Lucerne University of
Applied Sciences and Arts

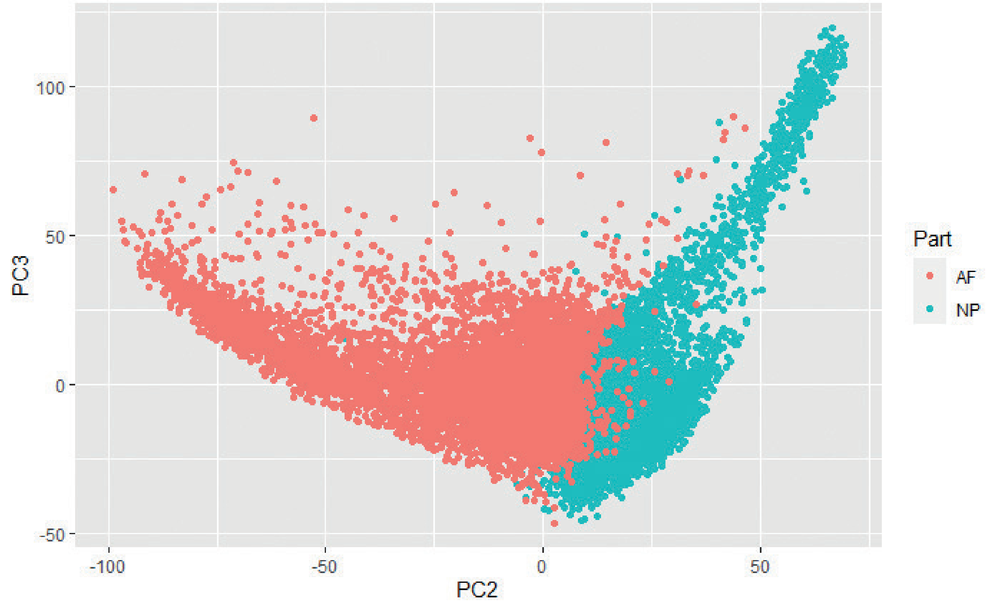
HOCHSCHULE
LUZERN

Technik & Architektur
FH Zentralschweiz

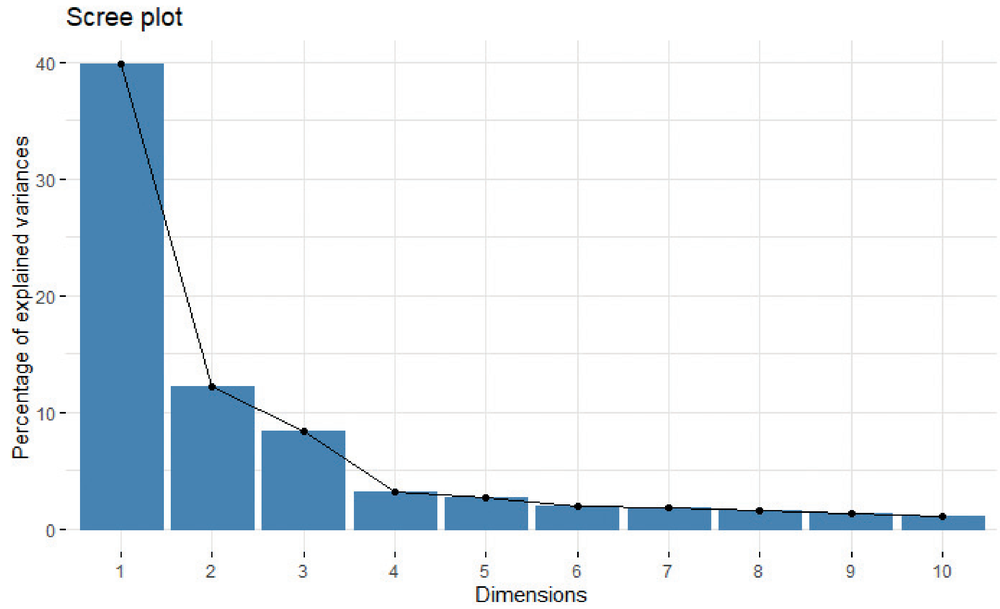
CC Bioscience and Medical Engineering
Obermattweg 9
6052 Hergiswil

Bachelor-Thesis Medizintechnik

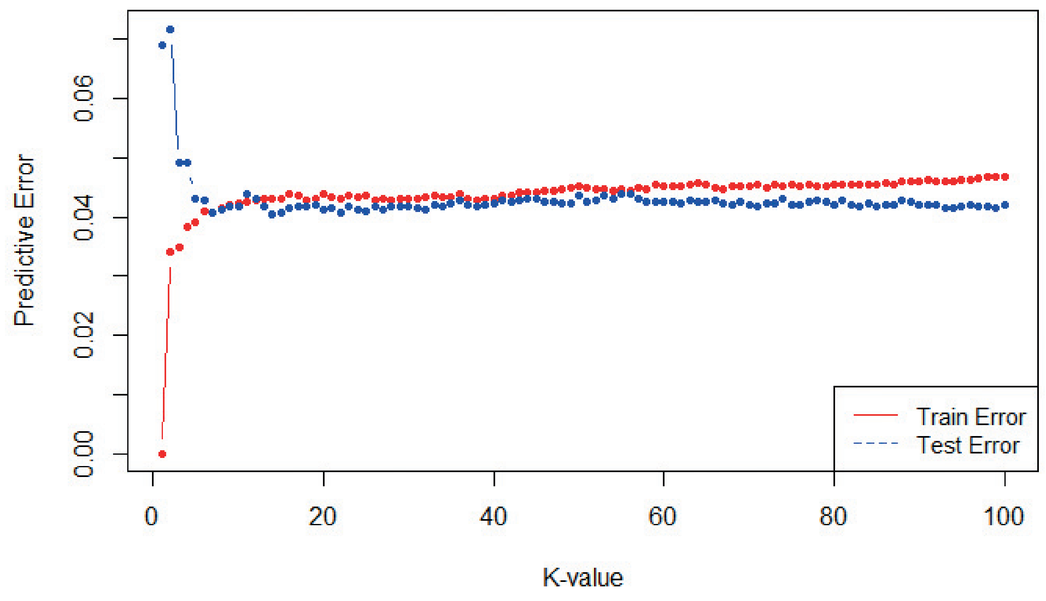
Validation of a Machine Learning Classifier based on Transcriptomics Data



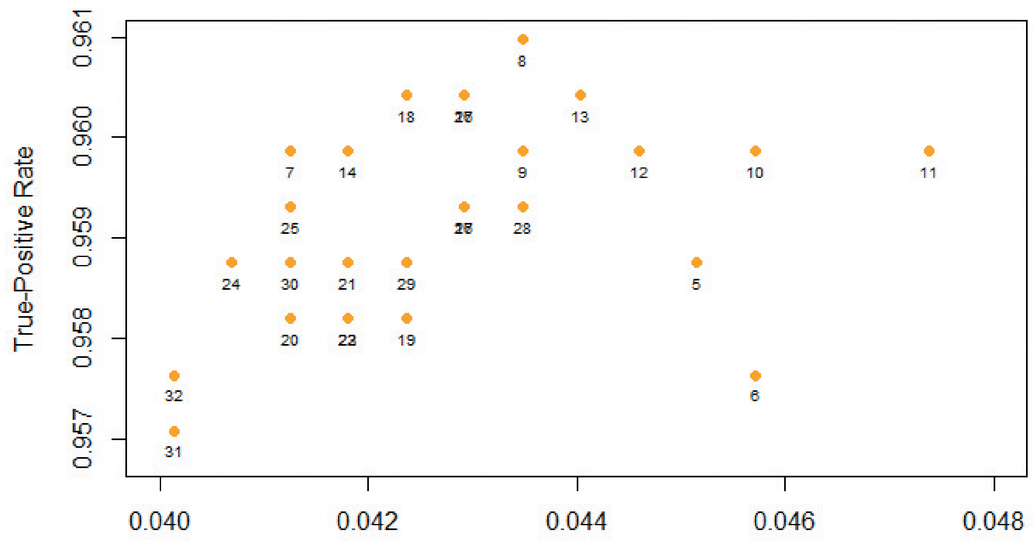
Principal component analysis for the training set



Scree plot showing explained variances after PCA



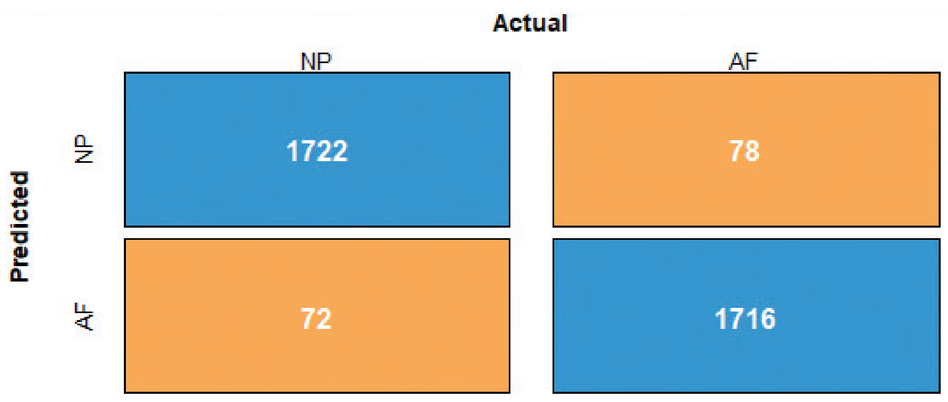
Learning Curve Analysis for Model fit



TPR vs FPR Analysis

Sensitivity	Specificity	Precision	Recall	F1
0.96	0.957	0.957	0.96	0.958

Accuracy	Kappa
0.958	0.916



Performance metrics for selected k = 9 model and confusion matrix (right). Here, positive class = NP and negative class = AF.

Background

This bachelor thesis deals with the validation of the k-nearest neighbour algorithm, involved in binary classification of real transcriptomics data. In a previous work, using artificial data, the classifier achieved perfect classification metrics. The ultimate goal is to develop a universal classifier, capable of detecting early stages of osteoarthritis. Transcriptomics data show information about the gene activity in a tissue sample, which are stored as counts of mRNA transcripts in a dataset.

Methodology

Initially, the project involved collection of data from published repositories. Due to major data scarcity in single-cell transcriptomics, a realistic dataset from bovine intervertebral discs (IVD) was used. The samples represented individual cells extracted from either of the two tissue types of the IVD: Nucleus pulposus (NP) and Annulus fibrosus (AF), which serve as the class labels for binary classification.

The pareto validation involved splitting the dataset in a 80:20 ratio, which followed Z-Score Normalization (Standardization) for attaining normal distribution of data and avoiding bias. This is a requirement for principal component analysis in reducing dimensions of the data, avoiding the curse of dimensionality and to efficiently use the available computation power.

The k-nearest neighbour algorithm uses the euclidean distance metric for calculating the distance between the points, which is specified by the hyperparameter k number of neighbours. Plotting the accuracy for different k-values gives a fundamental base for validation, however, comparing the true-positive (TPR) and false-positive rate (FPR) for different k-values helped to select k-values that showed higher TPR and lower FPR. Further, a learning curve was plotted between the training and testing error rate to gain an understanding about the model fit. Finally, the holdout validation, using an additional validation set for hyperparameter tuning, proved the applicability of the findings from the pareto validation.

Results

The pre-processing methods such as standardization and principal component analysis helped to reduce the bias and increase classification efficiency from an early stage. From the TPR vs. FPR plot, higher k-values proved to be better choices at the cost of increased computation time whereas the learning curves showed that k = 9 perfectly fits a good balance. Any value below that overfitted the model and after k = 41, the model transitioned to an underfit, leading to eventually an increased gap between the two learning curves and becoming more of an unrepresentative result. The holdout validation showed similar results using the same validation methods, where k = 9 with the euclidean distance metric showed an overall sensitivity of 96%, specificity of 95.7% and an F1 Score of 95.8%.

Noel Roy Palmgrove

Supervisor:
Prof. Dr. Fabian Ille

Expert:
PD. Dr. Philipp Stämpfli

Industrial Partner:
Prof. Dr. Marcel Egli