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

Niamh Belton[1,2], Aonghus Lawlor[3,4] and Kathleen M. Curran[1,2]

[1]Science Foundation Ireland Centre for Research Training in Machine Learning

[2]School of Medicine, UCD

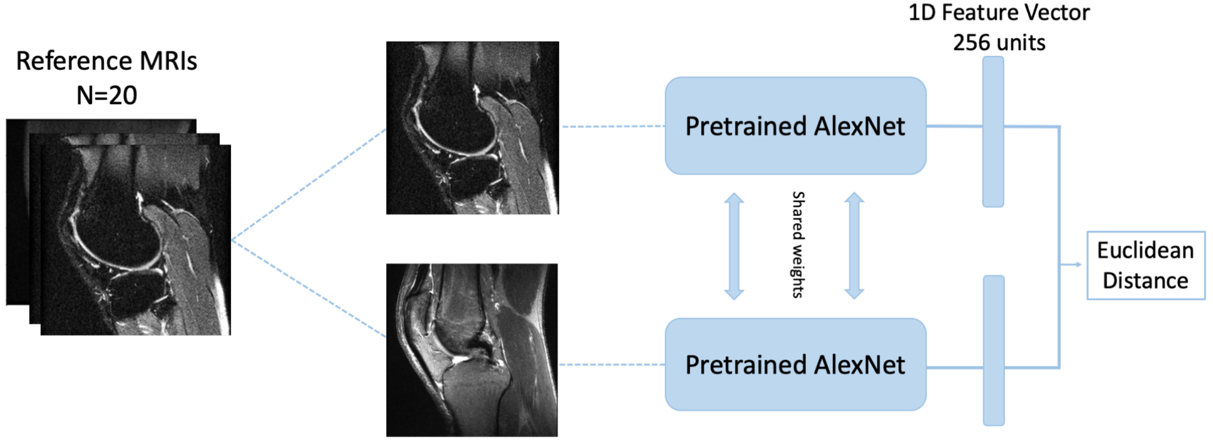
[3]School of Computer Science, UCD

[4]Insight Centre for Data Analytics, UCD, Dublin, Ireland

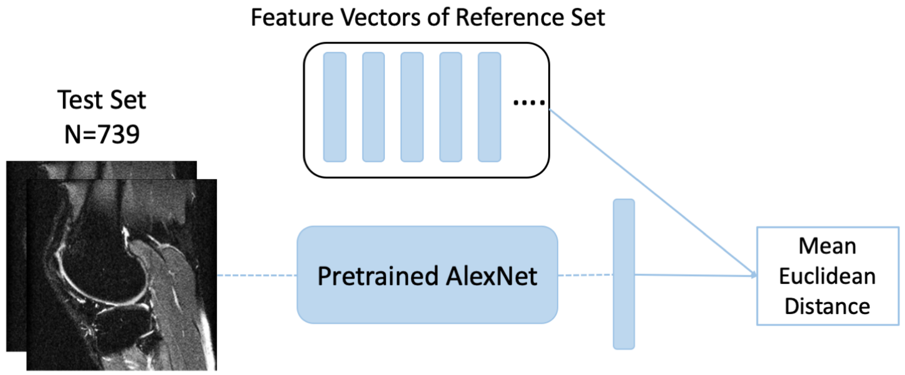
Proposed Method

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

(i) Training the Siamese Network

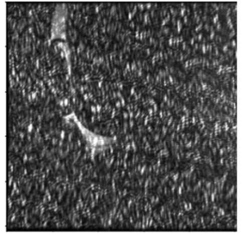


(ii) Testing

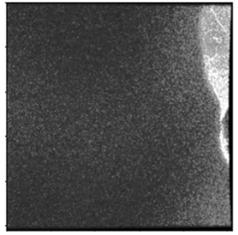


(iii) Mid-Slice of Cases with Largest MED

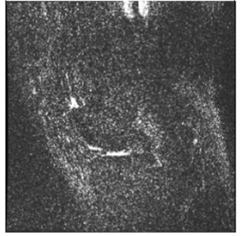
MED = 2.8
Rank 1



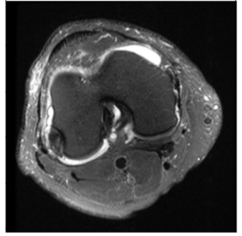
MED = 2.3
Rank 2



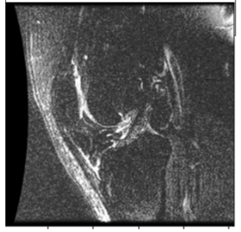
MED = 1.53
Rank 3



MED = 1.53
Rank 4

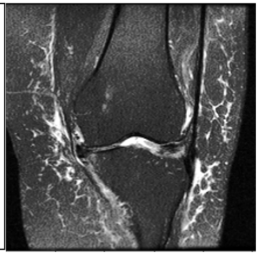


MED = 1.32
Rank 5

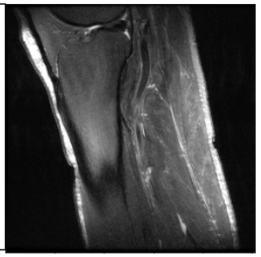


(iv) Additional Bad Data Examples

Example A
Class => Bad data
Siamese network, MED => 0.98
Siamese network, classification => Bad data
Isolation Forest, classification => Bad data



Example B
Class => Bad data
Siamese network, MED => 1.2
Siamese network, classification => Bad data
Isolation Forest, classification => Not bad data



Model Performance

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

	AUC	Sensitivity	Specificity
Siamese Network (proposed)	0.989	100%	89%
Isolation Forest	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.