\section{Motivation for machine learning}

For the breath samples measured, an analytical study based on p-values was performed to predict and to categorize the asthmatic and healthy samples by \cite{exhaled breath profiling}. This study did not get any definitive results for prediction. Both \citeauthor{current breathomics} and \citeauthor{limits of p-values} state their believe that methods more complicated than just using the p-values could give a better results. Finding the molecules which differentiate between a healthy and non-healthy subject, and predicting diseases of a patient are both mentioned as possibilities. Since neither of these sources applies machine learning to actually predict disease, and \cite{a review of volatiles} mentions that the application of machine learning performed on breath data is not widespread, an attempt to do just so is made. The analysis is performed on 70 absorbance spectra from healthy patients, and 70 absorbance spectra from asthmatic patients.

\section{Machine learning for classification}

Algorithms performing classification by machine learning do so by looking for pattern in the data of the different samples. Since for the current set of samples it is known whether they originate from healthy or asthmatic patients, supervised learning methods are applied. This means the machine learning algorithm is trained to recognize patterns of healthy and asthmatic samples by feeding it a part of all samples. In this case 98 of 140 samples are used to train the algorithms and the other 42 to test the accuracy of the algorithm.

\section{Classification by trial-and-error}

Since the scope of this project does not allow for learning the ins and outs of machine learning and the construction of an algorithm specifically made for these datasets, existing algorithms are used. Based on an extensive study by \citeauthor{current breathomics} of what kind of algorithms are most useful in the field of breath analysis, the focus initially lies on support vector machines (SVMs). A website specifically for uploading datasets, uploading machine learning algorithms, and running these algorithms on the datasets is used\cite{mlcomp}. This website expresses the rate of success in a single number, the error, which is the amount of falsely classified samples divided by the total amount of samples. Since most algorithm run in less than 5 minutes and they can run simultaneously, many different algorithms can be tested in a short amount of time. Aside from the complete set of absorbances of 140 samples over the wavelength range of 832 to 1263 $cm^{-1}$, various subsets of this are tested as well. Among these are the set of 36 wavenumbers for which the p-values are below 0.005, the set of concentrations of 167 compounds as derived from the 140 absorbances, and the set of 30 compounds with the highest concentrations as derived from the 140 absorbances.

\section{Results}

Results seem to generally be best for algorithms such as nearest neighbor, support vector machines, and rotation forests. A trend is also seen in that the larger data sets get lower misclassifications. The smallest errors are found for the unmodified data set of all absorbances over the entire wavelength range, as seen in \autoref{table}.

\begin{table}[h]

\centering

\caption{Best error rates for different algorithms run on the entire dataset of absorbances of 70 healthy and 70 asthmatic patients. Names of algorithms are as stated on the website\cite{mlcomp}.}

\begin{tabularx}{\textwidth}{|c|l|X|}

\hline

Error rate & Name of algorithm & Description of algorithm \\ \hline

0.167 & RBFNetwork\\_weka\\_nominal & Normalized Gaussian radial basisbasis function network \\ \hline

0.190 & sgd-hinge-stepsize0.5-iter5 +Chi2 & Stochastic gradient descent \\ \hline

0.190 & svmlight\\_multiclass-linear & Linear support vector machine for multiclass classification \\ \hline

0.190 & svmlight\\_multiclass & Support vector machine for multiclass classification \\ \hline

0.190 & IBk\\_weka\\_nominal & K-nearest neighbours classifier \\ \hline

0.190 & IB1\\_weka\\_nominal & Nearest neighbours classifier \\ \hline

0.214 & RotationForest\\_weka\\_nominal +Chi2 & Rotation forest \\ \hline

0.214 & svmlight-linear & Support vector machine for binary classification \\ \hline

\end{tabularx}

\label{table:ML}

\end{table}

The similarity of the error rates from \autoref{table:ML} can be explained by the small test set of 42 samples. The various nearest neighbour algorithms function quite similarly, and the same can be said for the support vector machines. As the results from \autoref{table:ML} are the best from a total of 41 different algorithms, they corroborate the use of SVMs.