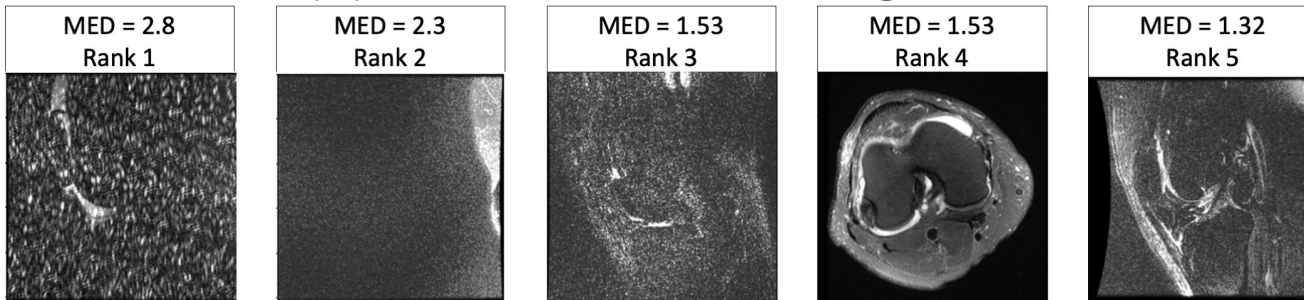
## (i) Training the Siamese Network



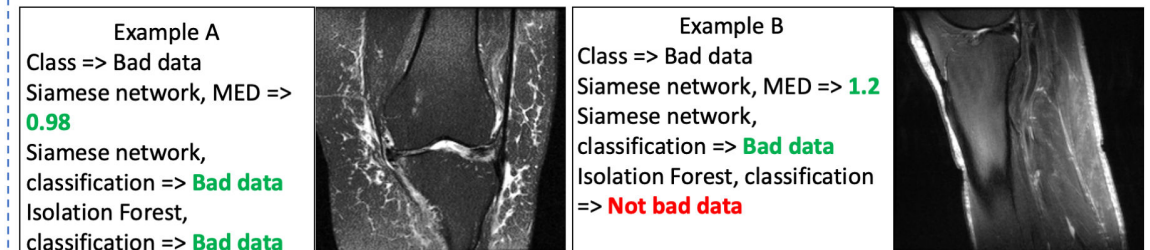
## (ii) Testing



## (iii) Mid-Slice of Cases with Largest MED



## (iv) Additional Bad Data Examples



# Model Performance

- Threshold chosen based on the largest Euclidean Distance between reference MRIs.

	AUC	Sensitivity	Specificity
<b>Siamese Network (proposed)</b>	0.989	100%	89%
<b>Isolation Forest</b>	0.802	71%	92%

## Advantages

- Achieves good performance.
- Identifies a wide variety of bad data.
- Requires only a fraction of the training data that previous methods require.
- Less tedious labelling process in comparison to other semi-supervised techniques.