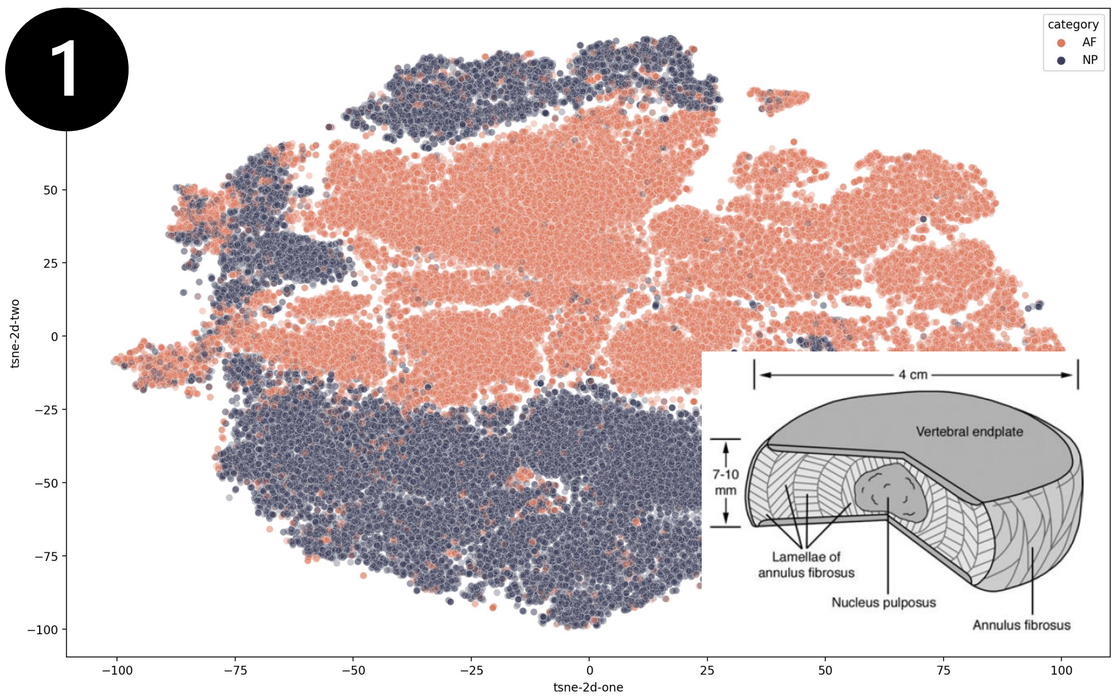
HSLU Hochschule Luzern

Technik & Architektur

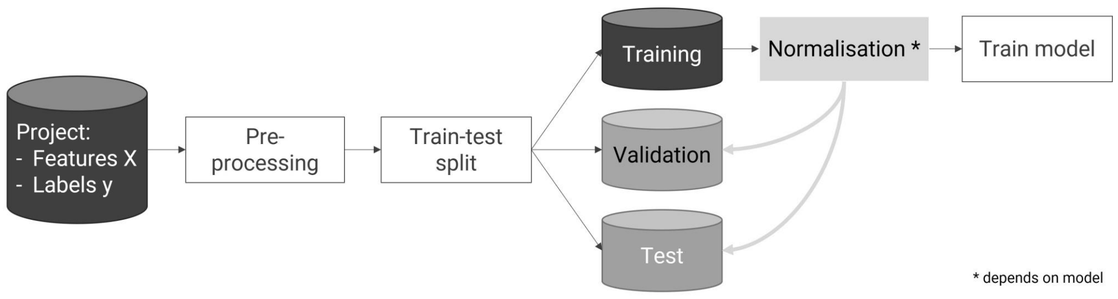
Bachelor thesis medical engineering

Intervertebral disc cell classifier

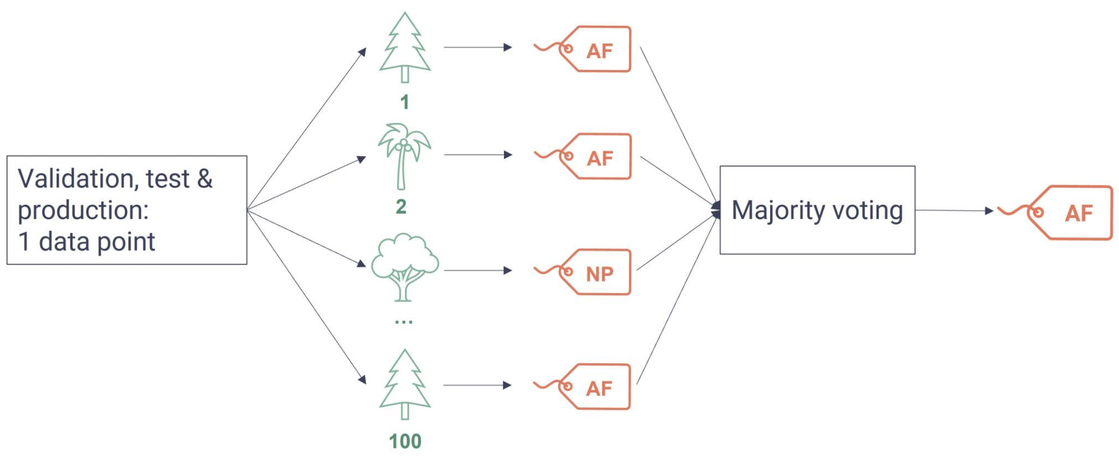
1



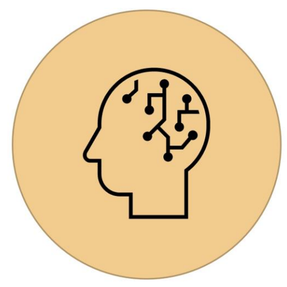
2



3



4



Accuracy of models in %

Dataset / model	RF2	RF3	RF4	RF5	RF6	NN2b	NN5	NN8	NN9
	Random forest					**Neural network**			
Validation RF	99.4	99.5	99.5	99.5	99.2	-	-	-	-
Validation NN	-	-	-	-	-	89.9	88.6	88.9	88.4
Test subset	99.1	99.1	98.8	99.1	98.6	95.0	96.6	95.1	95.2
Test	99.4	99.2	99.1	99.3	98.8	94.9	96.5	94.9	95.2
Bulk	50.0	58.3	66.6	50.0	75.0	100.0	100.0	100.0	100.0

Aim of the project

Lower back pain in many cases leads to early retirement from work or even to disability. One cause of lower back pain is intervertebral disc degeneration. As research on human tissue is more difficult to get approval for, intervertebral discs from bovine calf tails often serve as a model. Calió et al. (2021) performed a single-cell RNAseq analysis of the nucleus pulposus (NP) and the anulus fibrosus (AF) cells in the animal model. The data collected during this research project is now to be used to create a machine learning classifier. The machine learning model should be able to reliably assign cells to the anulus fibrosus or the nucleus pulposus based on their RNAseq data.

Solution concept

In this project, two distinct machine learning approaches, namely random forest and neural networks, have been employed and compared. The rationale behind selecting these two approaches stems from their fundamentally different working mechanisms.

The data provided by Calió et al. (2021) of approximately 97 000 cells was pre-processed, labeled and then split into different datasets. The random forest models, which are based on decision trees, were developed using the Python library scikit-learn. The neural network models were implemented with PyTorch. The models can be fed single-cell RNA data of bovine intervertebral discs and predict if the cells belonged to the anulus fibrosus or the nucleus pulposus.

Results

In total 14 models with different hyperparameters have been trained. A selection of these models has been evaluated not only on the validation set but also on the project test set. Additionally, a small independent test set consisting of bulk RNA data was used to investigate the generalisation of the models. The metric used to determine the model's performance is accuracy which indicates the percentage of correctly classified datapoints.

Overall, the random forest models performed better on the single-cell RNA data than the neural networks. However, the neural networks had a better performance on the bulk RNA data.

Reference: Calió, M., Gantenbein, B., Egli, M., Poveda, L., & Ille, F. (2021). The cellular composition of bovine coccygeal intervertebral discs: A comprehensive single-cell RNAseq analysis. *International journal of molecular sciences, 22(9)*. https://doi.org/10.3390/ijms22094917

Image credit IVD: Raj, P. P. (2008). Intervertebral disc: Anatomy-physiology-pathophysiology-treatment.

Tina Salvisberg

Instructor
Prof. Dr. Fabian Ille

Expert
PD Dr. Philipp Stämpfli, University of Zurich UZH

Industry partner
Prof. Dr. Marcel Egli, Institute of Medical Engineering IMT HSLU