Serum tumor markers for diagnosis of lung cancer using machine learning techniques

Project Computational Biology 8P320

Supervisor: Eduati, F.

Eindhoven, 22-4-2019

De Vries, R. 1004141

Abstract

Research has indicated the benefit of the early detection and treatment of lung cancer, as it provides patients with a better 5-year survival probability. The diagnosis of lung cancer is besides that invasive and thus stressful for the patients, who are also at risk for surgical complications. This study therefore evaluates seven blood serum tumor markers (carcinoembryonic antigen, carbohydrate antigen 15.3, squamous cell carcinoma–associated antigen, cytokeratin-19 fragment, neuron-specific enolase, pro–gastrin-releasing peptide and human epididymis protein 4) and their possibilities for the diagnosis of lung cancer and classification of cancer type. Previous research has shown the potential of threshold-defining algorithms and machine learning classifiers for similar cancer diagnostic tasks. This study will therefore include thresholds from the literature, as well as optimization algorithms for the definition of thresholds. The following machine learning models will also be included: Decision Tree Classifier, Logistic Regression, Supported Vector Machines, Naïve Bayes Classifier, Random Forest Classifier and k-Nearest Neighbour. Evaluation of these algorithms with cross-validation is done to assess the performance of each of the approaches. This has shown the potential for the tumor markers in combination with certain models for several clinical applications. First of all the defined thresholds, which used bootstrap for the optimization of the precision, that have shown to be useful for confirming the presence of lung cancer in patients. The classification of NSCLC has shown to be most accurately with the 5-Nearest Neighbour classifier, whilst classifying SCLC with high probabilities is best done with the random forest model. Only the exclusion of lung cancer has proven to be not viable. The classification of SCLC versus NSCLC did rely on a oversampling algorithm (SMOTE), as the labels in the dataset were highly imbalanced. It can therefore not be ensured that similar performances will be displayed in a clinical setting. Testing this as well as evaluating the performance of the algorithms for an increased number of patients and tumor markers will be needed in future studies, before implementation of the models in the clinic takes place.

Introduction

Cancer is shown to be the second leading cause of death worldwide, with an estimated 9.6 million deaths in 2018 according to the World Health Organization (WHO, 2018). Lung cancer [LC] is responsible for most of these deaths and therefore the most fatal cancer with 1.76 million deaths in 2018 alone (WHO, 2018). These cases of LC can be split into two types of cancer: small cell lung cancer [SCLC] which is accountable for 10-15% of all lung cancers and non-small cell lung cancer [NSCLC]. The survival rates for these two LC types differ since SCLC patients have a 5-year survival rate of 6.5%, while NSCLC patient have a 22.1% chance of survival over a 5-year period (Dayen, et al., 2017).

The diagnostic process for LC often starts with an initial evaluation when there is reason to believe a patient has LC or when a patient is at risk of having LC. The first tests consist of history and physical examinations, including chest radiography. Even if these tests show no sign of LC the doctor will often follow up with a contrast-enhanced computed tomography [CT] and positron emission tomography. This is necessary to rule out the presence of small tumors (Sharma, 2015) (Kelly, 2015). If the presence of LC is concluded, the next step will be the diagnostic evaluation which includes the diagnosis of the unhealthy tissue. This tissue diagnosis step aims to obtain an adequate tissue sample, which is imperative for the treatment plan. The acquisition of such a tissue sample can be performed in multiple ways, where the least invasive method is preferred for the patient (Kelly, 2015). An example of the many methods are a bronchoscopy or tissue biopsy (Kajatt, 2013). After this staging will be determined and treatment will be set up.

It is of importance that the presence of LC is found as early as possible as research has shown (Wang-Yu, et al., 2016). Patients with a NSCLC tumor of 1.0 cm or less are reported to have a 5-year survival rate of 95.0%. When a detected NSCLC tumor has a size ranging between 1.1 and 2.0 cm, the patients 5-year survival rate already drops to 87.9% (Baba, et al., 2011). Another study has shown similar results for NSCLC patients with 5-year survival rates of 80.20, 85.07 and 100% for tumors of 1.6–2.0 cm, 1.0–1.5 cm and less than 1.0 cm in diameter respectively (Shi, Zhang, & Han, 2011). Patients with NSCLC tumors with a diameter of 2.0 cm or larger display the same trend, with 5-year survival rates of 76.5% and 57.9% for a diameter of 2.1-3.0 cm and greater than 3.0 cm respectively (Morihito, 2005). It must be noted that all the presented survival rates are focused primarily on patients where the cancer has not yet spread to distant locations in the body.

Current detection methods for LC can be invasive for the patients, while early detection is crucial to optimize the survival chances of the patients. The development of an easy-to-use, fast and accurate method for the diagnosis of LC is therefore essential to reduce LC deaths, while also using a non- or minimally-invasive method to reduce patient stress and prevent complications which can occur during surgery. A technique which can provide information about the patient, while being minimally invasive, is blood sampling and testing. It is also easy to perform and therefore enables frequent check-ups. Machine learning can then be used to analyse the supplied patient data and make a prediction about the patients well-being.

Using blood sampling, to acquire information about the patients current state, allows the minimally-invasive conditions and enables early detection due to the possibility of frequent screenings. Research has shown that the concentration of several serum tumor markers [TMs] extracted from these samples can suffice for the diagnosis of LC. The six TMs which have been identified are: carcinoembryonic antigen [CEA], carbohydrate antigen 15.3 [CA15.3], squamous cell carcinoma–associated antigen [SCC], cytokeratin-19 fragment [CYFRA 21-1], neuron-specific enolase [NSE], and pro–gastrin-releasing peptide [ProGRP] (Molina, et al., 2016). Another TM which has been shown to differentiate LC from healthy patients is human epididymis protein 4 [HE4] (Liu, 2013). This is backed up by another study, which suggests that HE4 qualifies as a TM for men with lung cancer in combination with other tumor markers (Jra, et al., 2014).

The first six markers have been shown to individually associate with LC presence and LC type, NSCLC vs SCLC (Molina, et al., 2008) (Molina, et al., 2009), while a combined assessment also shows promising results (Molina, et al., 2016). These studies have used empirical thresholds and basic approaches to classify the data. However, a more complex and data-driven approach such as machine learning has also proven to be useful in this context. For example by using the mutations in cell-free DNA as input for the models for LC classification (Aliferis, 2002). Over the past decades cancer research has also shown varying results when using machine learning for the diagnosis of different types of cancer or predicting patient survival (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2014). This study will apply similar machine learning techniques for the classification of LC, while also defining empirical thresholds to perform classification. These two approaches can then be compared to evaluate whether machine learning algorithms can outperform the current state of the art techniques based on empirical thresholds.

The Molina (2016) paper has already shown the capability of classifying LC for a large dataset (N=3,144) by defining threshold levels (Molina, et al., 2016). However, this research will use a dataset with 225 patients (N=225) and is thus significantly smaller. It is therefore not certain that the dataset will show a clear difference in the distributions of the TMs for the two labels. Furthermore, the small collection of data might not yet be able to capture the general patterns which are present in reality and thus a large dataset (Tipton, Hallberg, Hedges, & Chan, 2017). Predictions can hence not be made if the defined thresholds and their optimization will provide similar results.

The aim of this research will be to evaluate the use of the seven serum tumor markers for the diagnosis of lung cancer and cancer type. To do this multiple methods will be set up, starting with the unsupervised approach of clustering and histograms. This is to get an insight into the coherence of the data and the distribution of the labels for each TM. After this threshold levels from the Molina (2016) paper will be used for each TM, where concentrations above the threshold are indicated as abnormal. When considering the combination of TMs, one abnormal TM level will suffice to indicate the presence of LC. The next step will be the implementation of algorithms, which will focus on the optimization of different metrics, to find optimal threshold values. The final approaches to assess the diagnostic potential of the TMs are focused on machine learning algorithms which include: Decision Tree Classifier, Logistic Regression, Supported Vector Machines, Naïve Bayes Classifier, Random Forest Classifier and k-Nearest Neighbour. The performance of each of the approaches will eventually be evaluated and comparisons will be made. However, this will be explained more thoroughly in the Materials and Methods section of the report.

To put the results into clinical perspective two cases will be assessed: Can the algorithm help to diagnose LC with high certainty, while also minimizing the number of false positives? Furthermore, can the algorithm exclude LC in patients with a high probability and thus minimize the number of false negatives?

Materials and Methods

Patient data

The patients who participate in this study have been referred to the Catharina Hospital Eindhoven, where the diagnosis of LC has taken place using standard clinical workup procedures. Blood samples have also been obtained before treatment has taken place, from which the TM concentrations could be determined. Age and smoking history have also been registered. This procedure has resulted in a dataset of 225 patients (N=225). 129 of these patients have been diagnosed with LC, while LC has been excluded in 72 patients. LC has not been excluded nor confirmed in the remaining patients. Splitting the patients by cancer type displays 122 patients which have been diagnosed with NSCLC, while 13 patients where confirmed to have SCLC. Data for cancer type is not available in the remaining patients and these will thus be excluded when the classification process will focus on cancer type. This current study is still taking place and it is expected that the dataset will include 1000 patients by mid-2020.

Statistics

To enable the comparison of the methods their performances will be presented with the following statistics: Positive Predictive Value [PPV], Negative Predictive Value [NPV], sensitivity, specificity and Area-Under-Curve [AUC]. The approach of defining thresholds will use the entire dataset to optimize the thresholds and then calculate the statistics. The machine learning classifiers will also use the complete dataset for the model fitting. However, cross-validation is used as well to provide a mean and standard deviation for the statistics. Additionally, a Receiver operating characteristic [ROC] curve will be displayed when fitting the classifiers to the entire dataset. The precision metric, which is equal to the PPV, is also used for one of the optimal thresholding methods. Calculating these statistics will be done with the following formulas:

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

With TP=True Positives, TN=True Negatives, FP=False Positives and FN=False Negatives.

Clustering and histograms

The initial step will focus on the visualization of the data with histograms for each TM to show the distributions of the labels. The distributions shown in the histogram for each label can already help to predict the performance of thresholding methods, as an overlap of distributions will complicate the optimization of this method since there is no clear cut-off concentration.

A basic unsupervised approach, to separate the patients by label, will also be used to evaluate whether this can already distinguish the patterns in the data. This will be done with a heatmap of the TMs with clustering of the rows, where each row is a patient, by optimizing the Euclidian distances. This clustering algorithms does thus not take the labels into account, but only the values of the features. The concentration of each separate TM column will also be normalized to zero mean and unit variance. This method provides an indication whether a complex algorithm needs to be used for the classification process or whether a simple method like this is already able to do so. A positive result for this method is indicated by clear clusters of the different labels, which thus indicates that the normalized TM concentrations of each label are closely related to each other. This label can be LC versus no LC or SCLC versus NSCLC.

Optimal thresholding

The next method for the classification process will focus on the definition of thresholding levels, where TM concentrations above or equal to the threshold are indicated as abnormal in case of the LC classification. When the classification process focuses on cancer type, the threshold level serves as a split for the two different cancer types: NSCLC and SCLC. To evaluate the performance of this method it has been chosen to evaluate the statistics on the individual level of each TM and for the combined assessment of the TMs. This implies that each TM with its corresponding threshold is used to classify the category of choice without considering the other TMs. Statistics will determine the performance of each of the individual TMs. After this all the TMs will be used in a combined assessment, where one abnormal concentration is considered to indicate LC or to indicate the positive cancer type over the defined default cancer type. To elaborate on the cancer type classification, this implies that one cancer type is fixed as the default cancer type. In this study the default cancer type will be SCLC and concentrations above the thresholds will indicate the other cancer type which is NSCLC.

The last part which is essential for this method is to define the threshold values, which will be done using multiple approaches. Each of these approaches is unique and will thus be described in the subsections below.

Molina paper

The first set of thresholds will be obtained from another research with similar TMs. All the TMs, except of HE4, were included in this study. This TM will therefore be left out of the classification process when using these thresholds. This study also considered the same method for the classification, as concentrations above the threshold levels were classified as abnormal. A combined assessment has also been included with a comparable methodology.

The threshold for each of the TMs are fixed to: CEA, 5 ng/ml; CYFRA 21-1, 3.3 ng/ml; SCC, 2 ng/ml; CA15.3, 35 U/ml; NSE, 25 ng/ml; and ProGRP, 50 pg/ml (Molina, et al., 2016).

Optimization of AUC

The second approach of defining thresholds is the optimization of the AUC by calculating the AUC for a set of discrete threshold values. A collection of threshold values will be set up for each TM, where the range of the TM concentrations and the threshold set are the same. Calculating the AUC value for each of the thresholds in this set will provide a continuous curve of the AUC. Localizing the maximum of this curve will provide a threshold. This procedure will be done for each of the TMs, after which these thresholds are used in the classification process.

Optimal thresholding with train and test split

Another set of thresholds can be produced using a training and test split of the data. Splitting the data into a training and test set will be done using cross-validation and specifically Stratified K-Fold with K=10. This ensure that each set contains approximately the same percentage of samples of each class as the complete set (scikit-learn, 2018). This assures that the cancer type SCLC is also present in each of the sets, which is not guaranteed with normal cross-validation due to its small sample size. Using this cross-validation method will return 10 folds containing training and test index, which can be used to loop over and repeat the methodology of this approach. The training set consists of 80% of the data, which will be used to generate a curve for each TM, which is identical to the previous method. The optimal threshold, and thus the cut-off concentration with the maximum AUC, will likewise be determined for each of the individual TMs. The thresholds from this loop will then be tested with the 20% test data by calculating the AUC. This will result into 10 optimal thresholds for each TM, which do not have to be unique, each with a corresponding AUC value. These values will be shown in a scatterplot. This is to assess whether the defined thresholds depend on the used data or whether they are clearly defined regardless of the composition of the dataset.

Optimal thresholding with Bootstrap

The final approach to define thresholds will be based on the resampling method known as Bootstrap. With bootstrap, at each iteration we select N random patient with replacement, where N is equal to the number of patients. Therefore, a patient can be included multiple times and it can occur that other patients are not included in one of the loops. This will be used to create multiple datasets. Each of these datasets will be used to calculate one of the following metric: AUC or precision; for each threshold from the discrete set of thresholds. This will be done for multiple datasets which thus result in a collection of statistical values for each of the thresholds. The means and standard deviations of the collections will be calculated for all threshold values. This enables the visualization of a curve which depicts all the thresholds versus the mean metric value and the corresponding standard deviation.

Machine learning algorithms

The previous classification options have been focused on a simple cut-off concentration, referred to as a threshold. However, more complex algorithms have been developed which are more data-driven and are known as machine learning algorithms. These algorithms rely on finding patterns and inference in the data using mathematical models (Witten & Frank, 2005). Many of these models have been developed from which a few will be used in this paper.

Because the models are so data-driven it is possible to add other features to the list of inputs. In this case the age of the patients and their smoking history will also be used as a feature. The category of the smoking history has three options: the patient has never smoked, has smoked in the past or is still smoking. Normalization of all these features will be done when preparation of the data is taking place. Patients will also be removed if any of the data is incomplete.

Evaluation of the models will take place using two approaches. The first approach will consists of training, also known as fitting, the models to the full dataset. These classifiers will then be used to assign probabilities to each labels of the same full dataset, after which the classification process is evaluated.

The second method for the evaluation of the classifiers will use cross-validation. The specific cross-validation technique that will be used is Stratified Shuffle Split. This method returns stratified randomized folds and thus ensures that each fold contains the same percentage of samples for each class compared to the full dataset (scikit-learn, 2018). The chosen split will be a 80%/20% split for the training and validation dataset, which will be used to calculate the mean and standard deviation of the defined statistics. This will be done by looping over all the folds that have been made, which are 100 folds in this case. Each loop will use the corresponding training data to fit the classifier, after which predictions are made for the validation set. These predictions, which are presented in the form of probabilities for each label, are then evaluated. Doing this for a number of splits will return a list of statistics for each fold. The mean and corresponding standard deviation of each statistic will be calculated and returned. It must be noted that due to the imbalance of the cancer type it might occur that a statistic could be returned as Not a Number. In this case the statistic will be set to 0 and will be incorporated in the list, which will thus effect the final mean and standard deviation. This is done since a Nan value for the statistic often occurs due to the division with 0, which can only be the case if the classification process has had a low performance.

Oversampling

As mentioned before the data for the cancer type is highly imbalanced which can cause problems for the classifiers. It can therefore be beneficial to artificially generate extra data for the class with the lowest sample size using an oversampling method. The method which will be used is Synthetic Minority Over-sampling Technique [SMOTE]. The samples which are artificially generated are a linear combination of two samples from the same minority class ($\mathbf{x}$ and $\mathbf{x}^R$) and are defined as:
$s = x + u \cdot (x^R - x)$ with $0 \leq u \leq 1$; $\mathbf{x}^R$ is randomly chosen among the 5 minority class nearest neighbors of $\mathbf{x}$ (Blagus & Lusa, 2013).

Research has shown that this method can increase the performance of classifiers when the feature space has a low dimensionality (Blagus & Lusa, 2013), which is the case for this study. The cited study also mentioned that under sampling can yield better results, however this cannot be done for this dataset due to the low number of patients which are included. Under sampling would result in such a low sample size that overfitting would always occur.

This technique will be applied for the classification of cancer type due to the imbalanced classes, but also for the classification of LC. This is to evaluate if increasing the number of patients benefits the performance of the models and thus if expanding the number of patients, which are included, can be beneficial for further research.

The machine learning models

The first classifier to be implemented is a Decision Tree [DT] classifier. This model classifies patients by posing a series of questions about the features associated with each patient. Each question is contained in a node, and every internal node points to one child node for each possible answer to its question. The questions thereby form a hierarchy, encoded as a tree (Kingsford & Salzberg, 2008). Classification of a patient is then done following the path and answering each of the questions, which have been set up when training the classifiers. These nodes and questions can be visualized to provide a clear overview of the tree's construction and all of the decision processes. Because of this it is not necessary to normalize the features as it does not influence the outcome. This kind of classifier can be made so specific that overfitting will almost always occur on the training set and it is thus likely that it will perfectly perform when fitting and predicting on the same dataset.

A random forest [RF] classifier has a similar approach as the DT classifier as it construct multiple decision trees for the classification process. Previous research has shown the performance of the DT and RF classifier when predicting LC survivability. These are therefore included in the list of classifiers (Hashi & Almamlook, 2018).

Another research which also focused on lung cancer classification proposed the use of the following machine learning algorithms: DT, k-Nearest Neighbour [KNN] and Support Vector Machines [SVM]. This study focused on the differences in the expression of genes and the possibility to use this information as features. This method has provided significant results when using the KNN or SVM classifier (Aliferis, 2002). These classifiers will hence be included in the set of implemented classifiers.

Two popular classifiers which will also be included to evaluate their performance are a Gaussian Naïve Bayes [NB] classifier and a Logistic Regression [LR] classifier. The LR classifier is also used in the Molina paper (Molina, et al., 2016) and can hence be used as a comparison for when the sample size is lower. The decision has been made to include a NB classifier as well, since this assumes the independence of each of the features which true for the selected TMs. The NB classifier has also shown to achieve a significant performance for other classification problems in the medical field (Witten & Frank, 2005) (Shortliffe & Cimino, 2006).

Code

All of the listed methods and models will be constructed and implemented in Python 3.6 using the Anaconda Navigator and Spyder. Visualization of the results will be done with the use of Excel. The completed code with all of the described functions can be found and downloaded from the following GitHub repository: https://github.com/RyanVries/Project-Computational. This code also shows the hyperparameters for each of the classifiers. The DT, NB and NN classifiers are set to default settings, while the LR classifier uses a l2 penalty and the liblinear solver due to the low sample size. The remaining models are mostly prescribed with the default settings except for the SVM classifier which will present the output as probabilities, while the RF algorithm is fixed to use all features and to set up 200 trees in the forest.

Results

Classification of LC

Clustering and histograms

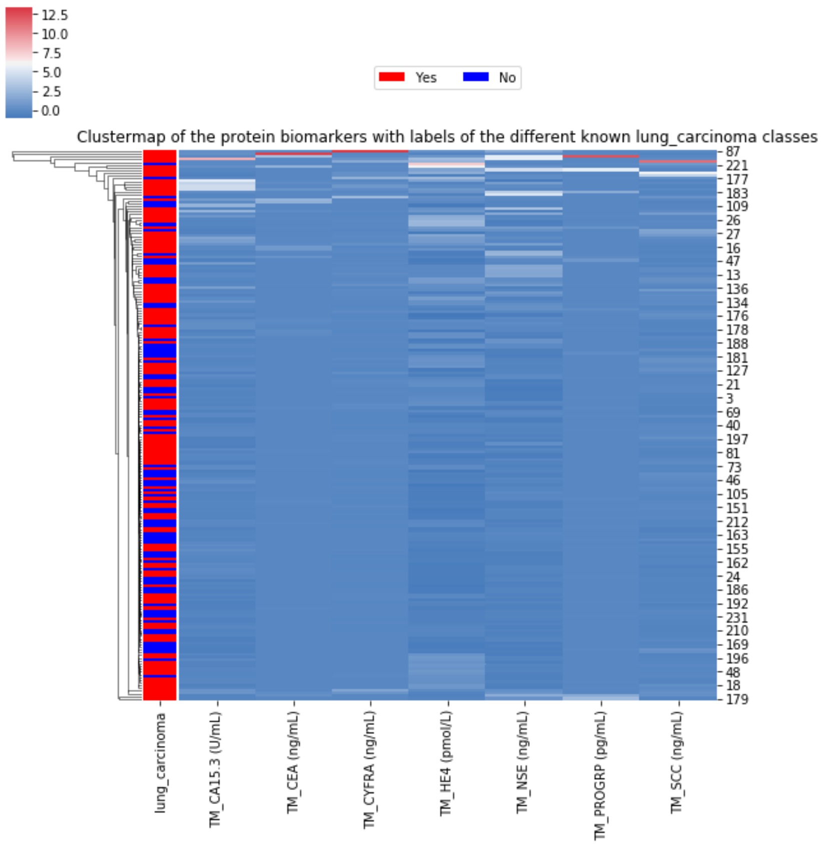


Figure 1 *Clustermap of the TMs with LC labels for the patients*

The first method to be applied is clustering of the data, which is an unsupervised approach for the classification process. Doing this resulted in the heatmap shown in Figure 1, with the legend showing the labels and thus colours for LC or no LC. When inspecting the map the distribution of the labels on the left side of the heatmap do not show clearly defined clusters of the different labels. The heatmap also shows various outliers for some of the TMs.

Visualization of the distribution of the TM concentrations for both labels is done in the histograms shown in Figure 2. The legends of each histogram indicate the label of each distribution and also show threshold values which are of associated with the optimal thresholding methods. When examining the histograms the distributions of the two classes have a significant overlap for all TMs. It is also not possible to distinguish a clear mean and standard deviation for each of the labels if a normal distribution is considered. This can therefore impede the defining process of the thresholds, as there is no visual cut-off point which separates the distributions. The tumour markers CEA, CYFRA and PROGRP also include multiple outliers which substantially increase the range of the concentration axis.

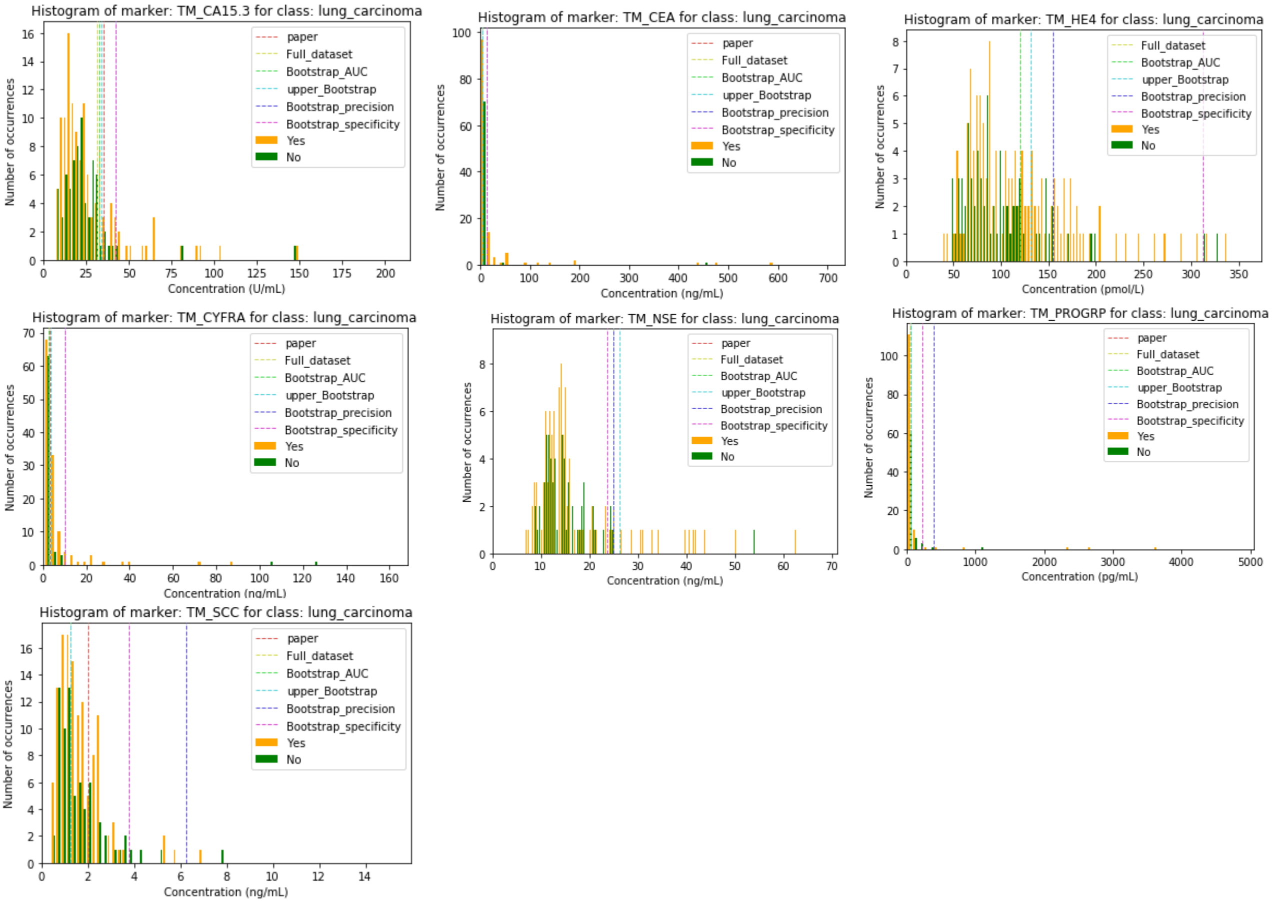


Figure 2 *Histograms of all TM concentrations for the classification of LC, including all thresholds*

Optimal thresholding

The first approach for the supervised classification of LC is the optimal thresholding methodology. The difference in each of these methods is the acquiring of the thresholds. The performances of these approaches are visualized in Figure 3, where each bar plot includes a single statistical value. Only the statistics for the HE4 threshold are missing for the (Molina, et al., 2016) paper, as the paper did not include this TM. The optimization curves which are associated with each of the approaches can be found in Appendix A1, A2 and A3. The thresholds, for each TM, that corresponds with the maximum mean metric value of these curves are captured as a set of thresholds. These threshold values are attached in Appendix A5, while Appendix A4 shows the scatterplots when using the train/test split.

The first bar plot shows that the PPV, when considering the NSE marker, is equal for all methods (96%) and that the number of false positives is therefore minimal. This is due to the single green outlier which is present in the corresponding histogram. The same histogram also shows that the thresholds are chosen after the main 'No' distribution, which ensures only the outlier is classified wrongly. This can also be seen when inspecting the NPV and sensitivity of this marker as these are significantly lower than the PPV. This is because a large number of LC patients are classified as not having LC and the number of FN is thus high. It must be noted that Bootstrap with precision optimization performs better for the PPV and specificity as it essentially minimizes the number of false positives in the optimization process.

A similar trend is presented for the other individual markers where the PPV and specificity are significantly higher than the NPV and sensitivity. This results in the low AUC's which are presented since the focus is too much on not having false positives and more false negatives are hence present.

When considering the combined assessment the PPV's are all above 60% and the PPV of Bootstrap with precision is even 84%. This therefore ensures that most of the positive predictions are correct. The specificity also shows that Bootstrap with precision is much better at minimizing the FP's compared to other methods. However, these values are lower than the statistics for most of the individual markers and more healthy patient are thus classified incorrectly when combining the markers. The combined assessment does show a better result for the NPV and sensitivity relative to the individual markers. Combination of the TMs therefore minimizes the number of false negatives. LC patient are hence less likely to be classified as having no LC, but healthy patients are classified with LC more often. This can be seen when looking at the AUC of the combined assesment, as these are lower for all of the methods compared to the single markers.

The Bootstrap methods also provided a range for each of the thresholds which are listed in Appendix A5. These ranges are determined by taking the maximum of the chosen metric and subtracting the corresponding standard deviation. A vertical line will then be drawn in the optimization curves, which intersects with this point. The closest points to the left and right of this point, which intersect with the line as well, will be captured. These two points show the range of the threshold from which it is 68% certain to contain the true optimal threshold. A large range will thus indicate that it is not possible to define a optimal threshold value for the TM, while a small range shows it is possible to define a clear threshold. Assessing these ranges show a large dispersion for some individual markers. These thresholds do thus highly depend on the composition, which is similar to the dispersion shown in the scatterplots for the train/test split. It is therefore likely that these threshold values in the scatterplot are not reproducible and these will therefore not be used.

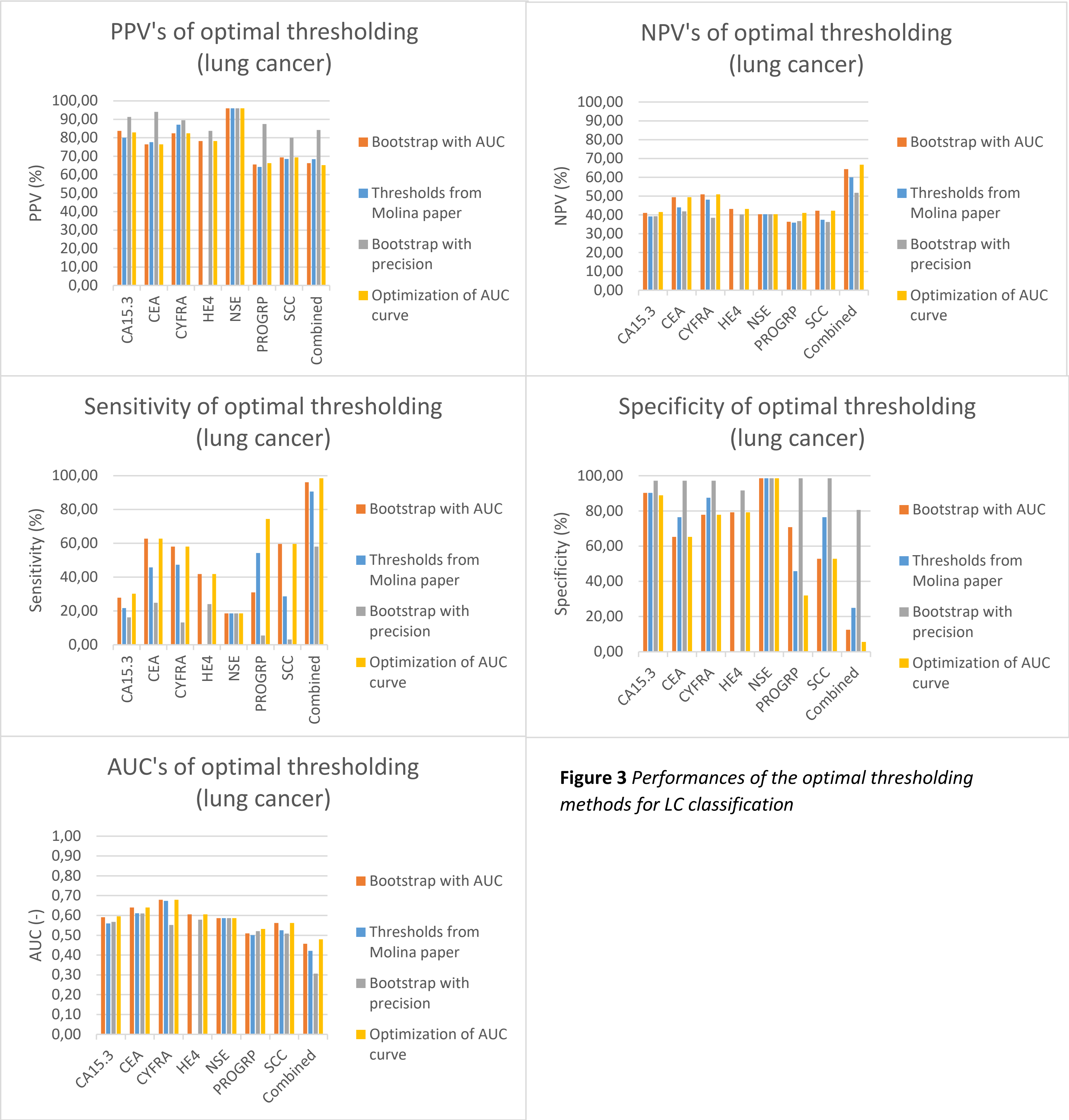


Figure 3 *Performances of the optimal thresholding methods for LC classification*

Machine learning models

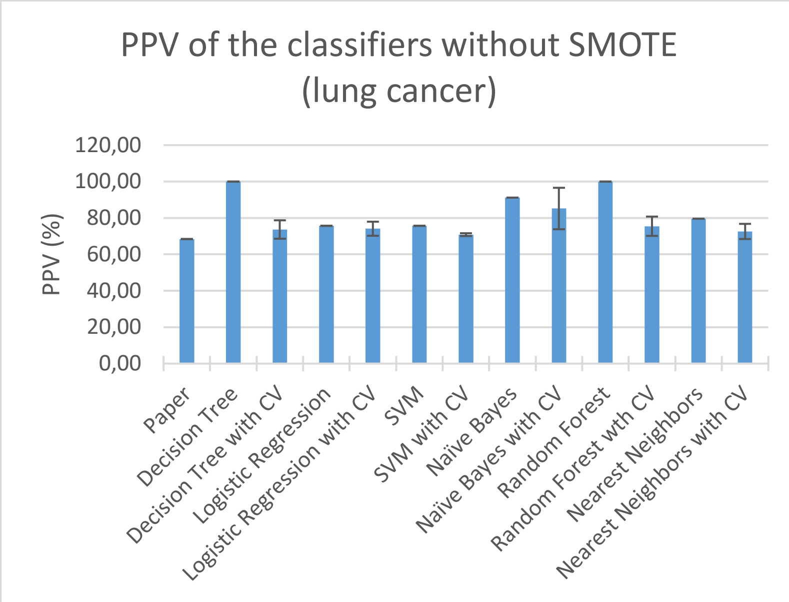
The more data-driven approach of machine learning algorithms are evaluated using the same statistics as the thresholding methodology to enable comparisons. The performances of these models are shown in Figure 4, which also include the performance of the threshold from the (Molina, et al., 2016) paper. The ROC curves for each of the models are attached in Appendix A6. The DT classifier is also visualized in Appendix C1. Figure 4 can be separated in two parts: training on data with SMOTE disabled and SMOTE enabled. Each of the models is then evaluated using training and predicting on the full dataset as well as with the cross-validation technique [CV]. The CV results also include an error equal to the standard deviation and thus shows the 68% interval of the statistic.

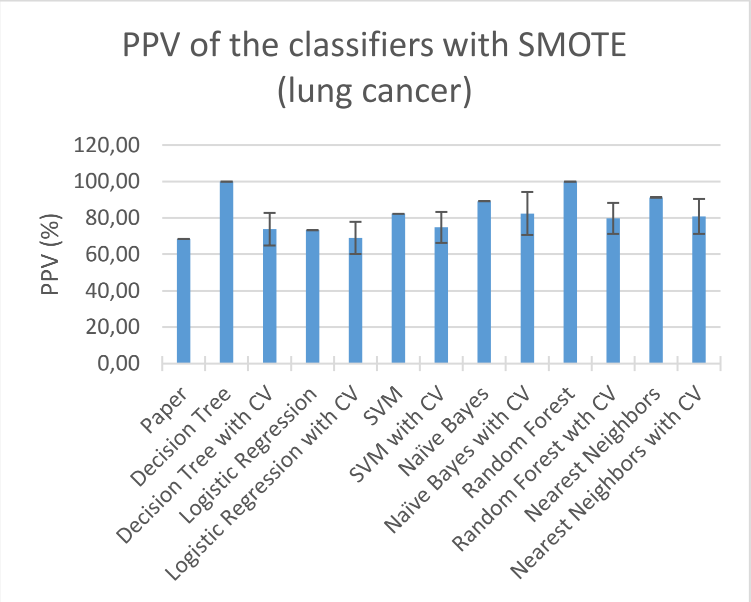
First of all the PPV, which is 100% for the DT and RF classifier when fitted to the full dataset without SMOTE. This shows that these classifiers can easily overfit to the provided dataset, as evaluating with CV drops this statistic to 74% and 76%. Enabling SMOTE does not significantly influence these results for these classifiers. The error is also relatively small compared to the absolute value in both cases. The LR, SVM and NN classifiers show similar values for the PPV with and without SMOTE. Only the training on the full dataset has a reduced performance as the classifiers do not overfit on the data. The NB classifier performs slightly better than the previous three models when trained on the full dataset. Evaluation with CV shows the highest values with 85% and 83% respectively without and with SMOTE of all the models. However, the standard deviation is higher compared to the others. But all in all this NB classifier appears to outperform the other models for the PPV and thus most of the predicted positives are classified correctly. All of these models also perform significantly better than the thresholding approach from the paper which has the lowest PPV.

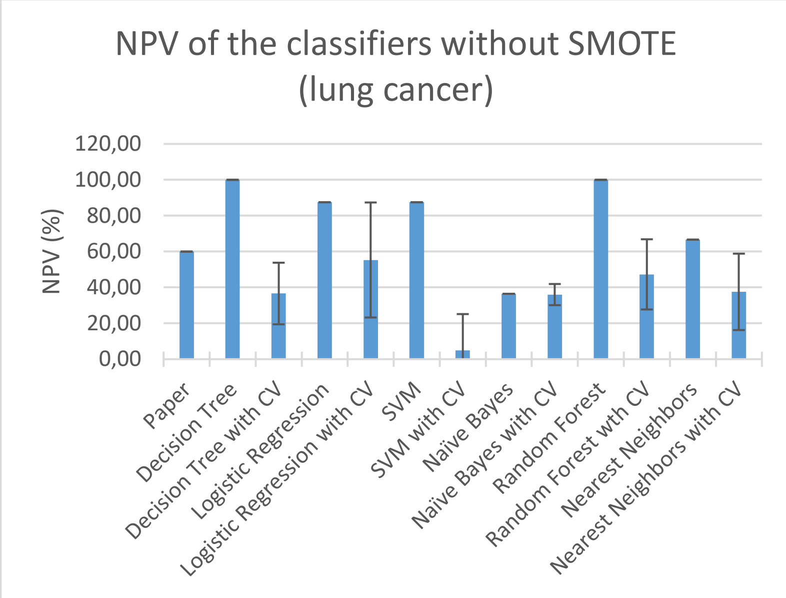
The NPV also shows that the DT and RF classifier overfit when using the full dataset, but they perform worse when using CV as the NPV drops to around 40% without oversampling. The NB and NN models also show poor NPV values in both cases without SMOTE. The SVM does perform well when presented with the full dataset, but has the lowest NPV of all due too many 0 values. This is because of the Nan values. The LR classifier perform the best (55%) when evaluation is done with CV, but this presents a large 68% interval. Enabling SMOTE provides better result for the NPV as the values now range between 60% and 80% with small standard deviations. The best performance is now of the RF classifier with a NPV of 77%. The performances are now also better compared to the thresholds from the paper when considering the NPV. The number of negatives which are classified correctly do thus increase when enabling SMOTE. Furthermore, the LR and RF classifier appear to perform the best with SMOTE respectively disabled and enabled.

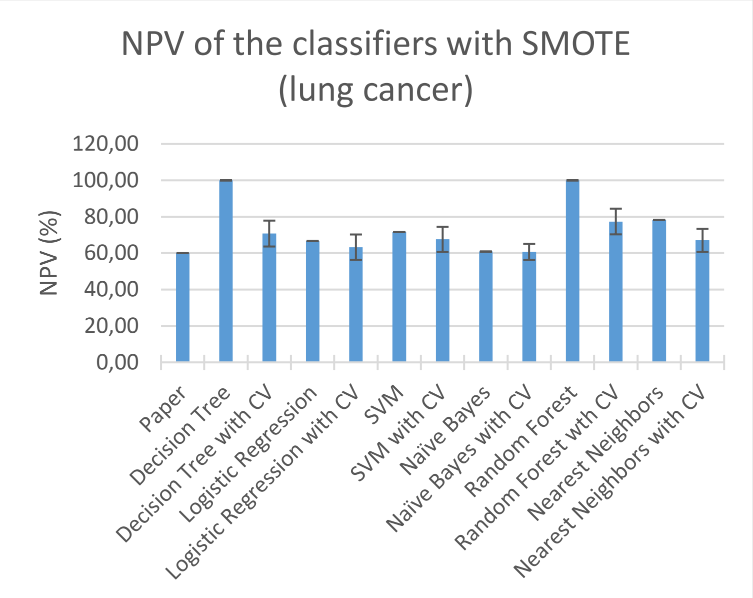
The sensitivities of the models show a different trend, since enabling SMOTE decreases the performance of all of the models. Disabling SMOTE shows that the DT, LR, SVM and RF classifier have an optimal performance when presented with the full dataset. Evaluation with CV shows that the SVM model (sensitivity of 99%) is able to minimize the number of false negatives and thus classify almost all of the LC patients. This is slightly better than the results from the thresholds.

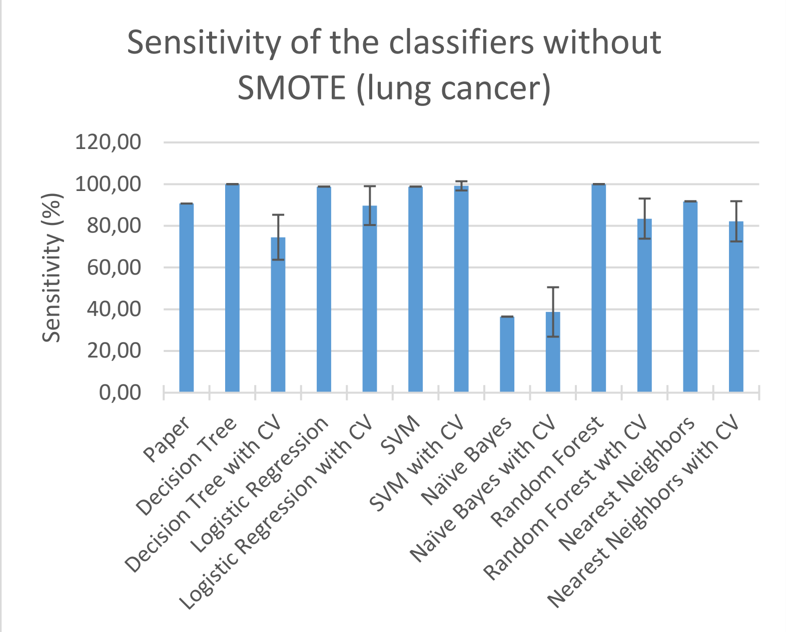
Enabling SMOTE does show to benefit the specificity of the predictions. Only the NB model is able to provide a significant specificity with SMOTE disabled (91% and 83% with CV). Enabling SMOTE still increases these value, but this has a more significant effect on the other models. Training on the full dataset shows the best results for each model, but CV evaluation is only marginally lower with small standard deviations. The NB classifier still performs the best (91%) when considering the CV score and is thus able to classify most of the healthy patient correctly. This is significantly better than the specificities from the optimal thresholding approaches.

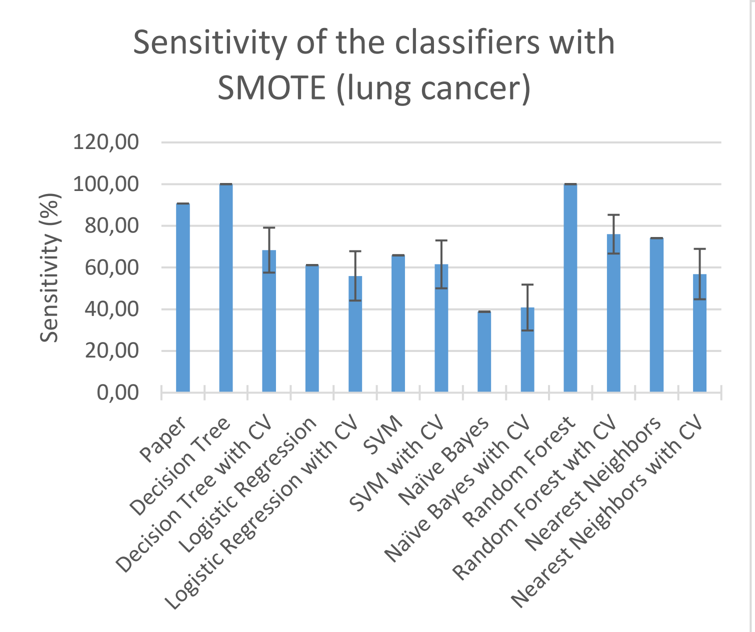

PPV of the classifiers without SMOTE (lung cancer)


PPV of the classifiers with SMOTE (lung cancer)


NPV of the classifiers without SMOTE (lung cancer)


NPV of the classifiers with SMOTE (lung cancer)


Sensitivity of the classifiers without SMOTE (lung cancer)


Sensitivity of the classifiers with SMOTE (lung cancer)

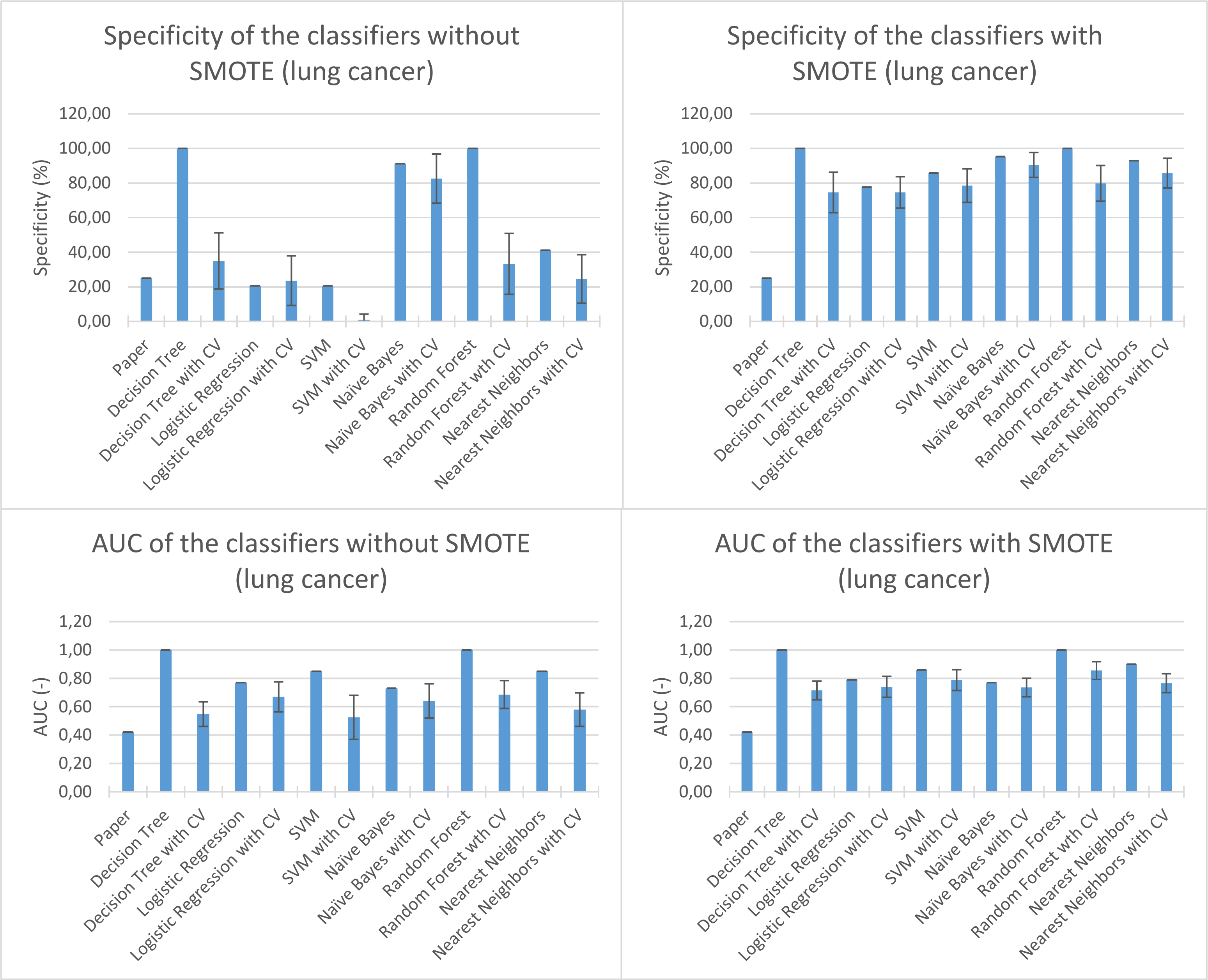


Figure 4 *Performances of the machine learning models for LC classification without and with oversampling*

The AUC of the models also show better performances for the models compared to the paper thresholds. Disabling SMOTE shows that the DT and Rf classifier perform optimally when using the full dataset. The CV score shows a similar trend with the best AUC for the RF model (0.69) and a relatively small 68% interval. Enabling SMOTE increases the AUC of the models, where the best AUC again belongs to the RF model with an AUC of 0.86. The other models also provide significant AUC scores, for both evaluation methods, when compared to the thresholding approach.

Overall the classifiers score significantly better when considering the NPV and sensitivity as statistics and do thus not have the problem of optimization to the positive class. Some of the models still perform better, when considering the specificity on its own, when compared to the threshold performances for the individual markers and the combined assessment. Furthermore, only the PPV for the NSE marker is able to outperform the machine learning models.

Classification of cancer type

Clustering and histograms

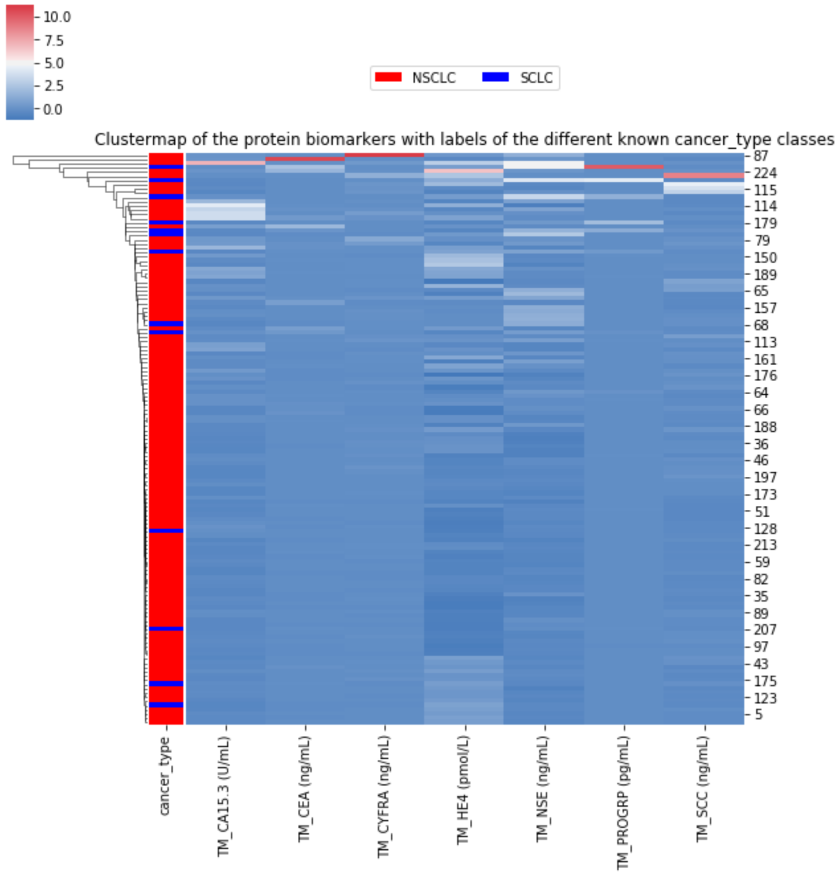


Figure 5 *Clustermap of the TMs with cancer type labels for the patients*

Generating a heatmap for the class cancer type is done likewise to LC classification, which provides the clustermap in Figure 5. The distribution of the labels on the left of the figure do not show clearly defined clusters and the heatmap itself include some outliers for some TM values. The labels on the left also show the low sample size of SCLC in the original dataset.

Histograms have also been generated for the cancer type class and these distributions are shown in Figure 6. The legend once again includes the label of each distribution and threshold values from the optimal thresholding methodology. The markers CEA, CYFRA and PROGRP still show outlier concentrations. The distributions for the two classes also have an overlap for all TMs and the distribution of the class SCLC is considerably smaller. It is therefore expected that the thresholding methods can have problems with finding cut-off concentrations.

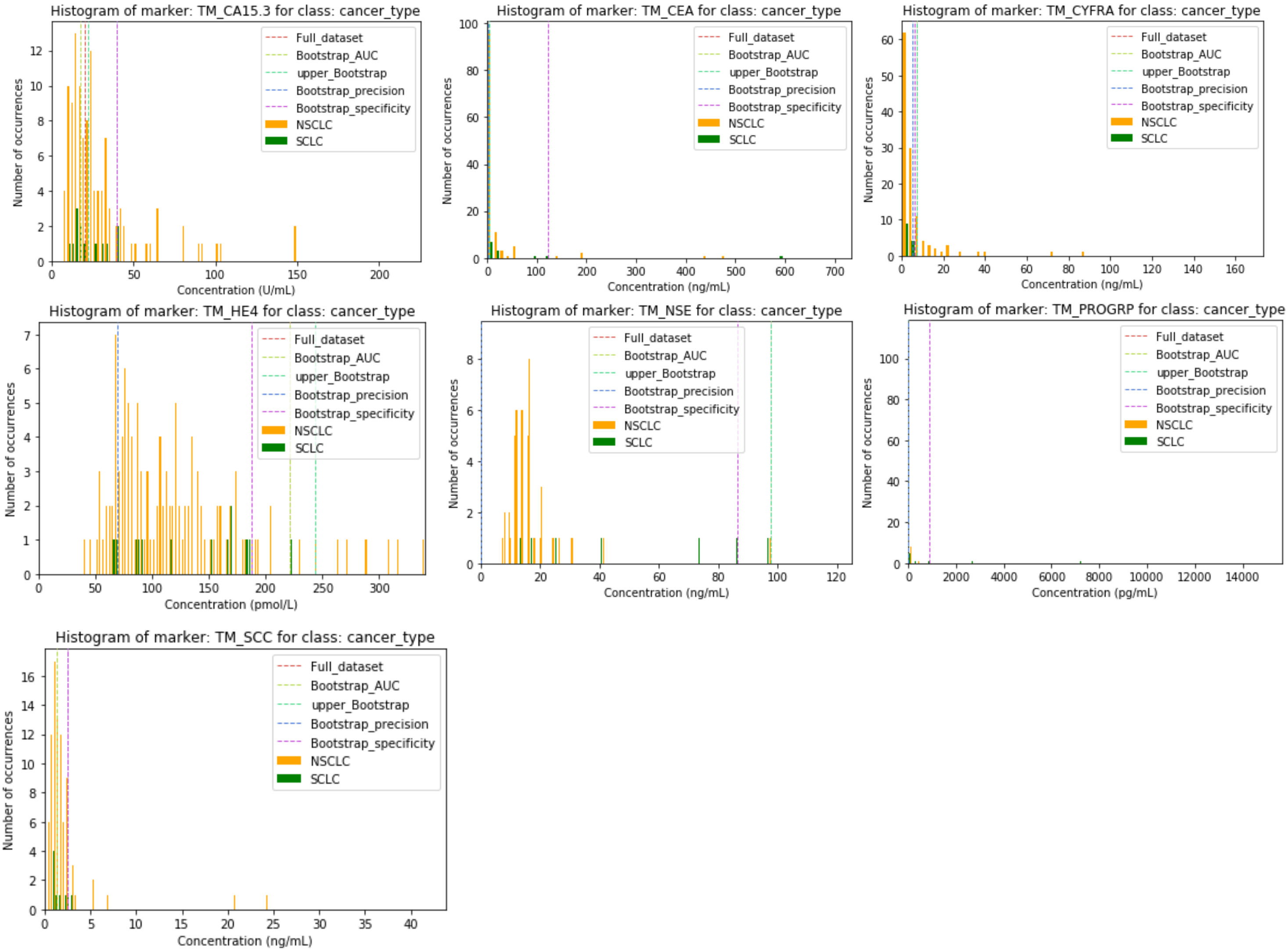


Figure 6 *Histograms of all TM concentrations for the classification of cancer type, including all thresholds*

Optimal thresholding

The first classification technique for the cancer type is the optimal thresholding methodology. The performances of these approaches are visualized in Figure 7. This classification process has not included any threshold values from the literature, as research has not yet focussed on the combination of these TMs for cancer type predictions. The optimization curves which are associated with each of the approaches can be found in Appendix B1, B2 and B3. The defined threshold values are attached in Appendix B5, while Appendix B4 shows the scatterplots when using the train/test split. It must be noted again that the NSCLC class is set as the positive class and the SCLC class thus as the negative class for the entire process.

First of all the PPV values, which all range from 90% to 100% for the individual TMs and the combined assessment. The thresholding methods are therefore able to ensure that most of the positive predictions are correct and thus detect the NSCLC patients. However, the NPV shows a different trend with values ranging from 0 to 15%. The approaches are hence not sufficient to provide reliable negative predictions, since most of those are false negatives. The PRoGRP marker is not even able to correctly identify a single SCLC patient, which reflects in the combined assessment.

The sensitivities also show that the NSCLC patients can be classified correctly when using the combined assessment or the PRoGRP marker. However, the PRoGRP marker never predicts the SCLC label and will thus always be correct for the NSCLC patients. This is not usable in the clinic. What is also interesting is the result for the NSE marker which is 100% when using the precision metric and Bootstrap, but 1% for the other two methods. This indicates that the other methods do provide significantly more false negatives, while the Bootstrap with precision is able to avoid this. This is again due to no SCLC predictions. A similar trend is visible for the HE4 marker. The other two methods do thus rely on the performance of the PRoGRP marker in their combined assessment, as the other marker are not sufficient in correctly identifying the NSCLC patients.

The specificities are in line with the previous results, as the PRoGRP marker has a specificity of 0% as it never predicts the SCLC class. Other markers like CYFRA and SCC are able to correctly classify the SCLC patients. However, due to the results for the PRoGRP marker the combined assessment results in a specificity of 0%. Interesting is the results for HE4, as Bootstrap with precision shows low result while the other two methods have a 100% score.

The AUC of all of the methods are remarkably low due to the problems with the NPV score and thus presenting reliable negative predictions. The CYFRA marker has the highest AUC of 0.61 together with the CA15.3 marker as they have the best performances for the combination of statistics. The NSE and PRoGRP marker have an AUC of 0.50 and their predictions are thus equal to making random guesses. This results in a similar AUC for the combined assessment due to the method of the classification process.

The Bootstrap methods also provided a range for each of the thresholds which are listed in Appendix B5. These ranges are calculated using the same method and again show the 68% interval of the threshold. These ranges show that each of the Bootstrap methods have a big uncertainty for at least one of the markers which reflects in the results of the combined assessment. The thresholds do thus highly depend on the composition of the dataset as the optimization curves show large standard deviations for the metric values. The scatterplots for the train/test split confirm this as this technique was not able to find a well-defined threshold for all of the TMs.

Machine learning models

The previous optimal thresholding methodology is thus able to perfectly predict most of the NSCLC patient by sometimes just not prediction the SCLC class at all. A more complex approach like machine learning algorithms are hence used together with the SMOTE oversampling method to artificially generate additional data. It must be noted that classification without SMOTE cannot be done due to the fact that the dataset only contains 7 SCLC patients when including the TMs, age and smoking history as features. The dimensionality of the feature space is hence larger than the sample size of this class. Using SMOTE resulted in a dataset with 83 patients with SCLC and 84 patients with NSCLC and thus overall a balanced dataset. Evaluation of the models in combination with this method results in the statistics shown in Figure 8. The ROC curves for the models are also displayed in Appendix B6. Furthermore, the complete tree of the DT classifier is visualized in Appendix C2.

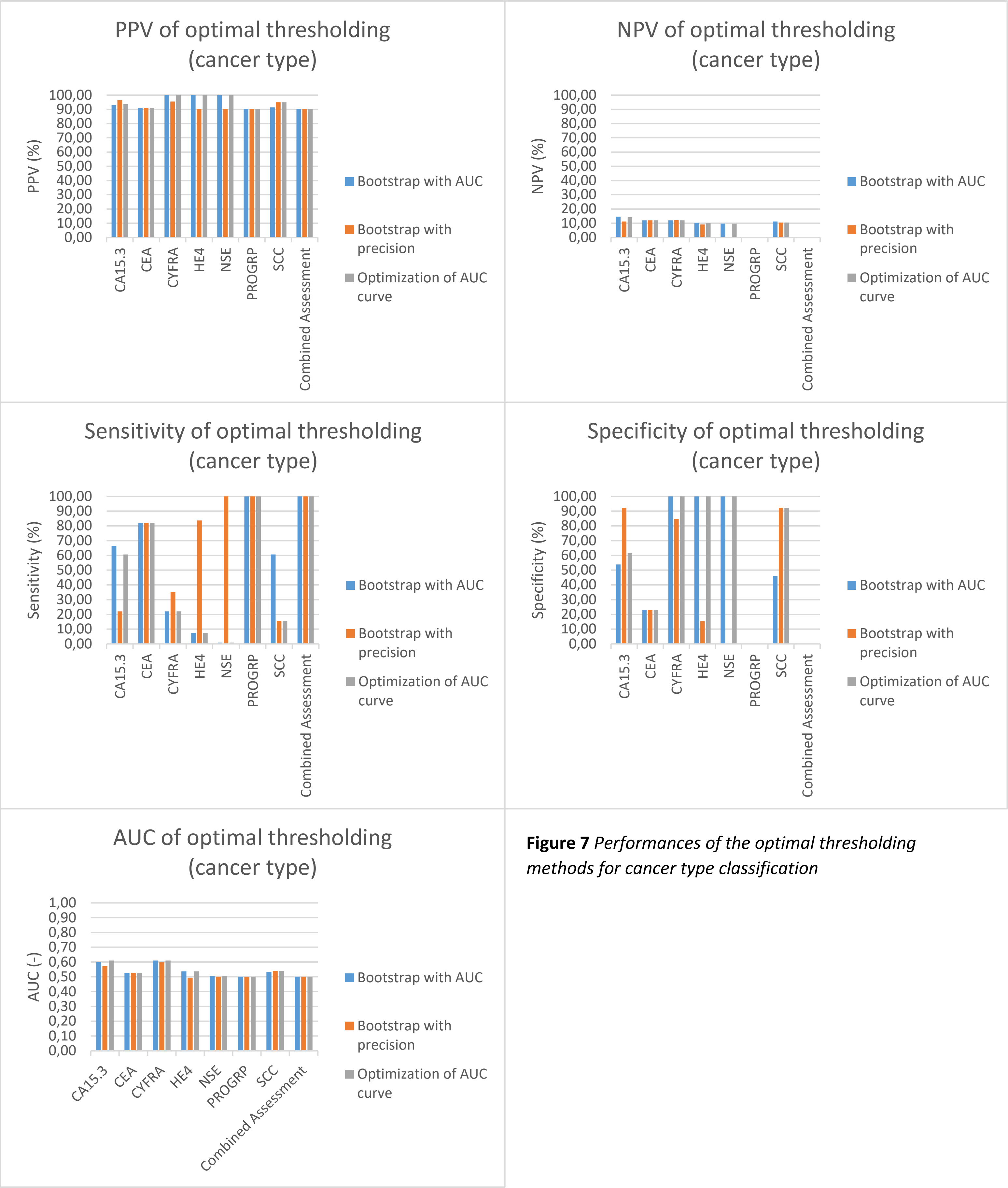


Figure 7 *Performances of the optimal thresholding methods for cancer type classification*

First and foremost the PPV's shown in Figure 8, which are 100% for the DT, RF and NN classifier when evaluated on the full dataset. These are thus perfectly able to perform the classification process without delivering any false positives. The other models have a marginally smaller PPV for the full dataset, with a minimum of 93% for the LR classifier. Evaluation of the CV score affirms these results as the standard deviations are small and the values range between 92% and 100%. The NN classifier scores the 100% and has thus been able to provide perfect NSCLC predictions every fold. The robustness of this classifier is thus excellent.

Evaluation of the NPV also shows optimal results as the values range between 83% and 100%. The DT and RF again are able to fit perfectly to the complete dataset. Only the NN model appears to provide some unreliable negative predictions with a NPV of 87%. The CV score shows a similar result with the NN model having some problems with the SCLC predictions. However, other models are able to perform better together with a high robustness which can be concluded due to the small 68% intervals. The NB algorithms is able to provide the most reliable negative predictions for all of the folds with a NPV of 96% and can thus minimize the number of false SCLC classifications.

The sensitivity shows a similar trend with values ranging between 83% and 100%. The problem of reducing the number of false negatives is again visible for the NN classifier. The other models also shows similar results with the perfect scores for the DT and RF models. The RF, SVM and NB classifier show the best CV scores with a NPV of 96% and can thus recognize almost all of the NSCLC patients in the dataset.

Grading the specificity is presenting comparable performances with a spectrum of scores between 92% and 100%. The DT, RF and also the NN classifier show the maximum result when presented with the full dataset. The CV score also shows the maximum result for the NN model which is significantly higher than the second best model which is the RF with 95%. The NN model is thus able to classify all of the SCLC patient with their corresponding label, while other models make some mistakes for this class.

Combining the importance of each of the statistics into the AUC shows that all models are able to show significant results. Predictions on the full dataset are done perfectly for the DT, SVM, RF and NN. The other models show relatable performances with a minimum of 0.98 for the LR algorithm. The CV score also present high performances with a marginal outlier for the DT classifier with an AUC of 0.91. The best performing models for all the folds are the SVM and RF machine learning models with an AUC score of 0.99. This is also visible in the corresponding ROC curves which are close to optimal.

The machine learning models clearly show better results compared to the optimal thresholding methods. These models do not show the problem of not being able to detect the SCLC patients. However, it must be noted that the models have been assisted by the SMOTE method which presented them with a different dataset compared to the first approach. This does not neglect the fact that the AUC scores of the models (maximum of 0.99) are significantly higher than the AUC scores of the thresholding methodology (0.61). This is also the case for the NPV and specificity when taking the combined assessment and the machine learning models into account.

It must be noted that all of the statistics visualized in the bar plots are also listed in Appendix D, if the specific values are desired.

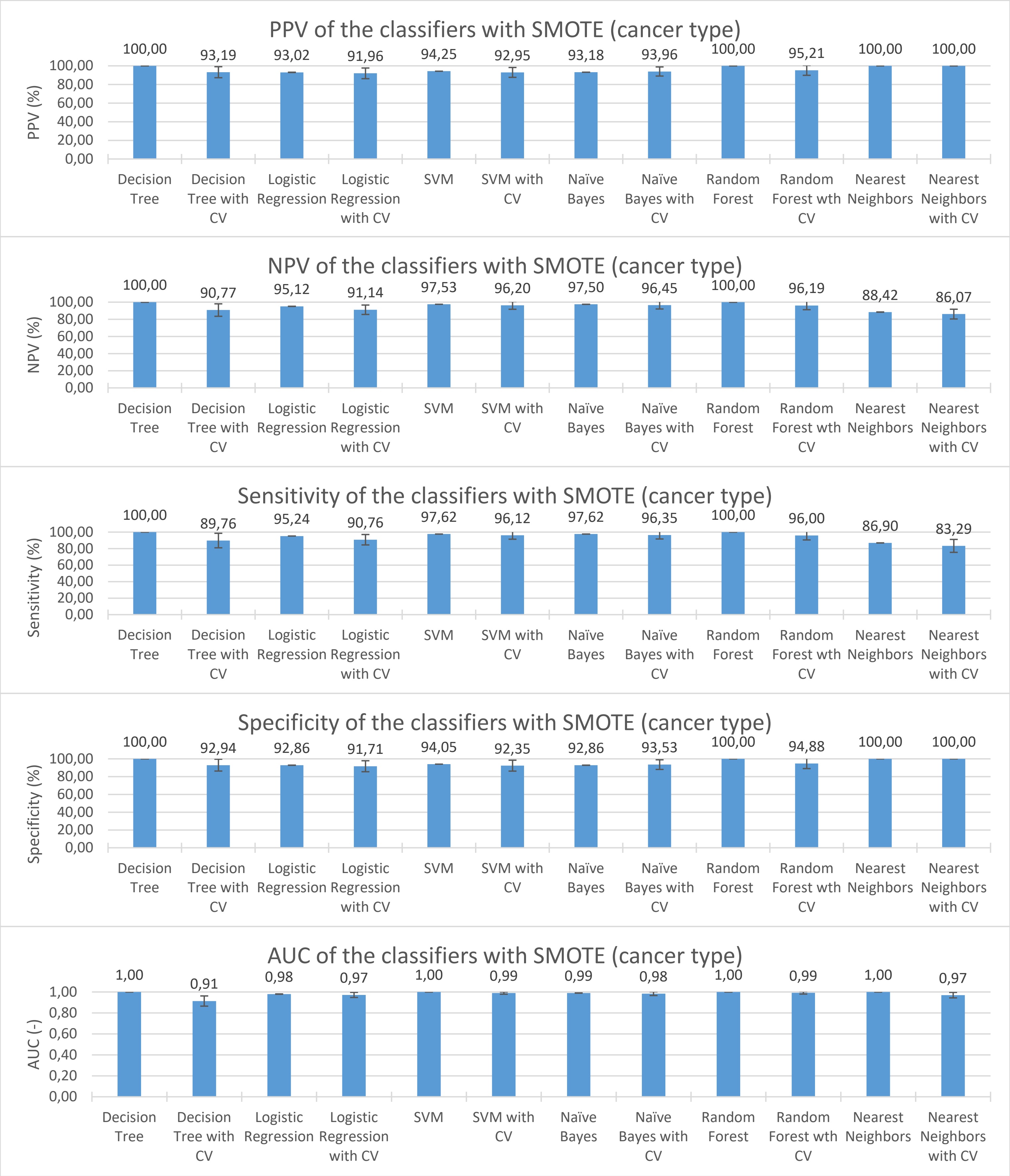


Figure 8 *Performances of the machine learning models for cancer type classification with SMOTE*

Discussion

To goal of this study is to assess the seven selected serum tumor markers in their function to diagnose lung cancer and cancer type in individuals. Classification of these two categories has been done with the use of algorithms for threshold definition and data-driven machine learning classifiers. Evaluation of each of these techniques has then taken place to enable comparisons and thus assess the possibilities for cancer detection using blood sampling. This provides the clinic with a minimally-invasive method of lung cancer diagnosis, while also enabling frequent examinations. This benefits the patients as tumors can be detected earlier, which increases the patients 5-year survival possibility.

Previous research

Previous research has already shown that thresholding approaches and machine learning models can be used in the medical field for cancer diagnosis. The combined assessment of defined thresholds for lung cancer classification has shown to result in a PPV, NPV, sensitivity and specificity of respectively 87.3%, 83.7%, 88.5% and 82% (Molina, et al., 2016). However, this study used a large cohort of individuals (n=3144) compared to the 225 patients which are included in this report. Decreasing the sample size can affect the ability of the algorithms to capture the general patterns, which are present in reality and thus a large dataset (Tipton, Hallberg, Hedges, & Chan, 2017). This can therefore affect the performance of the models when trained, but also when implemented in the clinic. A second problem is also present when classifying cancer type in particular, as the dataset is highly imbalanced for this class. This imbalance in the cancer type data will likely prevent the classification of SCLC due to the low sample size, because correctly classifying all the NSCLC samples can already be achieved by setting the thresholds to 0. Changing the default cancer type category will not solve this issue, as an infinite threshold value would then correctly classify all NSCLC patients.

Machine learning algorithms have been used in previous studies in the medical field. A study with similar TMs assessed several classifiers and thresholds for the classification of cancer. This provided that a SVM, LR and NN were able to achieve NPV's above 99% when evaluating men. The SVM model attained the highest sensitivity (76%), whereas the NN algorithm attained the highest specificity (86%) and PPV (3,9%). By contrast, for the women, the NN algorithm attained the highest sensitivity (65.5%), whereas the combined assessment of thresholds attained the highest specificity (88%) and PPV (2,2%). Moreover, all the methods attained high NPVs (all higher than 99%) (Wang, Hsieh, Wen, & Wen, 2016). Another study also showed that a decision tree classifier is able to diagnose lung cancer with a PPV, NPV, sensitivity, specificity and AUC of respectively 85%, 81%, 75%, 89% and 0.84 (Wang, et al., 2017).

An attempt to classify lung cancer type using machine learning models has not yet been presented by other studies. This prevents the comparison of the results. It must be noted that this study has used oversampling to artificially generate additional data, which is of essence for when comparisons will be made in the future. This can provide the issue that the presented data does not resemble the general patterns present in reality. However, the sample size of the SCLC class was lower than the dimensionality of the feature space which would otherwise prevent the classification of cancer type.

These studies also use more features for the classification process, which can benefit the performance of the models. Taken more features into account does result into a higher dimensionality of the feature space. However, if the dimensionality becomes too high, the performance of the classifiers will be start to decrease. This phenomenon is known as the Hughes phenomenon and shows that increasing the dimensionality of the feature space will improve the classifier performance at first to a certain extent. However, when the dimensionality becomes too high, overfitting will start to take place (Alonso, Malpica, & Martinez-Agirre, 2011).

Lung cancer classification

First of all the diagnosis of lung cancer using the various approaches. The clustermap indicated that separation of the classes was not possible using only the TM concentrations, and thus an unsupervised approach. The histograms also indicated that the thresholding approaches could encounter problems with defining a clear cut-off concentration, since there was significant overlap in the distributions. This was proven by the Bootstrap method and train/test split which indicated that the thresholds are not well-defined and depend highly on the composition of the presented dataset. This reflects in the optimal thresholding results, which presented low NPV values. The thresholds from the paper as well as the optimization of the AUC curve and Bootstrap with AUC were not able to provide consistent result, as one of the statistics would always drop when considering the combined assessment. Only the Bootstrap with optimization of the precision metric was able to achieve consistent results for all of the PPV, NPV, sensitivity and specificity. However, this method did show the lowest AUC value. The best AUC for the combined assessment is shown for the optimization of the AUC curve (0.48). This method also had the highest NPV (67%) and sensitivity (99%) of the thresholding methodology. However, this method essentially minimizes the number of false negatives by defining low threshold values and thus not predicting the ‘No’ class at all. This is confirmed by the low specificity. The Bootstrap with precision method is the only thresholding method which is able to avoid this given the maximum specificity (81%), while also having the best PPV (84%). However, this approach has a similar problem as the thresholds are chosen considerably high, which impedes the prediction of the ‘Yes’ class. This is confirmed by the sensitivity (58%).

A direct comparison with the machine learning algorithms cannot be made, which is why their performances are also presented with the same statistical values. The focus will be on the cross-validation as this provides the most robust results and thus the best comparisons. This shows that the models are able to achieve similar PPV values, with a maximum for the NB classifier (85%). However, this model does also have the biggest standard deviation. The model has thus been able to provide the most reliable positive predictions. This classifier is in spite of this not able detect the lung cancer that well, which can be seen at the low sensitivity. The exclusion of lung cancer is better for this algorithm with the maximum specificity of the models (83%). Confirmation of lung cancer was done best by the SVM model which achieved a sensitivity of 99%. The negative predictions are not that reliable however, as it has the lowest NPV and is cautious in assigning the lung cancer label. The other models also show this behaviour as the LR model has the highest NPV with 55% and a big 68% interval. Taking a look at the overall performance, by inspecting the AUC, shows that not all of the model are able to do that well for the small dataset. The RF has been able to achieve the highest AUC with a mere 0.69.

Enabling of SMOTE and thus increasing the number of patients does benefits the models and shows that increasing the number of patients in the study can benefit the outcomes. The highest AUC value is still for the RF model, but it has risen to 0.86. The oversampling method also benefits the specificity and NPV of the models, while decreasing the sensitivity and not affecting the PPV. It does thus increase the number of true negatives and false negatives, which implies that the number of positive predictions have decreased.

These values for the statistics shown in the bar plots are not similar to the performances of similar models in the literature. Only bootstrap with precision optimization is able to capture a similar PPV and specificity (84% and 81% versus 87% and 89% in the literature. The NPV and sensitivity are all significantly lower, but these studies did include significantly more individuals. The machine learning models also show a lower NPV and sensitivity compared to the literature, while the PPV and specificity are significantly higher (85% and 99% versus 4% and 86% in literature). Comparing the performance of the DT classifier with previous research shows that the DT which has been set up is not able to achieve comparable values, as all of the statistics are lower. However, these studies include a dataset with more patients which can influence the results. They also include more features which can benefit the performance of the machine learning algorithms.

Comparing the algorithms in this report can be done for each statistic. However, these models are intended to be used in clinic applications and it will therefore be more beneficial to evaluate them with the two clinical cases. The first case is that the doctor needs to diagnose lung cancer with a high certainty, while also minimizing the number of false positives. The positive predictions do therefore have to be reliable, while not too many cancer patients are missed. Taking into account the sensitivity is hence the first step, as this is the ratio of the detected lung cancer patients over the total of lung cancer patients. This shows that the optimal thresholding approaches with the combined assessment, except for Bootstrap with precision, may be used. However, these method classify almost every patient as cancer patient and do thus provide many false positives and thus unreliable positive predictions. The SVM does provide the best sensitivity (99%) and does thus recognize almost every cancer patient accordingly. However, the PPV (71%) is relatively low. More reliable positive predictions can be made with the NB classifier, which has an PPV of 85%. Most of the positive predictions do therefore include cancer patients and the least amount of healthy patients are hence incorrectly classified. The NB model can thus be a useful model if the positive predictions need to be of high certainty. However, bootstrap with optimization of the precision for the combined assessment can provide a similar certainty for the positive predictions (PPV is 84%). The sensitivity is also significantly higher with 58%, compared to the 39% of the NB model. This ensures that more of the cancer patients are recognized, which makes this the optimal method, since it also provides reliable positive predictions. If the reliability of the positive predictions is weighted less than the detection of the detection of cancer patients, the SVM model will become a better fit. It must be noted that the results for the machine learning classifiers with SMOTE are excluded, as it is not certain that the dataset resembles the population.

The second case focusses on excluding lung cancer with a high probability and at the same time identifying most of the healthy patients. Reliable negative predictions are indicated by a high NPV, which presents that the thresholds optimized with the AUC curve can be the best fit (NPV of 67%). However, this method has a specificity of 6% and does thus barely recognize any of the healthy patients. The method with a higher specificity is optimization of the precision metric to define threshold (81%), but this is accompanied with a lower NPV (52%). The machine learning models do also have a similar problem with the low NPV. The highest specificity is of the NB model (83%), but this has a NPV of 36%. Excluding most of the healthy patient can thus be done with the NB classifier, but the predictions are also unreliable. More reliable exclusions can be made with optimization of the AUC curve. However, the exclusion of lung cancer does not appear to be viable with the presented approaches.

Cancer type classification

The second set of approaches are focused on the classification of the cancer type SCLC versus NSCLC. The clustermap indicates that separation of these classes is not possible using only the TM concentrations, while the histograms show an overlap of the distribution of the classes. This indicates that the thresholding approaches can encounter problems when optimizing the cut-off concentration. This is proven by the Bootstrap method which shows a large range for some of the

TMs and thus an unreliable definition of the threshold. This is because it is 68% certain that the threshold will be in this range, which means that it can still be outside of this large range. The train/test split also show that some TMs do not have a clearly defined threshold, which implies that this cut-off concentration strongly depends on the composition of the presented dataset.

The uncertainty of the threshold values is observable in the results of the thresholding approaches, since the highest AUC is 0.61 for the CYFRA marker. This is especially due to the inability of these methods to provide reliable negative predictions as the maximum NPV is 15%. This also shows for the specificity, since the threshold for some of the markers are defined below the minimum concentration. Classification of the SCLC class does therefore never occur, which is optimal due to imbalanced dataset. The approaches are hence better in predicting the NSCLC classes, as some of the methods predict this type for every patient. This is confirmed by the sensitivity of 100% for the combined assessment of all of the approaches. The PPV is therefore also low as all the SCLC patients are incorrectly classified with the NSCLC label. These method are thus not usable, due the incompatibility of the thresholds to provide SCLC predictions.

Implementing machine learning algorithms with the same dataset resulted into problems as well, which were different of nature. The sample size of the SCLC class was lower than the dimensionality of the feature space and hence made it impossible to fit the classifiers. The SMOTE algorithm was hence used to artificially generate a balanced dataset. This provided high performances for each of the classifiers, which did not show any problems with reliably predicting either the positive or negative label. This can be deduced from the high PPV and NPV values of the cross-validation evaluation. The models were also able to correctly classify most of the cancer types as the sensitivity and specificity are above 90%, with exception of the sensitivity of the NN model. It must be noted that these performances can change if the models are presented with a balanced dataset containing real patients. The presented values can also not be compared to performances from the literature, since no similar research has yet been conducted and published.

Assessing the two clinical cases can deduce the optimal models for clinical application. First of all the necessity for a model which provides NSCLC predictions with a high certainty, which also minimizes the number of missed NSCLC patients. Providing NSCLC predictions with a high probability means that the ratio of the NSCLC patient over the predicted NSCLC patients, and thus the PPV, are high. The optimal model to do this is the NN classifier with a sensitivity of 100%, however this model is not able to recognize all of the NSCLC patients as the sensitivity is the lowest of all models (83%). The classifier with the best sensitivity is NB model which is thus able to identify more NSCLC patient, but the NSCLC prediction is less reliable. In this case the NN classifier is thus the best fit, as it provides NSCLC predictions which are of high certainty.

The second clinical case aims at providing reliable SCLC predictions, while missing a minimal number of SCLC patients. To provide the reliable predictions it is necessary to evaluate the NPV, which is the best for the NB model (96,5%), closely followed by the SVM and RF model. However, the NN model is able to detect all of the SCLC patients as it has a specificity of 100%. The RF algorithm is the second best algorithm for this with a specificity of 95%. This RF classifier is hence the best option in this case as it is able to provide reliable SCLC prediction, while also identifying most of the patients with the SCLC type for this artificially balanced dataset. This is all based on the cross-validation score as this provides the most robust performance evaluation.

Conclusions

This study shows the need for a supervised approach for the diagnosis of lung cancer and cancer type. Algorithms for the definition of thresholds as well as machine learning models were thus evaluated. This showed that the selected serum tumor markers in combination with specific models can be used for clinical applications. The confirmation of lung cancer with high reliability is most accurately done by the thresholds with optimized precision through cross-validation. Excluding lung cancer provided to be less viable with the used models. The detection of NSCLC in patients has been done most reliably with the 5-Nearest Neighbour classifier, while classifying SCLC with high probabilities is best done with the random forest model. The classification of SCLC versus NSCLC did rely on a oversampling algorithm (SMOTE), as the dataset was highly imbalanced. It is thus not assured that these models will be able to achieve similar performance when integrated in the clinic. The dataset also included a low number of individuals (n=225). Artificially increasing the sample size showed that extending the study can increase the performance of the classifiers, which can also be done with the inclusion of more tumor markers. Future studies will have to take this into account and evaluate whether increasing any or both of these parameters can benefit the performances. It can also focus on methods to combine the predicted probabilities of specific models, and thus take the strengths of each model into account, into an overall prediction, which will have to be evaluated.

References

Aliferis, C. (2002). Machine learning models for lung cancer classification using array comparative genomic hybridization. *Proc AMIA Symp*, 7-11.

Alonso, M., Malpica, A., & Martinez-Agirre, A. (2011). Consequences of the Hughes phenomenon on some classification Techniques. *American Society for Photogrammetry and Remote Sensing Annual Conference 2011.* Milwaukee.

Baba, T., Uramoto, H., Kuwata, T., Oka, S., Shigematsu, Y., & Nagata, Y. (2011). A study of surgically resected peripheral non-small cell lung cancer with a tumor diameter of 1.0 cm or less. *Scandinavian Journal of Surgery*, 153–158.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics*, 106.

Dayen, C., Debieuvre, D., Molinier, O., Raffy, O., Paganin, F., Virally, J., & Grivaux, M. (2017). New insights into stage and prognosis in small cell lung cancer: an analysis of 968 cases. *Journal of thoracic disease*, 5101–5111.

Hashi, Z., & Almamlook, R. (2018). Lung Cancer Survival Prediction Using Random Forest Based Decision Tree Algorithms. *Proceedings of the International Conference on Industrial Engineering and Operations Management* (p. 2602). Washington, DC: IEOM Society International.

Jra, B. N., Bhattoaa, H. P., Steiber, Z., Csobán, M., Szilasi, M., & Méhes, G. (2014). Serum human epididymis protein 4 (HE4) as a tumor marker in men with lung cancer. *Clin Chem Lab Med*, 1639–1648.

Kajatt, E. A. (2013). Lung cancer: a review of current knowledge, diagnostic methods and therapeutic perspectives. *Revista Peruana de Medicina Experimental y Salud Pública*, 85-92.

Kelly, M. (2015). Lung Cancer: Diagnosis, Treatment Principles, and Screening. *American Family Physician*, 250-256.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 1011–1013.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 8–17.

Liu, W. (2013). Evaluating the clinical significance of serum HE4 levels in lung cancer and pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 1346-1353.

Molina, R., Auge, J., Bosch, X., Escudero, J., Viñolas, N., & Marrades, R. (2009). Usefulness of serum tumor markers, including progastrin-releasing peptide, in patients with lung cancer: correlation with histology. *Tumour Biol*, 121–129.

Molina, R., JM, A., Escudero, J., Marrades, R., Viñolas, N., & Carcereny, E. (2008). Mucins CA 125, CA 19.9, CA 15.3 and TAG-72.3. *Tumour Biol*, 371–380.

Molina, R., Marrades, R., Auge, J., Escudero, J., Viñolas, N., & Reguart, N. (2016). Assessment of a Combined Panel of Six Serum Tumor Markers for lung cancer. *Am J Respir Crit Care Med*, 427–437.

Morihito, O. (2005). Effect of tumor size on prognosis in patients with non–small cell lung cancer: The role of segmentectomy as a type of lesser resection. *The Journal of Thoracic and Cardiovascular Surgery*, 87-93.

scikit-learn. (2018). *Cross-validation: evaluating estimator performance*. Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/cross_validation.html

Sharma, D. N. (2015). Lung cancer screening: history, current perspectives, and future directions. *Archives of medical science: AMS*, 1033–1043.

Shi, C., Zhang, X., & Han, B. (2011). A clinicopathological study of resected non-small cell lung cancers 2 cm or less in diameter: a prognostic assessment. *Med Oncol*, 1441.

Shortliffe, E., & Cimino, J. (2006). *Biomedical informatics: computer applications in health care and biomedicine.* New York, NY: Springer.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of Small Samples for Generalization: Adjustments and Rules of Thumb. *Evaluation Review*, 472–505. doi:https://doi.org/10.1177/0193841X16655665

Wang, H.-Y., Hsieh, C.-H., Wen, C.-N., & Wen, Y.-H. (2016). Cancers Screening in an Asymptomatic Population by Using Multiple Tumour Markers. *PLoS ONE*.

Wang, Z., Feng, F., Zhou, X., Duan, L., Wang, J., Wu, Y., & Wang, N. (2017). Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. *Oncotarget*, 94793–94804.

Wang-Yu, Z., Lin-lin, T., Zhao-yu, W., Shan-jun, W., Li-yun, X., Wei, Y., . . . Yong-Kui, Z. (2016). Clinical characteristics and advantages of primary peripheral micro-sized lung adenocarcinoma over small-sized lung adenocarcinoma. *European Journal of Cardio-Thoracic Surgery*, 1095–1102.

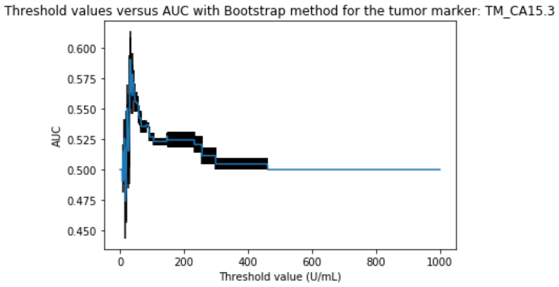
WHO. (2018, September 12). *Cancer*. Retrieved from WHO: https://www.who.int/news-room/fact-sheets/detail/cancer

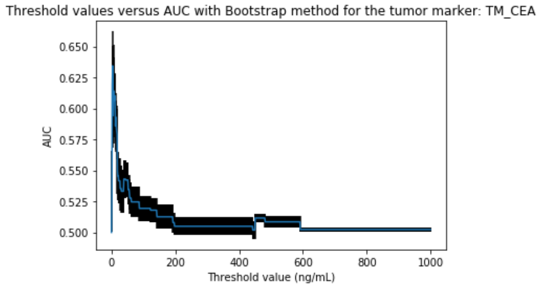
Witten, I., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques.* Amsterdam: Morgan Kaufman.

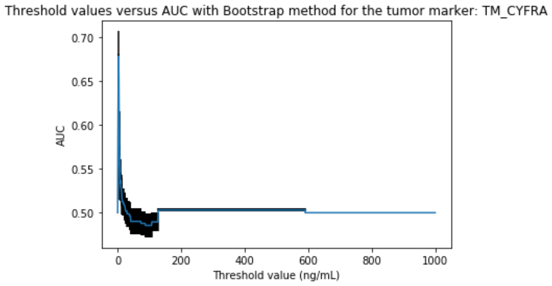
Appendices

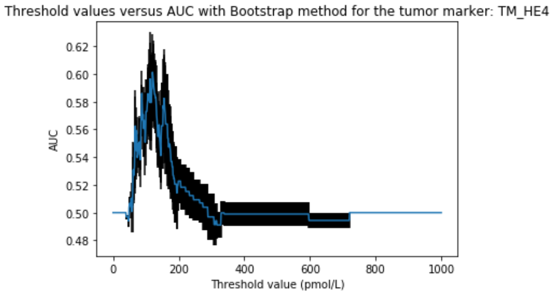
Appendix A: Classification of LC

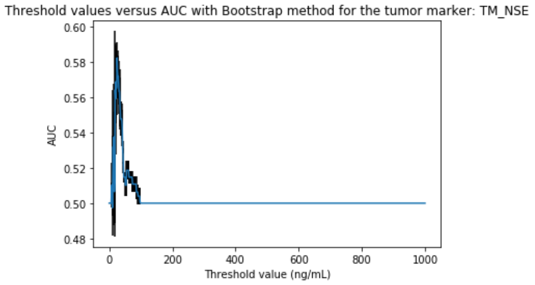
Appendix A1: Optimal thresholding curves of Bootstrap for AUC (lung cancer)

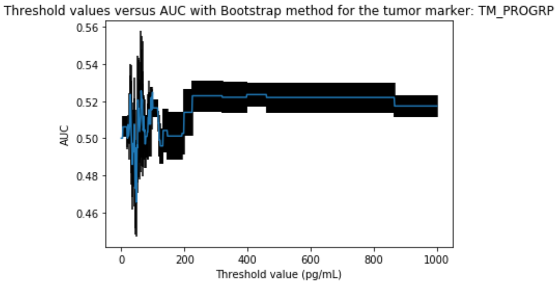


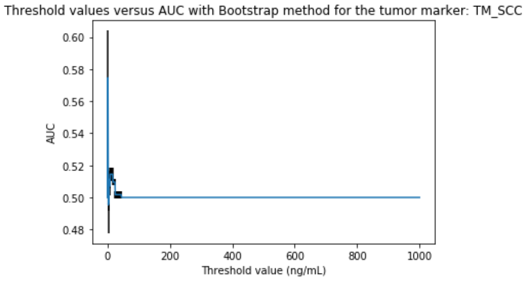






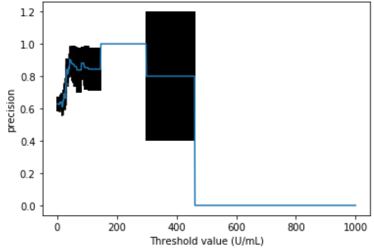




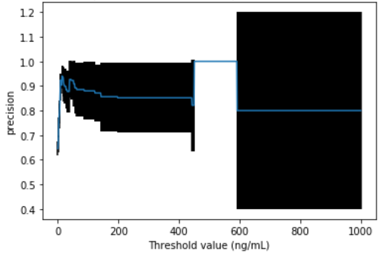


Appendix A2: Optimal thresholding curves of Bootstrap for precision (lung cancer)

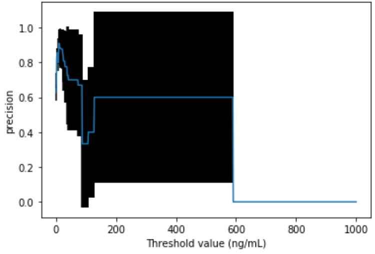
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CA15.3



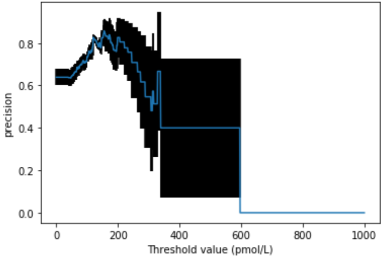
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CEA



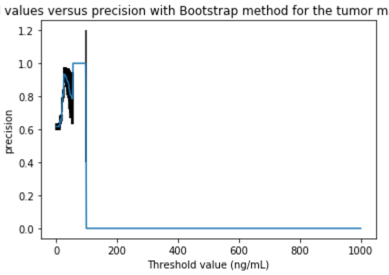
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CYFRA



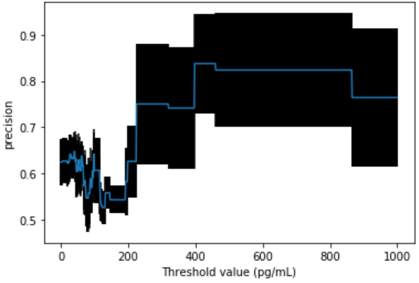
Threshold values versus precision with Bootstrap method for the tumor marker: TM_HE4



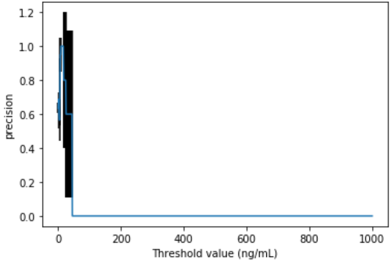
Threshold values versus precision with Bootstrap method for the tumor marker: TM_NSE



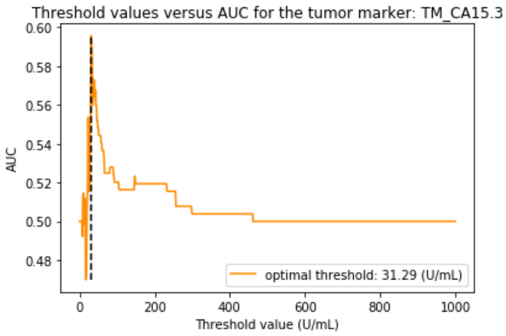
Threshold values versus precision with Bootstrap method for the tumor marker: TM_PROGRP

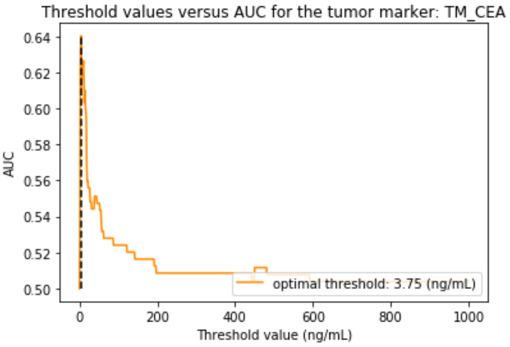


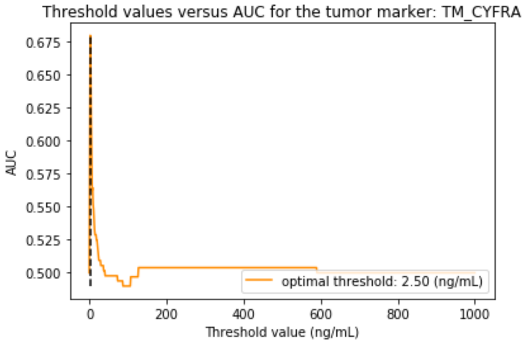
Threshold values versus precision with Bootstrap method for the tumor marker: TM_SCC

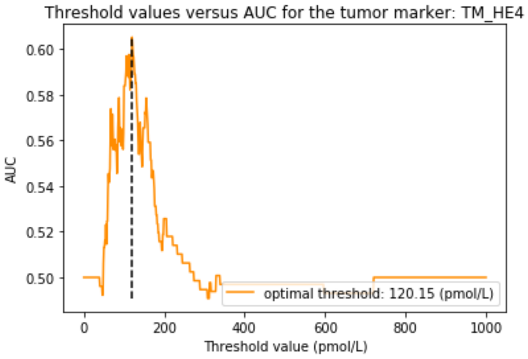


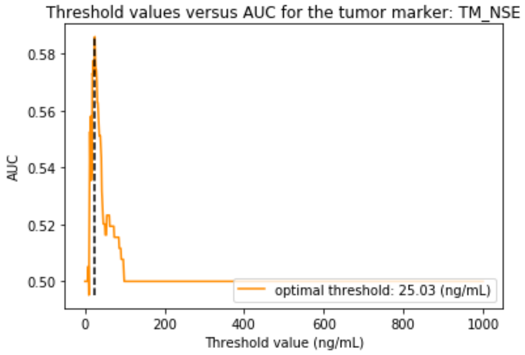
Appendix A3: Optimal thresholding curves for optimization of AUC (lung cancer)

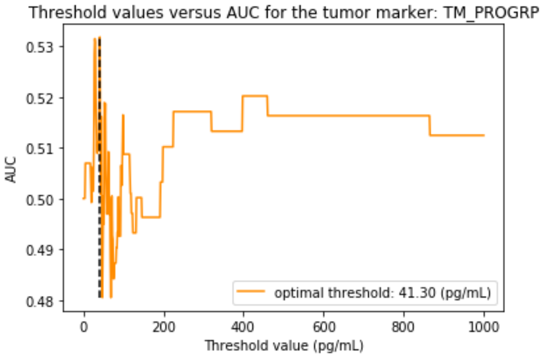

Threshold values versus AUC for the tumor marker: TM_CA15.3

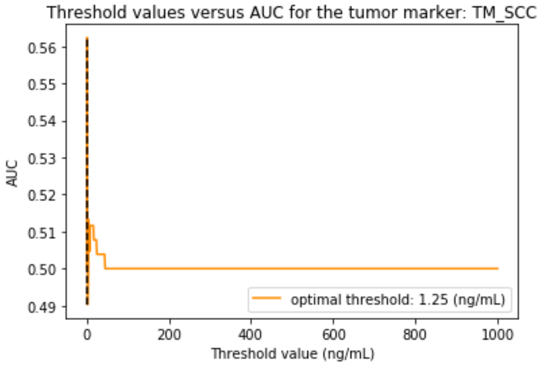

Threshold values versus AUC for the tumor marker: TM_CEA


Threshold values versus AUC for the tumor marker: TM_CYFRA


Threshold values versus AUC for the tumor marker: TM_HE4

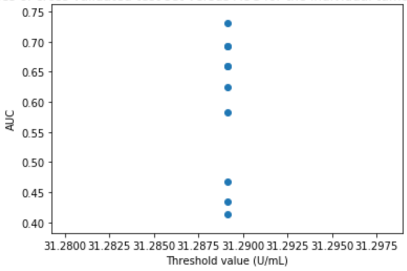

Threshold values versus AUC for the tumor marker: TM_NSE


Threshold values versus AUC for the tumor marker: TM_PROGRP

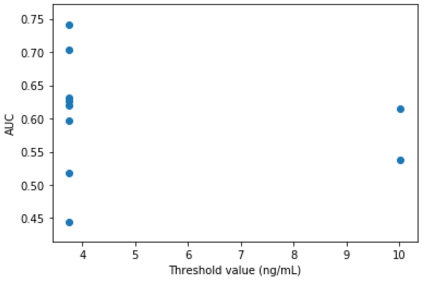

Threshold values versus AUC for the tumor marker: TM_SCC

Appendix A4: Optimal thresholding with train and test split (lung cancer)

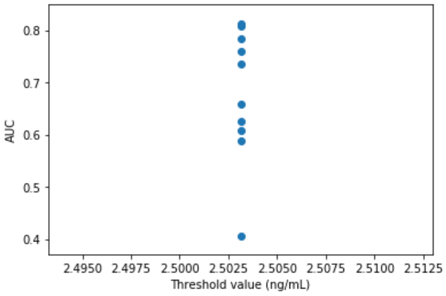
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CA15.3



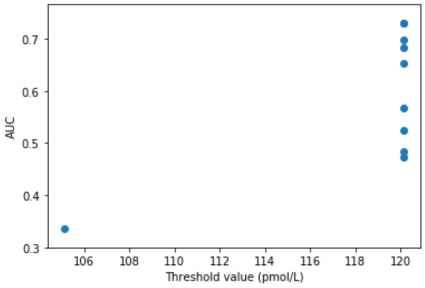
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CEA



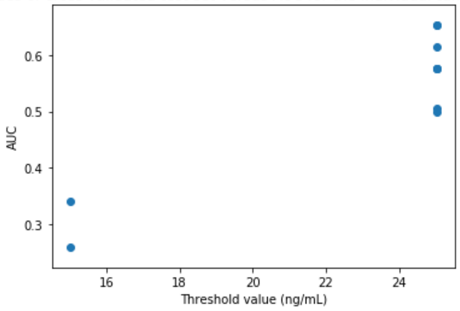
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CYFRA



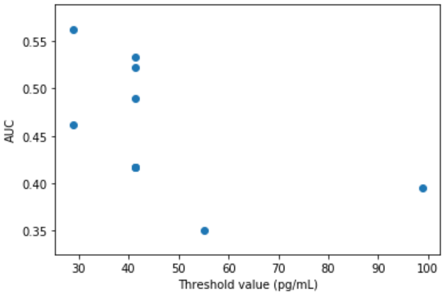
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_HE4



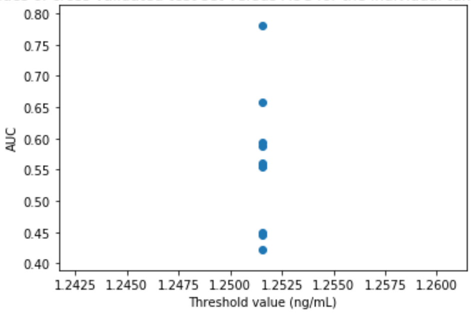
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_NSE



Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_PROGRP



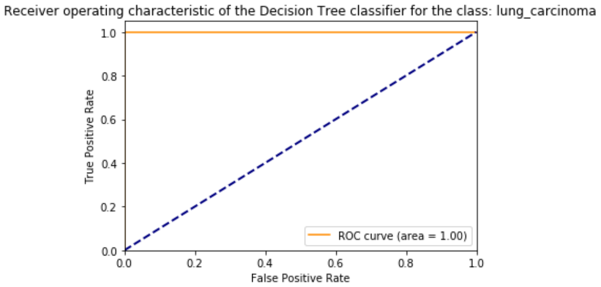
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_SCC

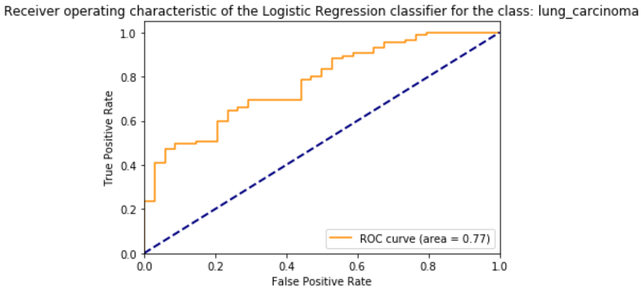


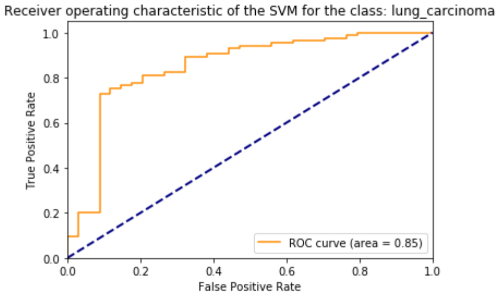
Appendix A5: Thresholds for all optimal thresholding methods (lung cancer)

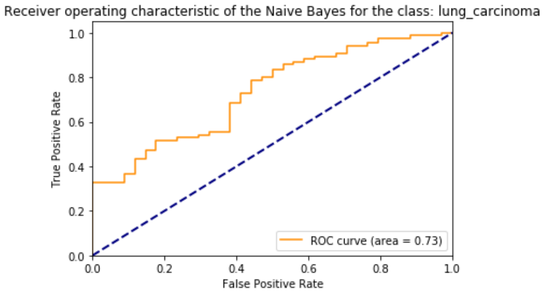
Thresholds	TM_CA15.3 (U/mL)	TM_CEA (ng/mL)	TM_CYFRA (ng/mL)	TM_HE4 (pmol/L)	TM_NSE (ng/mL)	TM_PROGRP (pg/mL)	TM_SCC (ng/mL)
Paper	35	5	3.3	-	25	50	2
Bootstrap AUC	32.54	3.75	2.50	120.15	25.03	62.58	1
Bootstrap AUC range	30.04-31.29	2.50-13.77	1.25-3.76	118.90-167.71	23.78-28.79	40.05-42.55	0.0-1.25
Bootstrap precision	42.55	13.77	10.01	155.19	25.03	398.00	6
Bootstrap precision range	41.30-145.18	12.52-21.28	8.76-23.78	153.94-178.97	23.78-35.04	396.75-396.75	5.00-7.51
Optimization of AUC	31.29	3.75	2.50	120.15	25.03	41.30	1

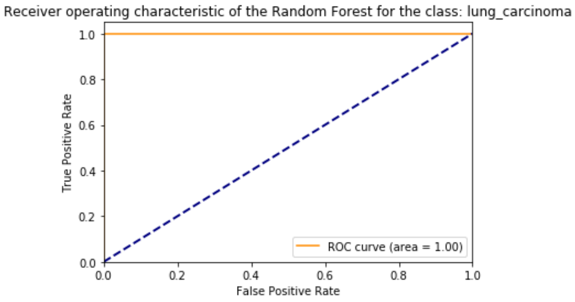
Appendix A6: ROC curves of the machine learning models (lung cancer)

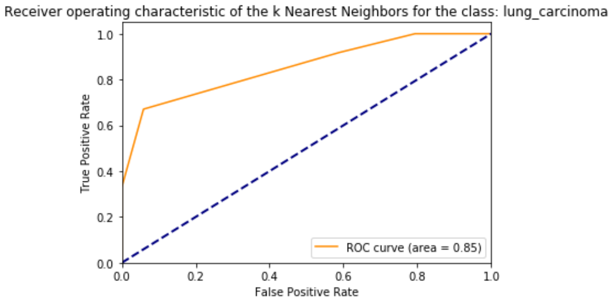

Receiver operating characteristic of the Decision Tree classifier for the class: lung_carcinoma


Receiver operating characteristic of the Logistic Regression classifier for the class: lung_carcinoma


Receiver operating characteristic of the SVM for the class: lung_carcinoma


Receiver operating characteristic of the Naive Bayes for the class: lung_carcinoma

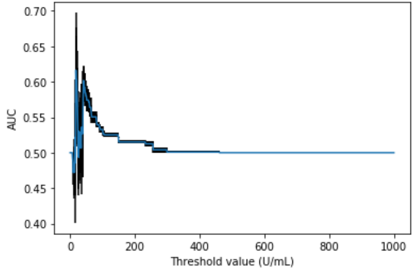

Receiver operating characteristic of the Random Forest for the class: lung_carcinoma


Receiver operating characteristic of the k Nearest Neighbors for the class: lung_carcinoma

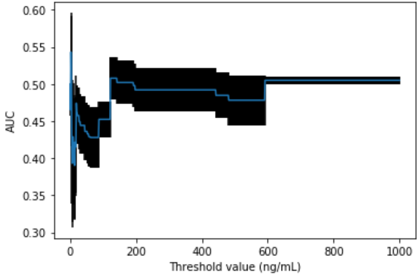
Appendix B: Classification of cancer type

Appendix B1: Optimal thresholding curves of Bootstrap for AUC (Cancer type)

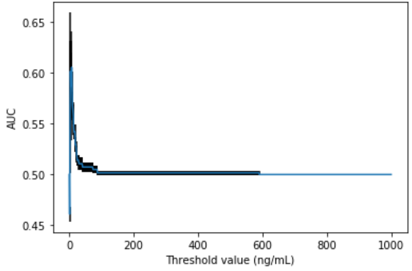
Threshold values versus AUC with Bootstrap method for the tumor marker: TM_CA15.3



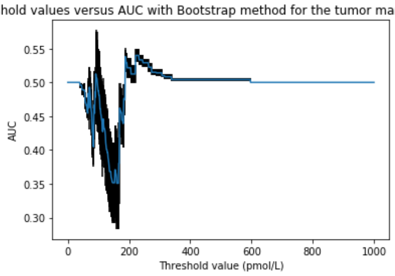
Threshold values versus AUC with Bootstrap method for the tumor marker: TM_CEA



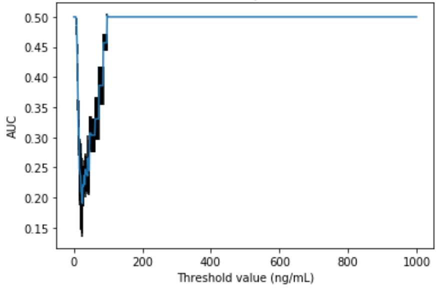
Threshold values versus AUC with Bootstrap method for the tumor marker: TM_CYFRA



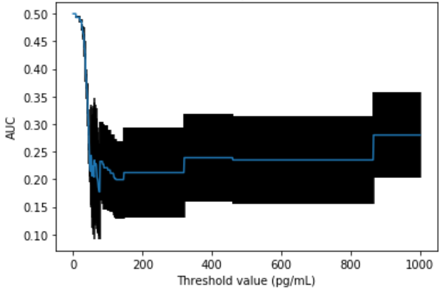
Threshold values versus AUC with Bootstrap method for the tumor marker: TM_HE4



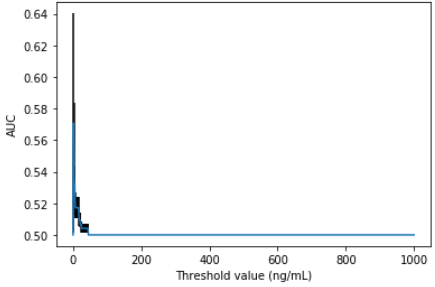
Threshold values versus AUC with Bootstrap method for the tumor marker: TM_NSE



Threshold values versus AUC with Bootstrap method for the tumor marker: TM_PROGRP

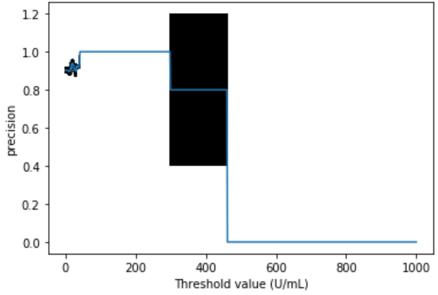


Threshold values versus AUC with Bootstrap method for the tumor marker: TM_SCC

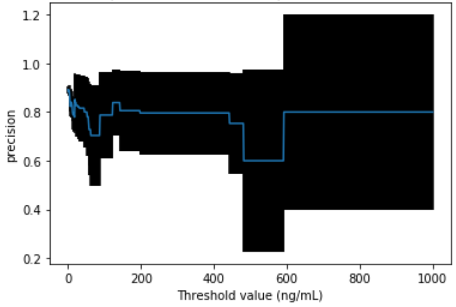


Appendix B2: Optimal thresholding curves of Bootstrap for precision (Cancer type)

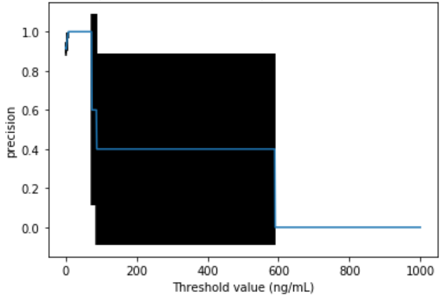
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CA15.3



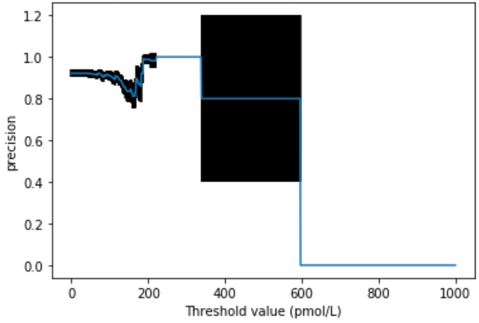
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CEA



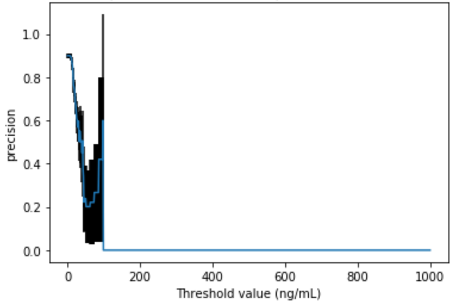
Threshold values versus precision with Bootstrap method for the tumor marker: TM_CYFRA



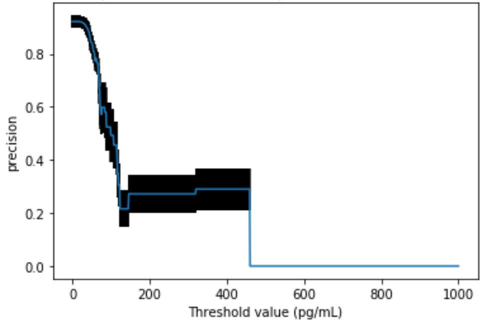
Threshold values versus precision with Bootstrap method for the tumor marker: TM_HE4



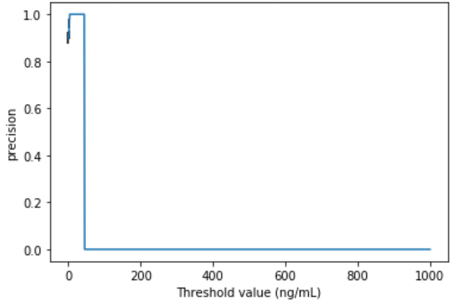
Threshold values versus precision with Bootstrap method for the tumor marker: TM_NSE



Threshold values versus precision with Bootstrap method for the tumor marker: TM_PROGRP

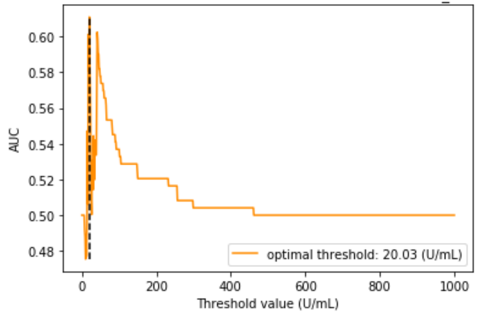


Threshold values versus precision with Bootstrap method for the tumor marker: TM_SCC

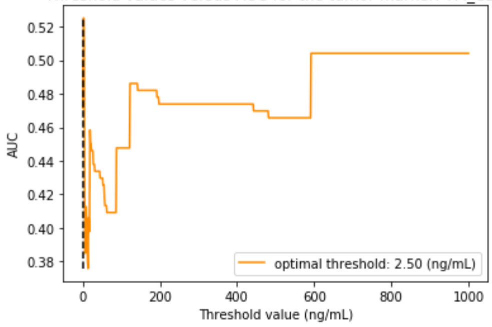


Appendix B3: Optimal thresholding curves for optimization of AUC (Cancer type)

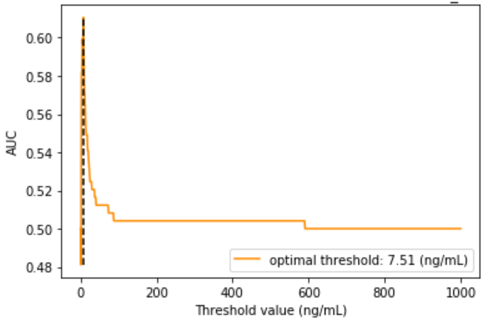
Threshold values versus AUC for the tumor marker: TM_CA15.3



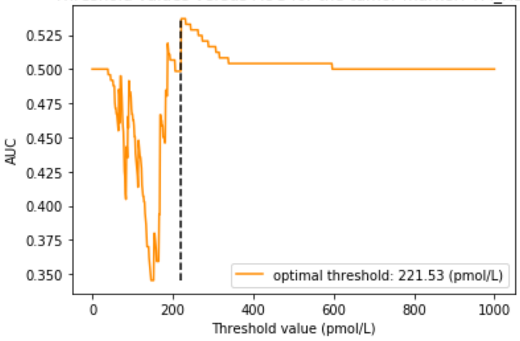
Threshold values versus AUC for the tumor marker: TM_CEA



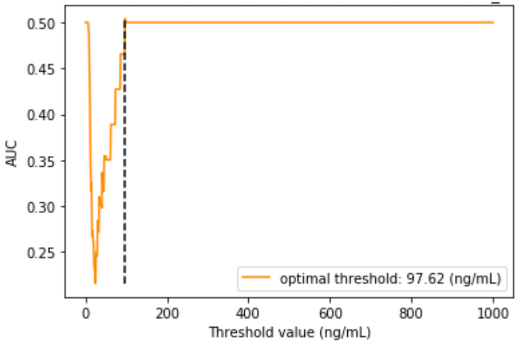
Threshold values versus AUC for the tumor marker: TM_CYFRA



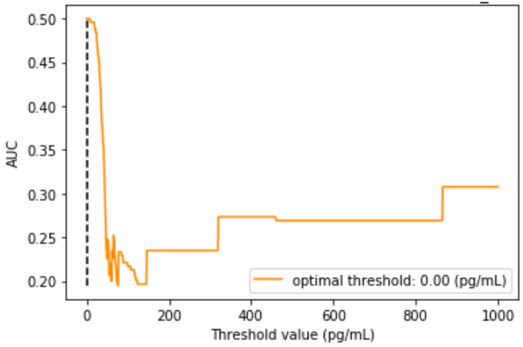
Threshold values versus AUC for the tumor marker: TM_HE4



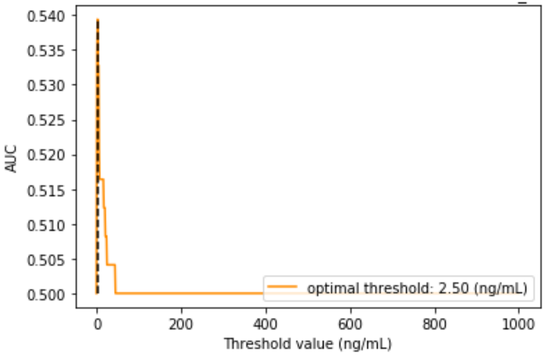
Threshold values versus AUC for the tumor marker: TM_NSE



Threshold values versus AUC for the tumor marker: TM_PROGRP

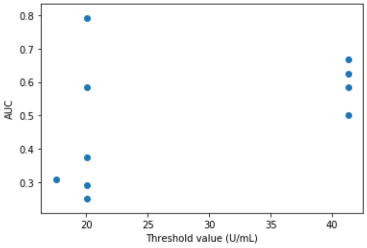


Threshold values versus AUC for the tumor marker: TM_SCC

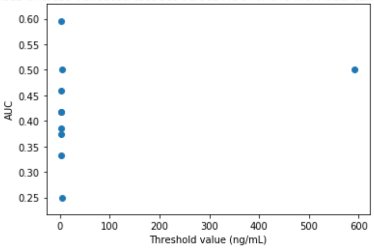


Appendix B4: Optimal thresholding with train and test split (Cancer type)

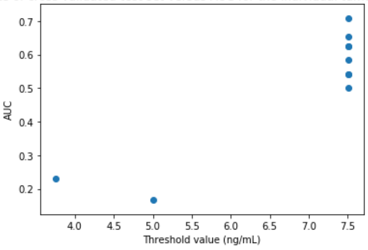
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CA15.3



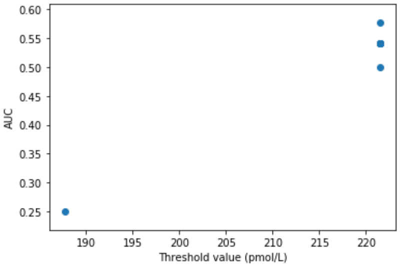
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CEA



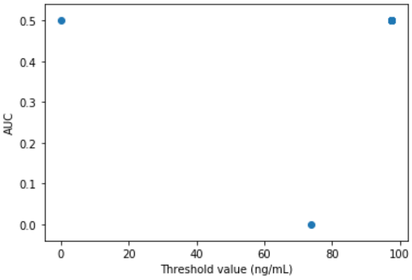
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_CYFRA



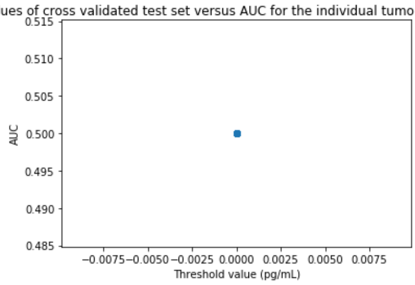
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_HE4



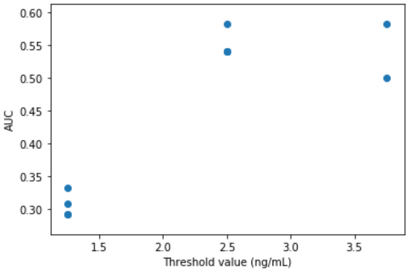
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_NSE



Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_PROGRP



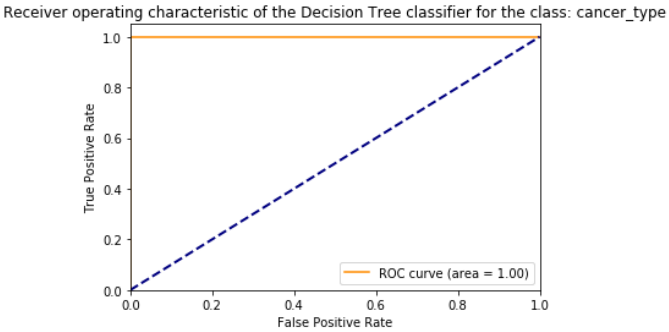
Threshold values of cross validated test set versus AUC for the individual tumor marker : TM_SCC

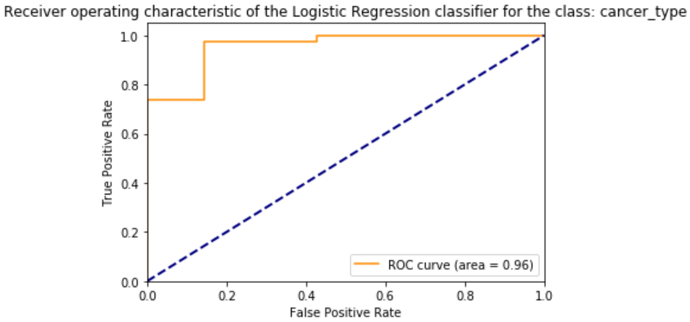


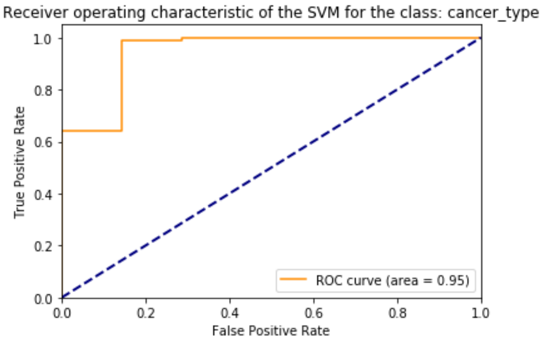
Appendix B5: Thresholds for all optimal thresholding methods (Cancer type)

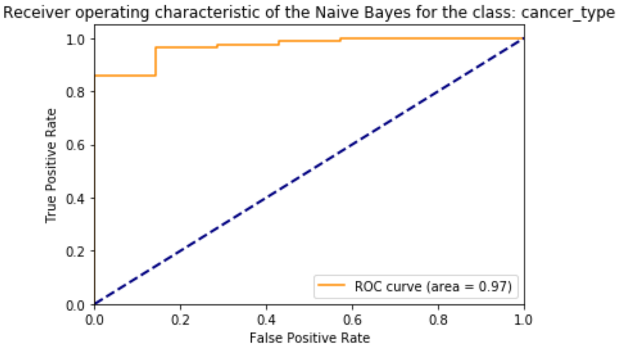
Thresholds	TM_CA15.3 (U/mL)	TM_CEA (ng/mL)	TM_CYFRA (ng/mL)	TM_HE4 (pmol/L)	TM_NSE (ng/mL)	TM_PROGRP (pg/mL)	TM_SCC (ng/mL)
Bootstrap AUC	17.52	2.50	7.51	221.53	97.62	0.00	1.25
Bootstrap AUC range	16.27-22.53	0.0-3.76	6.26-7.51	220.28-244.06	96.37-97.62	0.0-7.51	0.0-43.81
Bootstrap precision	40.05	2.50	5.01	70.09	0.00	0.00	2.50
Bootstrap precision range	38.80-40.05	1.25-3.76	3.76-5.01	68.84-106.38	0.0-8.76	0.0-37.55	1.25-2.50
Optimization of AUC	20.03	2.50	7.51	221.53	97.62	0.00	2.50

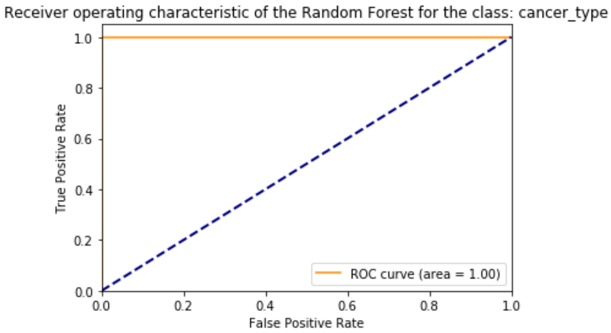
Appendix B6: ROC curves of the machine learning models (Cancer type)

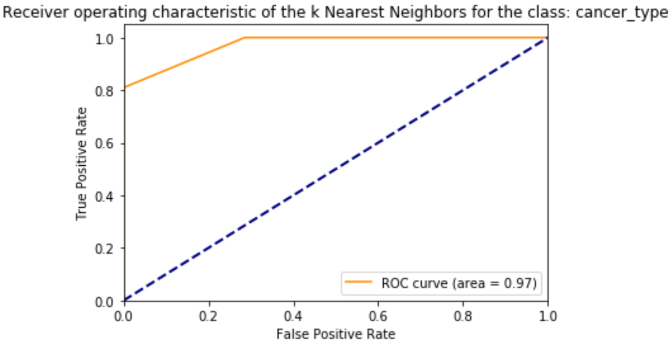






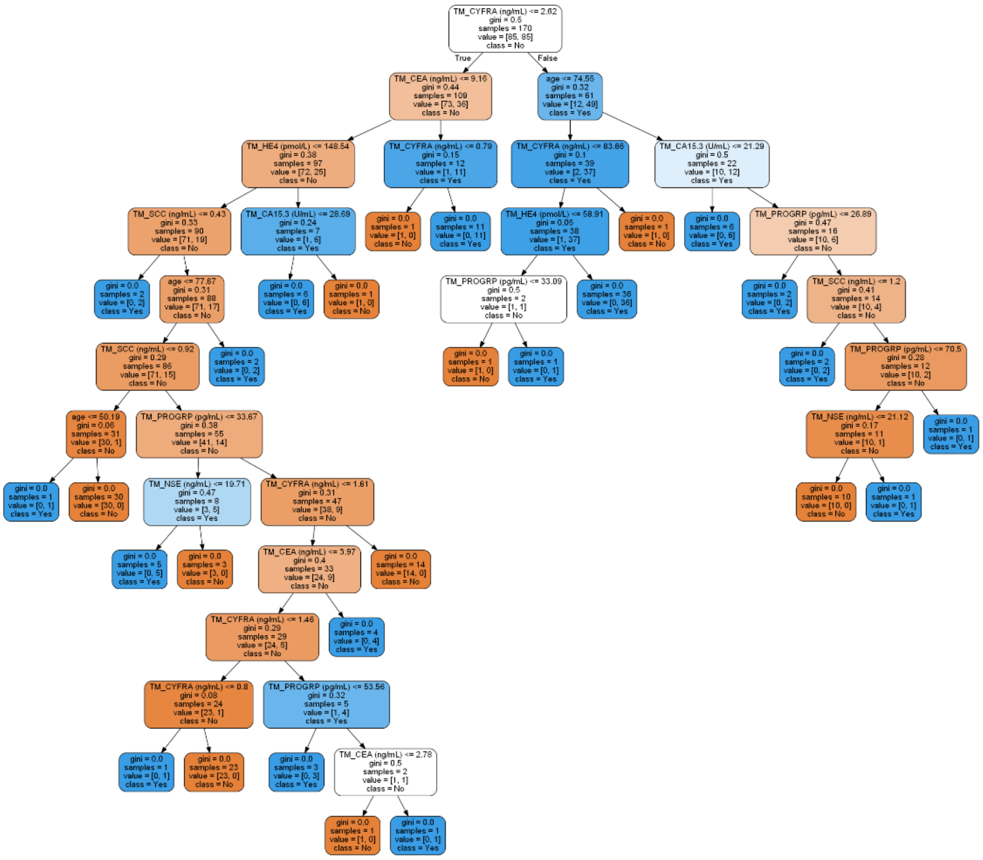




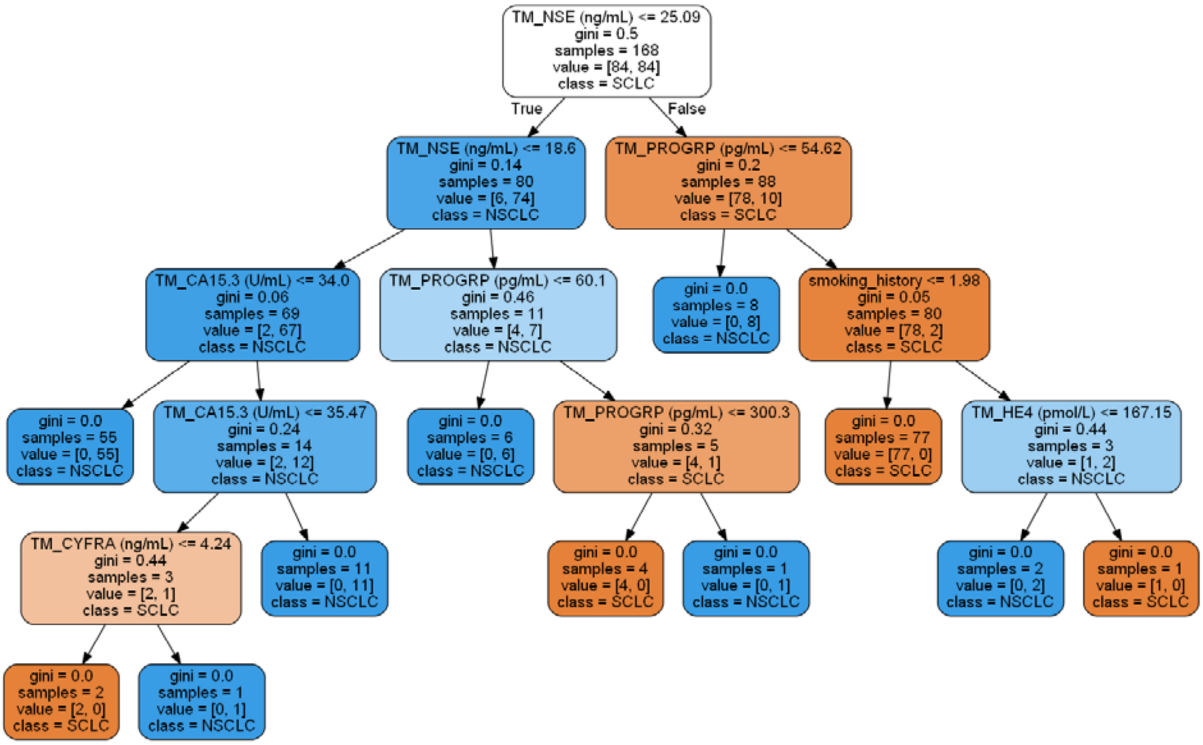


Appendix C: Visualization of DT classifiers

Appendix C1: Decision Tree for LC classification


TM_CYFRA (ng/mL) <= 2.62 gini = 0.5 samples = 170 value = [85, 85] class = No
True
False
TM_CEA (ng/mL) <= 9.16 gini = 0.44 samples = 109 value = [73, 36] class = No
age <= 74.55 gini = 0.32 samples = 61 value = [12, 49] class = Yes
TM_HE4 (pmol/L) <= 148.54 gini = 0.38 samples = 97 value = [72, 25] class = No
TM_CYFRA (ng/mL) <= 0.79 gini = 0.15 samples = 12 value = [1, 11] class = Yes
TM_CYFRA (ng/mL) <= 83.66 gini = 0.1 samples = 39 value = [2, 37] class = Yes
TM_CA15.3 (U/mL) <= 21.29 gini = 0.5 samples = 22 value = [10, 12] class = Yes
TM_SCC (ng/mL) <= 0.43 gini = 0.33 samples = 90 value = [71, 19] class = No
TM_CA15.3 (U/mL) <= 28.69 gini = 0.24 samples = 7 value = [1, 6] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 11 value = [0, 11] class = Yes
TM_HE4 (pmol/L) <= 58.91 gini = 0.05 samples = 38 value = [1, 37] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 6 value = [0, 6] class = Yes
TM_PROGRP (pg/mL) <= 26.89 gini = 0.47 samples = 16 value = [10, 6] class = No
gini = 0.0 samples = 2 value = [0, 2] class = Yes
age <= 77.67 gini = 0.31 samples = 88 value = [71, 17] class = No
gini = 0.0 samples = 6 value = [0, 6] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
TM_PROGRP (pg/mL) <= 33.09 gini = 0.5 samples = 2 value = [1, 1] class = No
gini = 0.0 samples = 36 value = [0, 36] class = Yes
gini = 0.0 samples = 2 value = [0, 2] class = Yes
TM_SCC (ng/mL) <= 1.2 gini = 0.41 samples = 14 value = [10, 4] class = No
TM_SCC (ng/mL) <= 0.92 gini = 0.29 samples = 86 value = [71, 15] class = No
gini = 0.0 samples = 2 value = [0, 2] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 2 value = [0, 2] class = Yes
TM_PROGRP (pg/mL) <= 70.5 gini = 0.28 samples = 12 value = [10, 2] class = No
age <= 50.19 gini = 0.06 samples = 31 value = [30, 1] class = No
TM_PROGRP (pg/mL) <= 33.67 gini = 0.38 samples = 55 value = [41, 14] class = No
TM_NSE (ng/mL) <= 21.12 gini = 0.17 samples = 11 value = [10, 1] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 30 value = [30, 0] class = No
TM_NSE (ng/mL) <= 19.71 gini = 0.47 samples = 8 value = [3, 5] class = Yes
TM_CYFRA (ng/mL) <= 1.61 gini = 0.31 samples = 47 value = [38, 9] class = No
gini = 0.0 samples = 10 value = [10, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 5 value = [0, 5] class = Yes
gini = 0.0 samples = 3 value = [3, 0] class = No
TM_CEA (ng/mL) <= 3.97 gini = 0.4 samples = 33 value = [24, 9] class = No
gini = 0.0 samples = 14 value = [14, 0] class = No
TM_CYFRA (ng/mL) <= 1.46 gini = 0.29 samples = 29 value = [24, 5] class = No
gini = 0.0 samples = 4 value = [0, 4] class = Yes
TM_CYFRA (ng/mL) <= 0.8 gini = 0.08 samples = 24 value = [23, 1] class = No
TM_PROGRP (pg/mL) <= 53.56 gini = 0.32 samples = 5 value = [1, 4] class = Yes
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 23 value = [23, 0] class = No
gini = 0.0 samples = 3 value = [0, 3] class = Yes
TM_CEA (ng/mL) <= 2.78 gini = 0.5 samples = 2 value = [1, 1] class = No
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes

Appendix C2: Decision Tree for cancer type classification


TM_NSE (ng/mL) <= 25.09 gini = 0.5 samples = 168 value = [84, 84] class = SCLC
True
False
TM_NSE (ng/mL) <= 18.6 gini = 0.14 samples = 80 value = [6, 74] class = NSCLC
TM_PROGRP (pg/mL) <= 54.62 gini = 0.2 samples = 88 value = [78, 10] class = SCLC
TM_CA15.3 (U/mL) <= 34.0 gini = 0.06 samples = 69 value = [2, 67] class = NSCLC
TM_PROGRP (pg/mL) <= 60.1 gini = 0.46 samples = 11 value = [4, 7] class = NSCLC
gini = 0.0 samples = 8 value = [0, 8] class = NSCLC
smoking_history <= 1.98 gini = 0.05 samples = 80 value = [78, 2] class = SCLC
gini = 0.0 samples = 55 value = [0, 55] class = NSCLC
TM_CA15.3 (U/mL) <= 35.47 gini = 0.24 samples = 14 value = [2, 12] class = NSCLC
gini = 0.0 samples = 6 value = [0, 6] class = NSCLC
TM_PROGRP (pg/mL) <= 300.3 gini = 0.32 samples = 5 value = [4, 1] class = SCLC
gini = 0.0 samples = 77 value = [77, 0] class = SCLC
TM_HE4 (pmol/L) <= 167.15 gini = 0.44 samples = 3 value = [1, 2] class = NSCLC
TM_CYFRA (ng/mL) <= 4.24 gini = 0.44 samples = 3 value = [2, 1] class = SCLC
gini = 0.0 samples = 11 value = [0, 11] class = NSCLC
gini = 0.0 samples = 4 value = [4, 0] class = SCLC
gini = 0.0 samples = 1 value = [0, 1] class = NSCLC
gini = 0.0 samples = 2 value = [0, 2] class = NSCLC
gini = 0.0 samples = 1 value = [1, 0] class = SCLC
gini = 0.0 samples = 2 value = [2, 0] class = SCLC
gini = 0.0 samples = 1 value = [0, 1] class = NSCLC

Appendix D: Tables with the performances of the algorithms

Appendix D1: Performance of the optimal thresholding models for lung cancer classification

Thresholds from Molina paper					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/mL)	80.00	39.16	21.71	90.28	0.56
TM_CEA (ng/mL)	77.63	44.00	45.74	76.39	0.61
TM_CYFRA (ng/mL)	87.14	48.09	47.29	87.50	0.67
TM_HE4 (pmol/L)	0.00	0.00	0.00	0.00	0.00
TM_NSE (ng/mL)	96.00	40.34	18.60	98.61	0.59
TM_PROGRP (pg/mL)	64.22	35.87	54.26	45.83	0.50
TM_SCC (ng/mL)	68.52	37.41	28.68	76.39	0.53
Combined assesment	68.42	60.00	90.70	25.00	0.42

Bootstrap with AUC					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/ml)	83.72	41.14	27.91	90.28	0.59
TM_CEA (ng/ml)	76.42	49.47	62.79	65.28	0.64
TM_CYFRA (ng/ml)	82.42	50.91	58.14	77.78	0.68
TM_HE4 (pmol/L)	78.26	43.18	41.86	79.17	0.61
TM_NSE (ng/ml)	96.00	40.34	18.60	98.61	0.59
TM_PROGRP (pg/mL)	65.57	36.43	31.01	70.83	0.51
TM_SCC (ng/ml)	69.37	42.22	59.69	52.78	0.56
Combined assessment	66.31	64.29	96.12	12.50	0.46

Bootstrap with precision					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/mL)	91.30	39.33	16.28	97.22	0.57
TM_CEA (ng/mL)	94.12	41.92	24.81	97.22	0.61
TM_CYFRA (ng/mL)	89.47	38.46	13.18	97.22	0.55
TM_HE4 (pmol/L)	83.78	40.24	24.03	91.67	0.58
TM_NSE (ng/mL)	96.00	40.34	18.60	98.61	0.59
TM_PROGRP (pg/mL)	87.50	36.79	5.43	98.61	0.52
TM_SCC (ng/mL)	80.00	36.22	3.10	98.61	0.51
Combined assesment	84.27	51.79	58.14	80.56	0.31

Optimization of AUC curve					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/mL)	82.98	41.56	30.23	88.89	0.60
TM_CEA (ng/mL)	76.42	49.47	62.79	65.28	0.64
TM_CYFRA (ng/mL)	82.42	50.91	58.14	77.78	0.68
TM_HE4 (pmol/L)	78.26	43.18	41.86	79.17	0.61

TM_NSE (ng/mL)	96.00	40.34	18.60	98.61	0.59
TM_PROGRP (pg/mL)	66.21	41.07	74.42	31.94	0.53
TM_SCC (ng/mL)	69.37	42.22	59.69	52.78	0.56
Combined assesment	65.13	66.67	98.45	5.56	0.48

Appendix D2: Performance of the machine learning models for lung cancer classification without SMOTE

Lung cancer classification without SMOTE										
metrics	PPV (%)	std	NPV(%)	std	sensitivity (%)	std	specificity (%)	std	AUC (-)	std
Paper	68.42	0.00	60.00	0.00	90.70	0.00	25.00	0.00	0.42	0.00
Decision Tree	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Decision Tree with CV	73.66	5.06	36.61	17.18	74.53	10.78	35.00	16.21	0.55	0.09
Logistic Regression	75.68	0.00	87.50	0.00	98.82	0.00	20.59	0.00	0.77	0.00
Logistic Regression with CV	74.09	3.87	55.31	32.06	89.71	9.31	23.57	14.34	0.67	0.11
SVM	75.68	0.00	87.50	0.00	98.82	0.00	20.59	0.00	0.85	0.00
SVM with CV	70.84	0.77	4.83	20.32	99.18	2.20	0.86	3.39	0.53	0.16
Naïve Bayes	91.18	0.00	36.47	0.00	36.47	0.00	91.18	0.00	0.73	0.00
Naïve Bayes with CV	85.18	11.39	36.02	5.90	38.71	11.86	82.57	14.23	0.64	0.12
Random Forest	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Random Forest wth CV	75.45	5.30	47.29	19.56	83.47	9.62	33.29	17.62	0.69	0.10
Nearest Neighbors	79.59	0.00	66.67	0.00	91.76	0.00	41.18	0.00	0.85	0.00
Nearest Neighbors with CV	72.60	4.17	37.54	21.30	82.18	9.65	24.57	14.01	0.58	0.12

Appendix D3: Performance of the machine learning models for lung cancer classification with SMOTE

Lung cancer classification with SMOTE										
metrics	PPV (%)	std	NPV(%)	std	sensitivity (%)	std	specificity (%)	std	AUC (-)	std
Paper	68.42	0.00	60.00	0.00	90.70	0.00	25.00	0.00	0.42	0.00
Decision Tree	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Decision Tree with CV	73.83	8.95	70.78	7.12	68.35	10.78	74.59	11.70	0.71	0.07
Logistic Regression	73.24	0.00	66.67	0.00	61.18	0.00	77.65	0.00	0.79	0.00
Logistic Regression with CV	69.03	8.94	63.33	6.94	56.00	11.84	74.59	9.07	0.74	0.07
SVM	82.35	0.00	71.57	0.00	65.88	0.00	85.88	0.00	0.86	0.00
SVM with CV	74.81	8.46	67.62	6.92	61.53	11.48	78.53	9.71	0.79	0.07
Naïve Bayes	89.19	0.00	60.90	0.00	38.82	0.00	95.29	0.00	0.77	0.00
Naïve Bayes with CV	82.44	11.79	60.71	4.42	40.82	11.00	90.47	7.19	0.74	0.07
Random Forest	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Random Forest wth CV	79.80	8.49	77.40	7.07	76.00	9.29	79.82	10.33	0.86	0.06
Nearest Neighbors	91.30	0.00	78.22	0.00	74.12	0.00	92.94	0.00	0.90	0.00
Nearest Neighbors with CV	80.87	9.54	67.06	6.34	56.88	12.09	85.76	8.57	0.77	0.07

Appendix D4: Performance of the optimal thresholding algorithms for cancer type classification

Bootstrap with AUC					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/ml)	93.10	14.58	66.39	53.85	0.60
TM_CEA (ng/ml)	90.91	12.00	81.97	23.08	0.53
TM_CYFRA (ng/ml)	100.00	12.04	22.13	100.00	0.61
TM_HE4 (pmol/L)	100.00	10.32	7.38	100.00	0.54
TM_NSE (ng/ml)	100.00	9.70	0.82	100.00	0.50
TM_PROGRP (pg/ml)	90.37	NaN	100.00	0.00	0.50
TM_SCC (ng/ml)	91.36	11.11	60.66	46.15	0.53
Combined Assessment for NSCLC	90.37	NaN	100.00	0.00	0.50

Bootstrap with precision					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/mL)	96.43	11.21	22.13	92.31	0.57
TM_CEA (ng/mL)	90.91	12.00	81.97	23.08	0.53
TM_CYFRA (ng/mL)	95.56	12.22	35.25	84.62	0.60

TM_HE4 (pmol/L)	90.27	9.09	83.61	15.38	0.50
TM_NSE (ng/mL)	90.37	NaN	100.00	0.00	0.50
TM_PROGRP (pg/mL)	90.37	NaN	100.00	0.00	0.50
TM_SCC (ng/mL)	95.00	10.43	15.57	92.31	0.54
Combined Assessment for NSCLC	90.37	NaN	100.00	0.00	0.50

Optimization of AUC curve					
Evaluated tumor marker	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	AUC
TM_CA15.3 (U/mL)	93.67	14.29	60.66	61.54	0.61
TM_CEA (ng/mL)	90.91	12.00	81.97	23.08	0.53
TM_CYFRA (ng/mL)	100.00	12.04	22.13	100.00	0.61
TM_HE4 (pmol/L)	100.00	10.32	7.38	100.00	0.54
TM_NSE (ng/mL)	100.00	9.70	0.82	100.00	0.50
TM_PROGRP (pg/mL)	90.37	NaN	100.00	0.00	0.50
TM_SCC (ng/mL)	95.00	10.43	15.57	92.31	0.54
Combined Assessment for NSCLC	90.37	NaN	100.00	0.00	0.50

Appendix D5: Performance of the machine learning classifiers for cancer type classification with SMOTE

cancer type classification with SMOTE										
metrics	PPV (%)	std	NPV (%)	std	sensitivity (%)	std	specificity (%)	std	AUC (-)	std
Decision Tree	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Decision Tree with CV	93.19	5.91	90.77	7.27	89.76	8.79	92.94	6.66	0.91	0.05
Logistic Regression	93.02	0.00	95.12	0.00	95.24	0.00	92.86	0.00	0.98	0.00
Logistic Regression with CV	91.96	5.72	91.14	5.49	90.76	6.29	91.71	6.17	0.97	0.02
SVM	94.25	0.00	97.53	0.00	97.62	0.00	94.05	0.00	1.00	0.00
SVM with CV	92.95	5.29	96.20	4.51	96.12	4.72	92.35	6.14	0.99	0.01
Naïve Bayes	93.18	0.00	97.50	0.00	97.62	0.00	92.86	0.00	0.99	0.00
Naïve Bayes with CV	93.96	4.89	96.45	4.42	96.35	4.69	93.53	5.42	0.98	0.02
Random Forest	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	1.00	0.00
Random Forest wth CV	95.21	5.37	96.19	4.98	96.00	5.51	94.88	5.80	0.99	0.01
Nearest Neighbors	100.00	0.00	88.42	0.00	86.90	0.00	100.00	0.00	1.00	0.00
Nearest Neighbors with CV	100.00	0.00	86.07	5.69	83.29	7.84	100.00	0.00	0.97	0.03