# Written text

A good overview of all the techniques used can be found in a survey article (96) explaining all the algorithms used, as well as several others.

Similar research has been done prior to this one. Numerous algorithms have been proposed for feature selection (87, 88, 89, 90, 99, 106) and simple performance tests are done to do research on both these and other algorithms (19, 91, 92, 94, 105). Most of these focus on a much smaller number of features than in this research however, not testing the algorithms for data with more than 1000 features. Other feature selection research focus more on one type of the feature selection, being wrapper methods (104, 107, 108) or embedded methods (113).

A select number of articles focused on bigger feature set sizes. One research looked at bigger not necessarily biomedical datasets and tried to use support vector machines for feature selection, one of these datasets also being the Arcene dataset (128). A similar research focused on feature selection for a microarray dataset of 7000 and showed that feature selection is beneficial (NEW). Another research looked at 35 datasets with a range of features between 37 and 50000 and what type of feature selection works best (NEW). They used five advanced algorithms and therefore harder to understand, whereas the ones used in our research are more intuitive. Another very similar research evaluated feature selection algorithms, too, but did not have the focus on the number of features.

Much research is done on feature selection algorithms, as can be found in the related work. This work was made for using feature selection for biomedical data however, containing several benefits:

* The main goal was to create a feature subset as small as possible with a performance as high as possible, which only rarely was a goal in other research (114). This resulted in both a score being made that incorporates both the feature subset size and the performance as the FS\_score and the usage of this score to evaluate the usefulness of every feature. While not being a main goal, the computation time is also stored for additional investigations on that.
* Feature selection is done on very big biomedical datasets of more than 1000 features, therefore the feature selection has a big influence. This is not often done, whereas these datasets will become more available as well as investigated in the future.
* Only basic feature selection algorithms are used so other users can easily understand the basic principles and many different ones are investigated, making this work a very extensive search for well performing feature selection algorithms. Moreover, these algorithms are also very different in their feature selection, also showing the benefits and drawbacks of every type of algorithm.
* At last automation of feature selection techniques is added, too. Automated machine learning is a very interesting approach to extracting information out of a dataset. Properly implementing feature selection in automated machine learning opens it up for high dimensional datasets, as well.

# Related work

115 Recursive svm feature selection and sample classi\_cation for

mass-spectrometry and microarray data

Similar to my research with other algorithms

# Contributions

Main differences to all these works:

* I look at computation time (Hardly ever done)
* I look at automation of the techniques. (Never done for biomedical world)

# Related work

## Algorithm proposals – Mainly for a smaller number of features

87 Planar arrangement of

high-dimensional biomedical data sets by isomap coordinates

88 A novel feature selection approach for biomedical data

classi\_cation,

89 Feature selection for high-dimensional data|a pearson redundancy

based \_lter,

90 Minimum redundancy feature selection from microarray gene expression

Data

99 Higher criticism thresholding: Optimal feature selection when useful

features are rare and weak

106 A new maximum relevance-minimum multicollinearity

(mrmmc) method for feature selection and ranking

## Simple performance tests

19 Data complexity assessment in undersampled classi-

\_cation of high-dimensional biomedical data,

91 Investigating the e\_ect of dataset size, metrics sets, and feature

selection techniques on software fault prediction problem

92 A comparative study on feature selection and classi\_cation

methods using gene expression pro\_les and proteomic patterns

94 Investigating the e\_ect of dataset size, metrics sets, and feature

selection techniques on software fault prediction problem,

105 Automated feature set selection and its application to mcc identi\_cation in digital mammograms

for breast cancer detection

## Better (non-biomedical dataset references)

128 Combining svms with various feature selection strategies

The only one that looks at 1000+ features. Comparing how well SVM performs before and after feature selection and shows that it differs per dataset, but some have improved performance. (Very well cited too, 880) times

NEW Feature Selection for High-Dimensional Genomic Microarray Data

This article only tries to find out if feature selection is possible by using a dataset of 7000 and one type of feature selection and says this is possible.

NEW A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data

A good one to read. It focuses more on known ways to do it and not on fundamental algorithms like I did. But looked at a lot of datasets with similar ideas. Also focusing a little bit on micro-array but not much.

95 Feature selection algorithms: A survey and

experimental evaluation,

Very similar too, but they do not talk about the number of initial features (and graphs are of terrible quality)

96 A review of feature selection techniques in bioinformatics

Shows a good overview of all of the techniques

104 Massively-parallel feature selection for big data

Shows that wrapper methods take a very long time for high dimensional data

107 new gene selection approach

based on minimum redundancy-maximum relevance (mrmr) and genetic algorithm

(ga),

Reasonable number of features and extension of wrapper method

108 Minimum redundancy maximum

relevance feature selection approach for temporal gene expression data,"

Similar to 107, mrmr tests

113 Feature selection in proteomic

pattern data with support vector machines

Interesting article for proteomic data using embedded svm

114 Mining mass

spectra for diagnosis and biomarker discovery of cerebral accidents,

Uses the same idea as me (finding a small subset of features that can explain the disease)

115 Recursive svm feature selection and sample classi\_cation for

mass-spectrometry and microarray data

Similar to my research with other algorithms

# Contributions

Main differences to all these works:

* I compare basic feature selection techniques, not advanced algorithms (some do this too, but for other data)
* I compare a lot of basic feature selection techniques (hardly any that look at this many different ones)
* I look at much higher numbers of features (usually <1000, some <10000, only a couple <50000) and biomedical data (only some micro-array or mass spect)
* I look at computation time (Hardly ever done)
* I look at automation of the techniques. (Never done for biomedical world)
* I propose a score that combines number of features and quality