Computational prediction of diagnosis and feature selection on mesothelioma patient health records

by **Davide Chicco**

davide.chicco@gmail.com www.DavideChicco.it

Toronto Deep Learning Series - TDLS

Peter Munk Cardiac Centre

2019-02-19

UNIVERSITY OF TORONTO

PLOS | ONE

RESEARCH ARTICLE

Computational prediction of diagnosis and feature selection on mesothelioma patient health records

Davide Chicco[1,2]*, **Cristina Rovelli**[3]

1 Peter Munk Cardiac Centre, Toronto, Ontario, Canada, **2** Princess Margaret Cancer Centre, Toronto, Ontario, Canada, **3** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

* davidechicco@davidechicco.it

Abstract

Background

Mesothelioma is a lung cancer that kills thousands of people worldwide annually, especially those with exposure to asbestos. Diagnosis of mesothelioma in patients often requires time-consuming imaging techniques and biopsies. Machine learning can provide for a more effective, cheaper, and faster patient diagnosis and feature selection from clinical data in patient records.

Methods and findings

We analyzed a dataset of health records of 324 patients having mesothelioma symptoms from Turkey. The patients had prior asbestos exposure and displayed symptoms consistent with mesothelioma. We compared probabilistic neural network, perceptron-based neural network, random forest, one rule, and decision tree classifiers to predict diagnosis of the patient records. We measured classifiers' performance through standard confusion matrix scores such as Matthews correlation coefficient (MCC). Random forest outperformed all models tried, obtaining MCC = +0.37 on the complete imbalanced dataset and MCC = +0.64 on the under-sampled balanced dataset. We then employed random forest feature selection to identify the two most relevant dataset traits associated with mesothelioma: lung side and platelet count. These two risk factors resulted so predictive, that decision tree focusing on them achieved the second top accuracy on the complete dataset diagnosis prediction (MCC = +0.28), outperforming all other methods and even decision tree itself applied to all features.

Conclusions

Our results show that machine learning can predict diagnoses of patients having mesothelioma symptoms with high accuracy, sensitivity, and specificity, in few minutes. Additionally, random forest can efficiently select the most important features of this clinical dataset (lung side and platelet count) in few seconds. The importance of pleural plaques in lung sides and blood platelets in mesothelioma diagnosis indicates that physicians should focus on these

OPEN ACCESS

Citation: Chicco D, Rovelli C (2019) Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PLoS ONE 14 (1): e0208737. https://doi.org/10.1371/journal.pone.0208737

Editor: Alfredo Vellido, Universitat Politecnica de Catalunya, UNITED STATES

Received: June 12, 2018

Accepted: November 22, 2018

Published: January 10, 2019

Data Availability Statement: The mesothelioma dataset file analyzed in the manuscript is currently available publically on the University of California Irvine Machine Learning Repository, at the following URL: https://archive.ics.uci.edu/ml/datasets/Mesothelioma%C3%A2%E2%82%AC%E2%84%A2s+disease+data+set+ The code software developed for the diagnosis prediction and feature selection described in this paper is publically available on Github.com at the following URL: https://github.com/davidechicco/mesothelioma-diagnosis-predictions.

Problem: mesothelioma

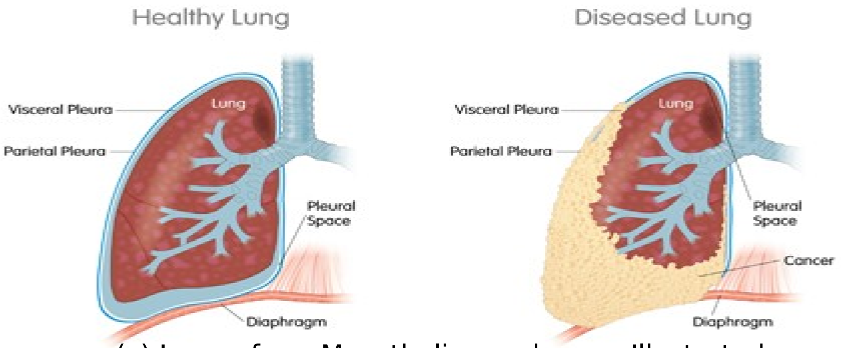
- problem
- dataset analysis
- machine learning diagnosis prediction
- diagnosis prediction results
- machine learning feature selection
- feature selection results

Mesothelioma

Mesothelioma is a lung cancer affecting millions of people around the world, especially those with historical exposure to asbestos and amianthus

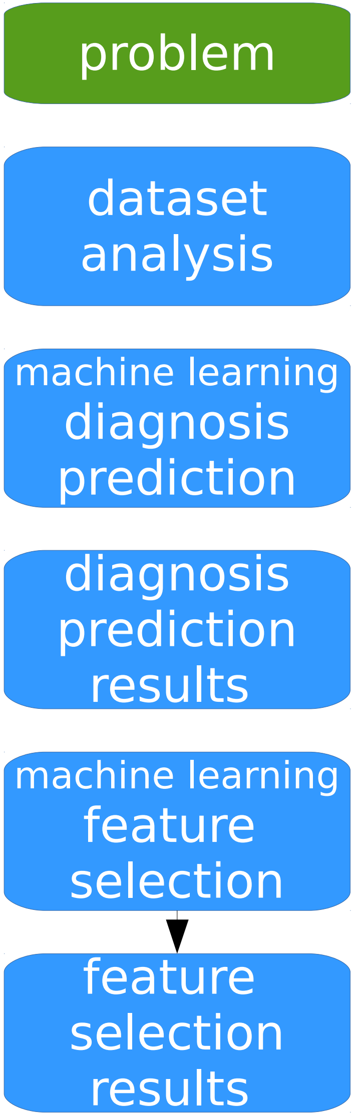
Mesotheliomas are always malignant, but some patients with mesothelioma symptoms might have pleural placques instead

The most important symptoms are pain, dyspnea (shortness of breath), cough, pain and dry cough, pleural effusion, chest pain, and shoulder pain



(c) Image from Mesothelioma.uk.com: Illustrated image of healthy and disease lung

(c) mesotheliomalawyercenter.org



problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Mesothelioma

In more advanced stages, other symptoms can show up: weakness, fever, hoarseness, hypoxemia (scarce oxygen in the blood), dysphagia (difficulty swallowing), fever, night sweats, and weight loss

Differently from symptoms, clinical features provide quantitative information about the trend of the disease in the patients

The most important feature is the occupational history of the patient, because it can show previous exposure to asbestos. In fact, if the worker has been working close to asbestos or amianthus for long, it is extremely likely that he might have developed pleural mesothelioma

(c) Image from GreeneEnvironmentalServices.com:
an amianthus/asbestos material

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

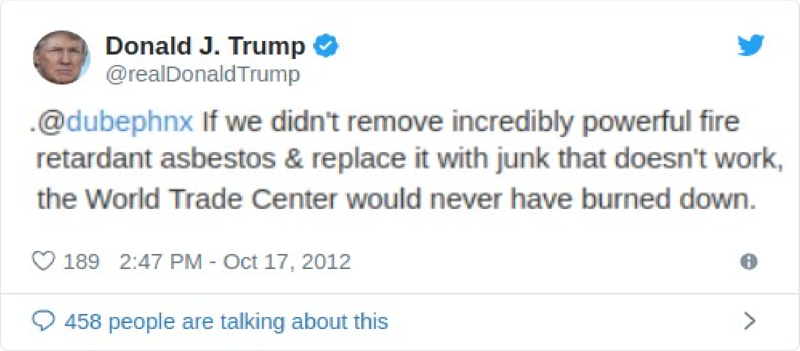
machine learning feature selection

feature selection results

Asbestos causes mesothelioma

Mesothelioma is a major type of lung cancer. Incidence varies markedly by country. Between 2004 and 2008, 23,869 people in the Americas, 49,779 people in Europe, and 12,012 people in Asia died of mesothelioma (World Health Organization reports).

Some people don’t undestand the lethal effects of asbestos:



Donald J. Trump
@realDonaldTrump

.@dubephnx If we didn't remove incredibly powerful fire retardant asbestos & replace it with junk that doesn't work, the World Trade Center would never have burned down.

189 2:47 PM - Oct 17, 2012

458 people are talking about this

This is wrong! The World Trade Centre would have collapes anyway and mesothelioma kills ∼ 5,000 people every year in the Americas

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

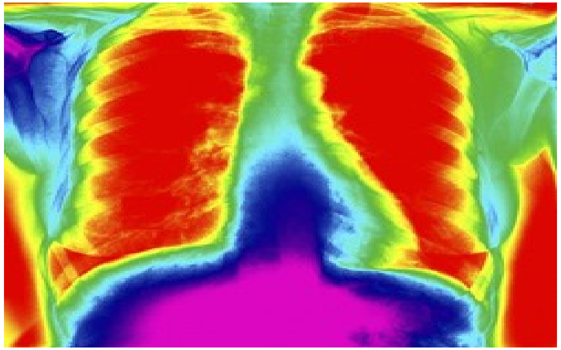
feature selection results

Mesothelioma

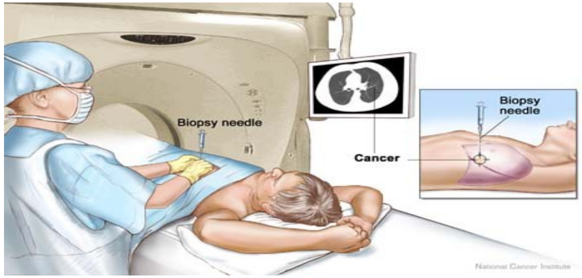
Mesothelioma diagnosis generally requires expensive imaging and laboratory medicine resources, such as medical imaging techniques (X-rays, magnetic resonance imaging (MRI) and positron emission tomography (PET) scans), biopsies, and blood tests.

(c) Lung x-rays exam
Medscape.com

(c) MRI for mesothelioma
radiopaedia.org



(c) PET scan for mesothelioma
asbestoslaw.com



(c) Lung biopsy
lhsc.on.ca

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Mesothelioma

Computational machine learning techniques can provide more effective, cheaper, and faster alternative methods for diagnosis of patients by processing their health record datasets, containing clinical data and laboratory test results

Mesothelioma dataset

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

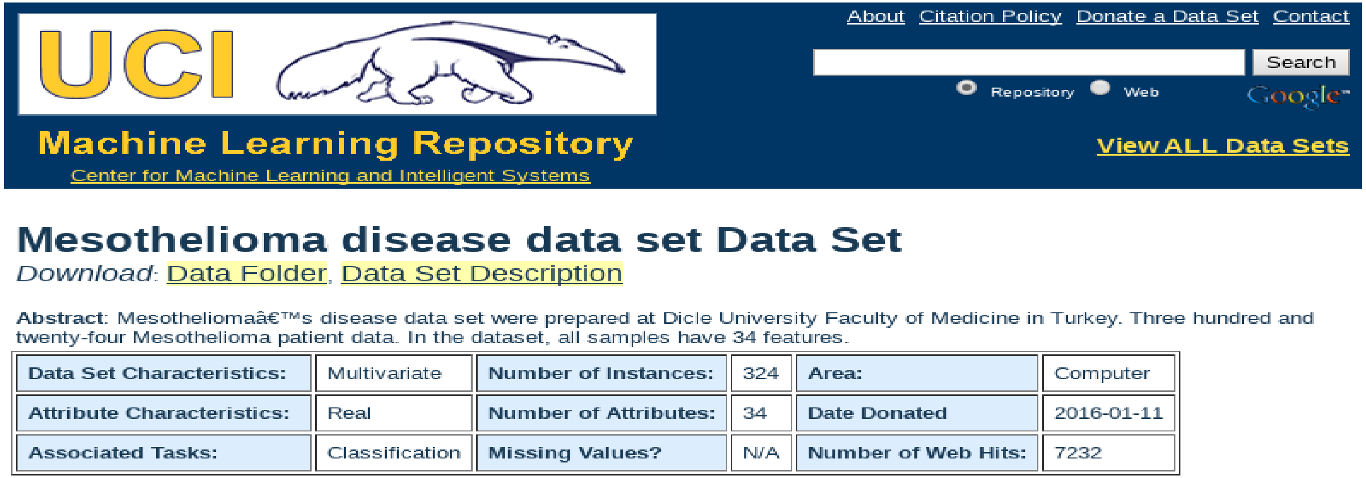
machine learning feature selection

feature selection results

Mesothelioma dataset

- Originally released with the paper "An approach based on probabilistic neural network for diagnosis of Mesotheliomas disease", by Orhan Er and colleagues. Computers & Electrical Engineering, 38(1):75–81, 2012

- Dataset from the Faculty of Medicine, Dicle University, Diyarbakır, Turkey



Mesothelioma disease data set Data Set

Download: Data Folder, Data Set Description

Abstract: Mesotheliomaâ€™s disease data set were prepared at Dicle University Faculty of Medicine in Turkey. Three hundred and twenty-four Mesothelioma patient data. In the dataset, all samples have 34 features.

Data Set Characteristics:	Multivariate	**Number of Instances:**	324	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	34	**Date Donated**	2016-01-11
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	7232

(c) Screenshot from the University of California Irvine Machine Learning Repository website https://archive.ics.uci.edu/ml/datasets

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

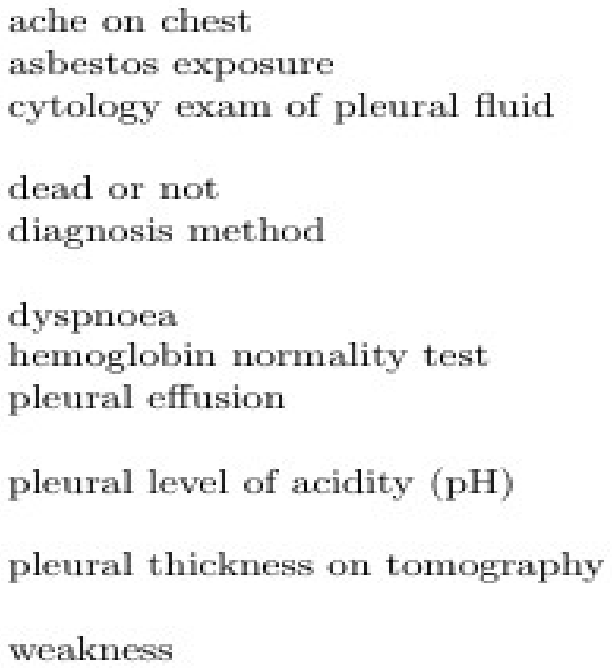
feature selection results

Mesothelioma dataset

The first step has been the analysis of the meaning of the dataset features (324 patients, 33 features, and 1 target=healthy/sick).

Imbalance: 96 sick patients (29.63%) and 228 healthy patients (70.37%)

boolean features


ache on chest
asbestos exposure
cytology exam of pleural fluid

dead or not
diagnosis method

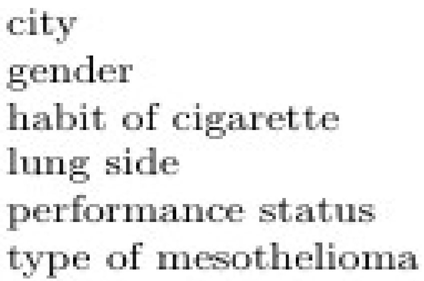
dyspnoea
hemoglobin normality test
pleural effusion

pleural level of acidity (pH)

pleural thickness on tomography

weakness

category features


city
gender
habit of cigarette
lung side
performance status
type of mesothelioma

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Mesothelioma dataset

The first step has been the analysis of the meaning of the dataset features (324 patients, 33 features, and 1 target=healthy/sick).

Imbalance: 96 sick patients (29.63%) and 228 healthy patients (70.37%)

time features

age

duration of asbestos exposure

duration of symptoms

target

healthy / unhealthy

real valued features

albumin
alkaline phosphatise (ALP)
C-reactive protein (CRP)

glucose
lactate dehydrogenase test (LDH)

platelet count (PLT)
pleural albumin
pleural fluid WBC count
pleural fluid glucose
pleural lactic dehydrogenase
pleural protein

sedimentation rate
total protein
white blood cells (WBC)

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Problem framing

target **healthy / unhealthy**

Each patient of the dataset has a target feature that can be:

- unhealthy = sick = mesothelioma
- healthy = non-mesothelioma

This case can be considered a typical binary classification supervised learning problem

- problem
- dataset analysis
- machine learning diagnosis prediction
- diagnosis prediction results
- machine learning feature selection
- feature selection results

Diagnosis prediction & feature selection

- First part of the project: machine learning methods to predict the patients diagnosis (healthy / unhealthy)

- Second part of the project: feature selection to detect the most relevant features in the dataset

Diagnosis prediction

problem

dataset analysis

machine learning diagnosis prediction

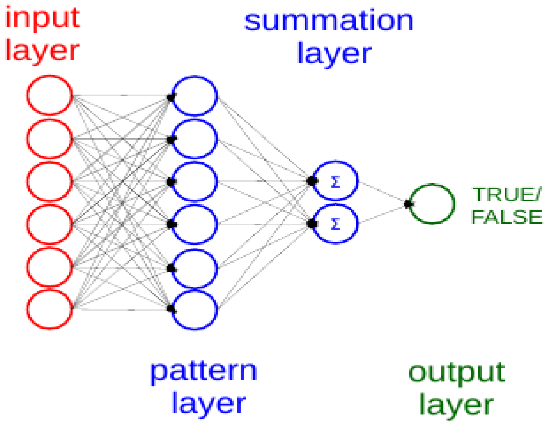
diagnosis prediction results

machine learning feature selection

feature selection results

Diagnosis prediction methods

- Probabilistic neural network (PNN), method used in the original paper



python
NeuPy

- Unlike the classical multi-layer perceptron, the probabilistic neural network computes the output values as an estimation of probability of class membership.

- A lazy learning model, meaning that it does include an iterative training procedure. When using a PNN, we do not train the neurons' weights, but rather assign values to them.

problem

dataset analysis

machine learning diagnosis prediction

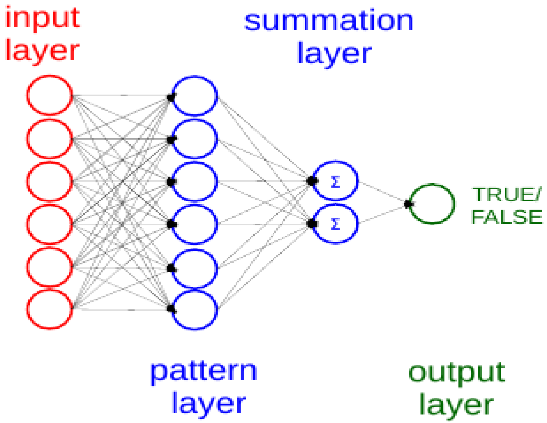
diagnosis prediction results

machine learning feature selection

feature selection results

Diagnosis prediction methods

- Probabilistic neural network (PNN), method used in the original paper



python NeuPy

- The input layer reads the input values, while the pattern layer computes the radial distance between each pair of input neurons, through a Gaussian function (standard deviation = 0.1).

- In the summation layer, the neural network sums all the values outputted by the previous layer, generating probability values that estimate the likely of class membership in the output layer. In a supervised binary case, the method assigns each value to the most likely boolean category (true or false).

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

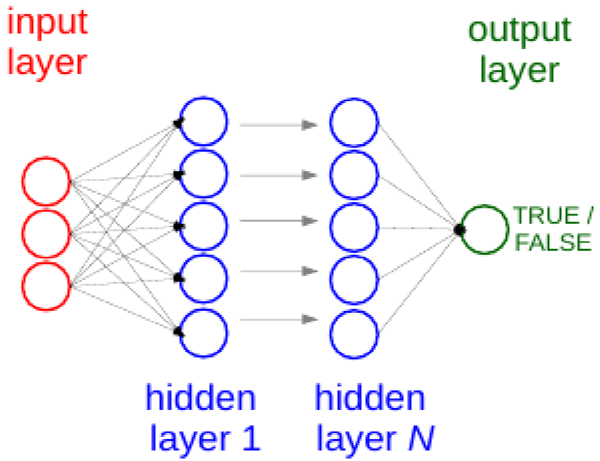
machine learning feature selection

feature selection results

Diagnosis prediction methods

- Perceptron neural network

torch
Optim package



- The main difference between a perceptron and a probabilistic neural network comes from back-propagation.
- In the perceptron, once the values cross the neural network and reach the output layer, the neural network computes the mean square error between the predicted values and the gold-standard values. Afterwards, the algorithm sends this error measure back to neurons of each hidden layer, through a technique called back-propagation, and they update their weights accordingly.

problem

dataset analysis

machine learning diagnosis prediction

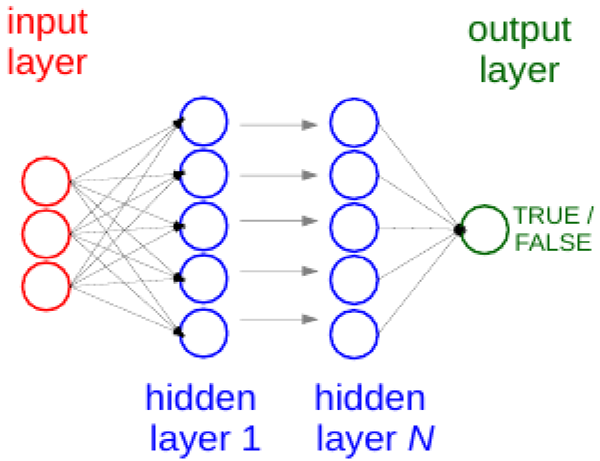
diagnosis prediction results

machine learning feature selection

feature selection results

Diagnosis prediction methods

- Perceptron neural network



torch
Optim package

- learning rate = 0.01
- iterations = 200
- momentum with alpha = 0.5
- dropout
- likelihood threshold = 0.5
- optimized hyper-parameters: hidden layers from 1 to 5, and hidden units 25 to 300 (step = 25)
- training set (179 patients), validation set (45), test set (100)

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

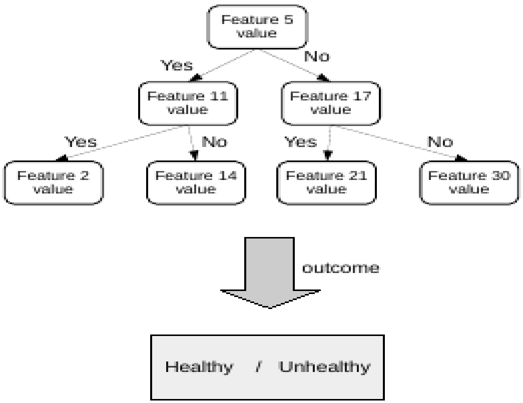
machine learning feature selection

feature selection results

Diagnosis prediction methods

randomForest package

Decision tree



- A decision tree is a classification model in which every node is a decision function, and the node child represents every potential choice from that decision

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

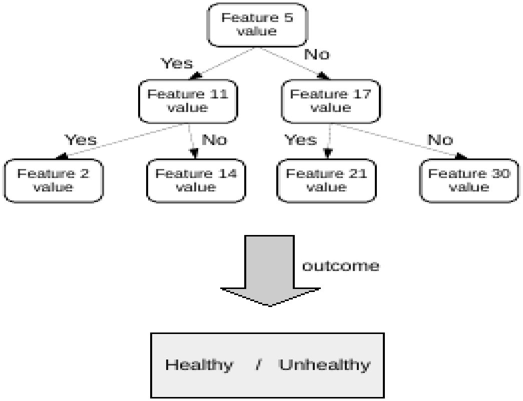
machine learning feature selection

feature selection results

Diagnosis prediction methods

randomForest package

Decision tree



- The tree applies the decision function of each node repeatedly to the input, and then associates the data sample to the corresponding child

- Afterwards, the child also applies its decision function to the input, and associates it to one of its children-nodes, and so on. The algorithm repeats this procedure until it reaches the end

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

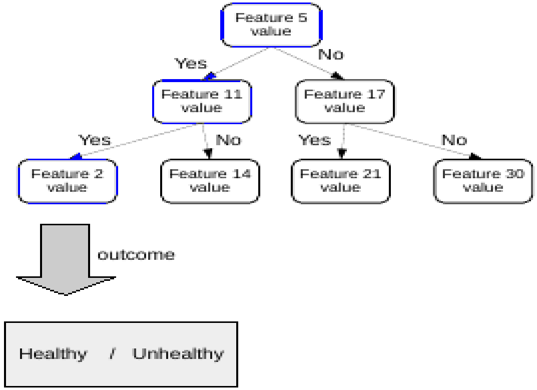
machine learning feature selection

feature selection results

Diagnosis prediction methods

randomForest package

One rule



This method generates one rule for each possible rule in the data, then selects the rule that generated the lowest error rate, and selects that as the “one rule” to follow.

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

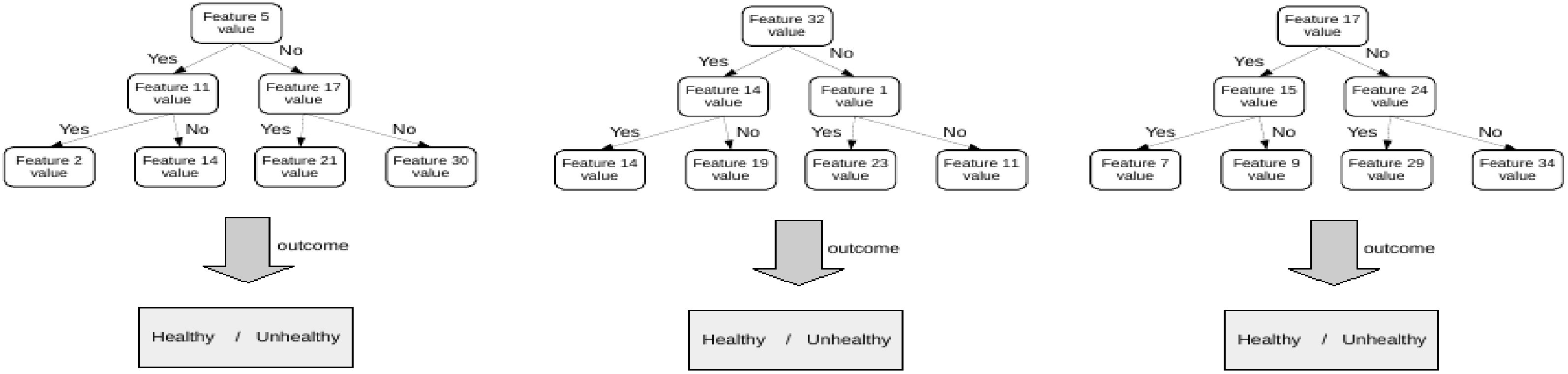
machine learning feature selection

feature selection results

Diagnosis prediction methods

Random forest

randomForest package



- Random forest is an ensemble learning method: a set of predictive decision trees maps each input into an output category, by processing it through its leaves

- Random decision trees are applied to subsets of features and subsets of data instances. In the end, the algorithm computes the majority vote between all the trees

problem

dataset analysis

machine learning diagnosis prediction

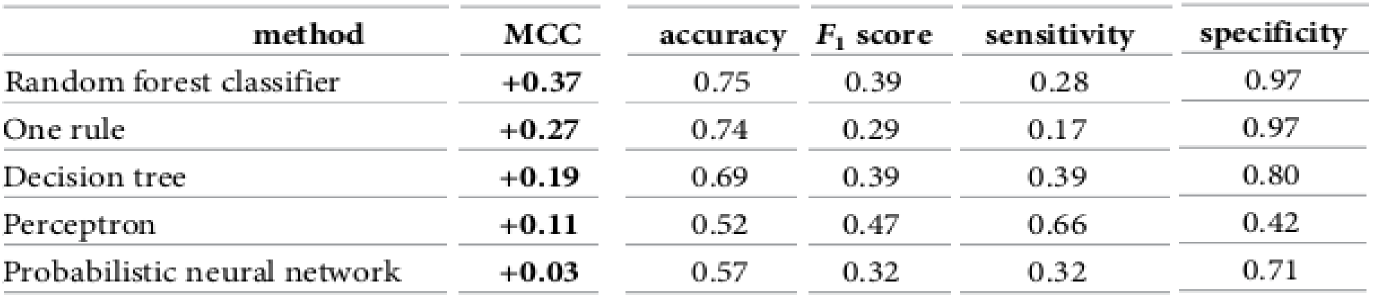
diagnosis prediction results

machine learning feature selection

feature selection results

Diagnosis prediction methods

On the complete dataset (324 patients)



				true positive rate	true negative rate
method	**MCC**	**accuracy**	**F_1 score**	**sensitivity**	**specificity**
Random forest classifier	**+0.37**	0.75	0.39	0.28	0.97
One rule	**+0.27**	0.74	0.29	0.17	0.97
Decision tree	**+0.19**	0.69	0.39	0.39	0.80
Perceptron	**+0.11**	0.52	0.47	0.66	0.42
Probabilistic neural network	**+0.03**	0.57	0.32	0.32	0.71
possible values	[-1; +1]	[0; 1]	[0; 1]	[0; 1]	[0; 1]

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

worst value $= -1$; best value $= +1$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

worst value $= 0$; best value $= 1$

problem

dataset analysis

machine learning diagnosis prediction

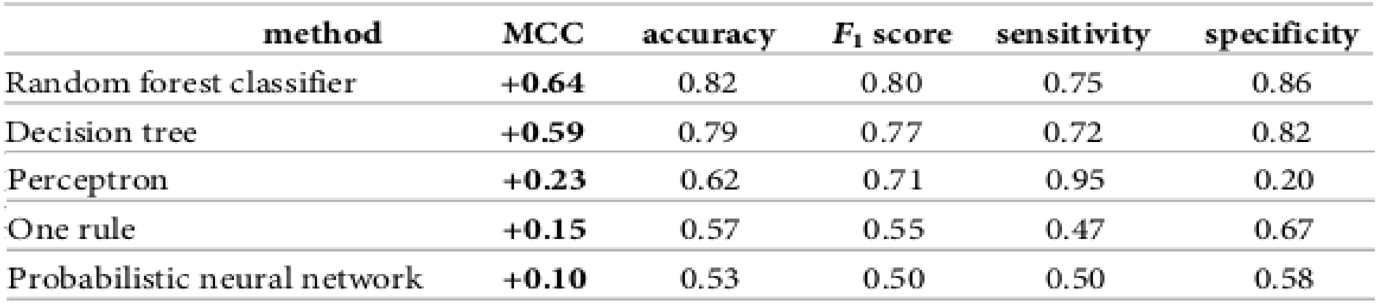
diagnosis prediction results

machine learning feature selection

feature selection results

Diagnosis prediction methods

On the undersampled dataset (192 patients: 50% positives and 50% negatives)



				true positive rate	true negative rate
method	MCC	accuracy	F_1 score	sensitivity	specificity
Random forest classifier	**+0.64**	0.82	0.80	0.75	0.86
Decision tree	**+0.59**	0.79	0.77	0.72	0.82
Perceptron	**+0.23**	0.62	0.71	0.95	0.20
One rule	**+0.15**	0.57	0.55	0.47	0.67
Probabilistic neural network	**+0.10**	0.53	0.50	0.50	0.58
possible values	[-1; +1]	[0; 1]	[0; 1]	[0; 1]	[0; 1]

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

worst value $= -1$; best value $= +1$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

worst value $= 0$; best value $= 1$

– TDLS break -

Feature selection

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

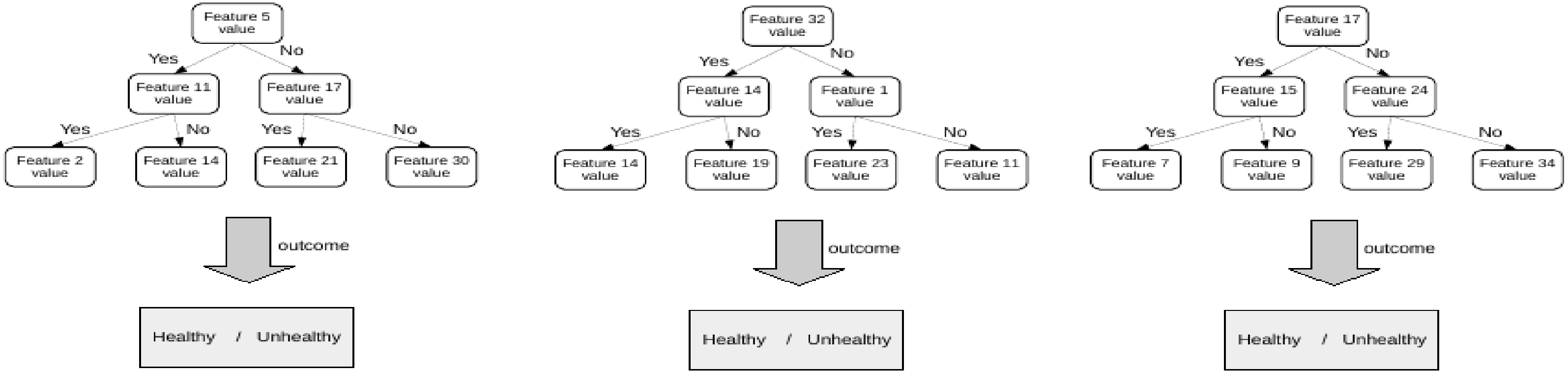
machine learning feature selection

feature selection results

Feature selection method

Random forest feature selection

randomForest package



- Random forest can be applied for feature selection, too
- The method predicts the diagnosis by excluding one feature at each time
- The decrease in the accuracy (mean square error) indicates the “statistical importance” of that feature in the dataset
- The decrease of the Gini purity indicates the “informative importance” of that feature in the dataset

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

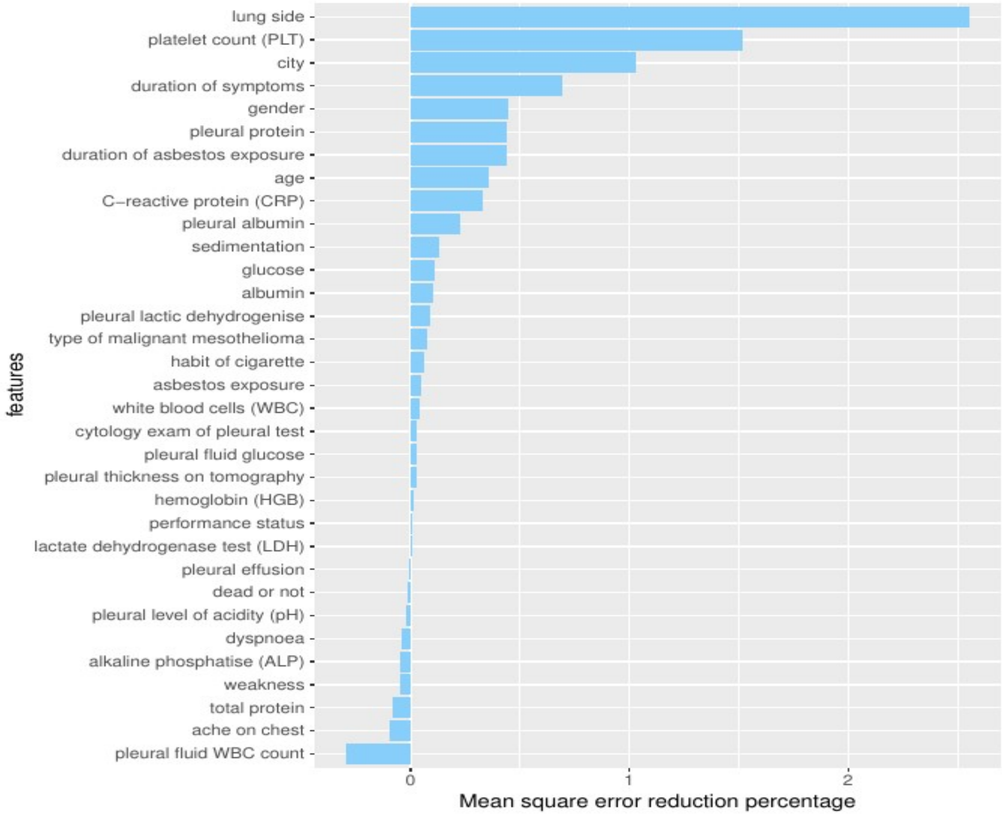
machine learning feature selection

feature selection results

Feature selection results

- Random forest: Mean square error (MSE) decrease in accuracy

- Measure of the accuracy decrease in the random forest prediction when a specific feature is removed



problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

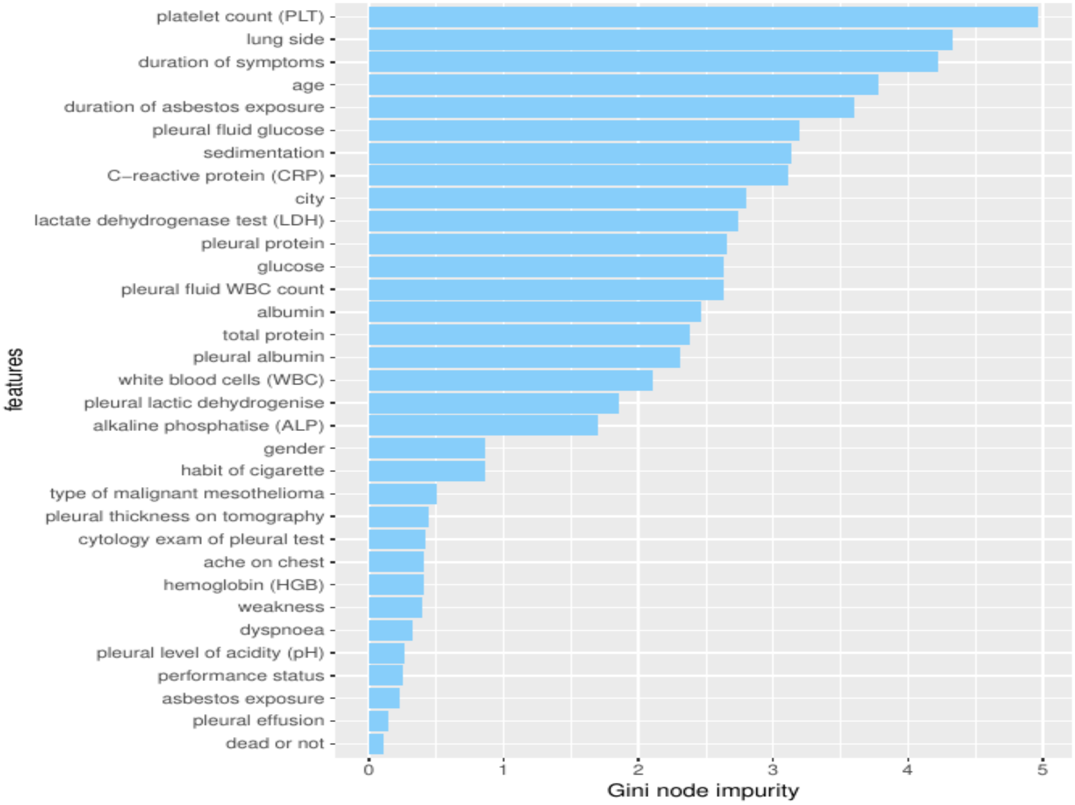
machine learning feature selection

feature selection results

Feature selection results

- Random forest: Gini impurity increase

- Measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset



problem

dataset analysis

machine learning diagnosis prediction

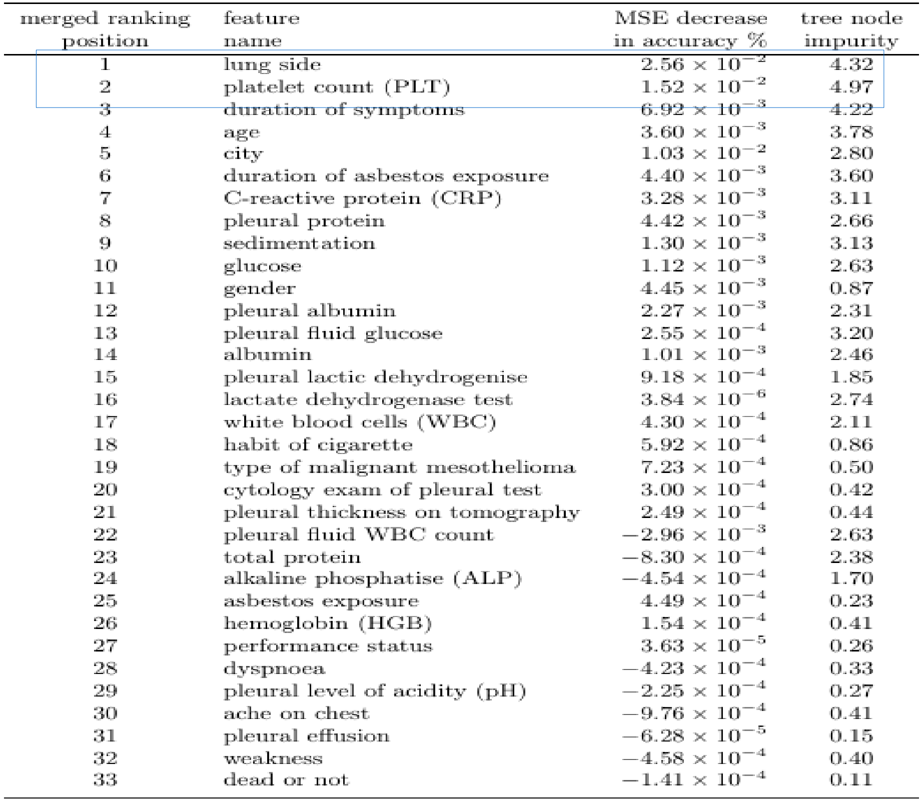
diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

- Random forest: merged ranking of selected features



merged ranking	feature	MSE decrease	tree node
position	name	in accuracy %	impurity
1	lung side	2.56×10^{-2}	4.32
2	platelet count (PLT)	1.52×10^{-2}	4.97
3	duration of symptoms	6.92×10^{-3}	4.22
4	age	3.60×10^{-3}	3.78
5	city	1.03×10^{-2}	2.80
6	duration of asbestos exposure	4.40×10^{-3}	3.60
7	C-reactive protein (CRP)	3.28×10^{-3}	3.11
8	pleural protein	4.42×10^{-3}	2.66
9	sedimentation	1.30×10^{-3}	3.13
10	glucose	1.12×10^{-3}	2.63
11	gender	4.45×10^{-3}	0.87
12	pleural albumin	2.27×10^{-3}	2.31
13	pleural fluid glucose	2.55×10^{-4}	3.20
14	albumin	1.01×10^{-3}	2.46
15	pleural lactic dehydrogenise	9.18×10^{-4}	1.85
16	lactate dehydrogenase test	3.84×10^{-6}	2.74
17	white blood cells (WBC)	4.30×10^{-4}	2.11
18	habit of cigarette	5.92×10^{-4}	0.86
19	type of malignant mesothelioma	7.23×10^{-4}	0.50
20	cytology exam of pleural test	3.00×10^{-4}	0.42
21	pleural thickness on tomography	2.49×10^{-4}	0.44
22	pleural fluid WBC count	-2.96×10^{-3}	2.63
23	total protein	-8.30×10^{-4}	2.38
24	alkaline phosphatise (ALP)	-4.54×10^{-4}	1.70
25	asbestos exposure	4.49×10^{-4}	0.23
26	hemoglobin (HGB)	1.54×10^{-4}	0.41
27	performance status	3.63×10^{-5}	0.26
28	dyspnoea	-4.23×10^{-4}	0.33
29	pleural level of acidity (pH)	-2.25×10^{-4}	0.27
30	ache on chest	-9.76×10^{-4}	0.41
31	pleural effusion	-6.28×10^{-5}	0.15
32	weakness	-4.58×10^{-4}	0.40
33	dead or not	-1.41×10^{-4}	0.11

1 lung side
2 platelet count (PLT)

problem

dataset analysis

machine learning diagnosis prediction

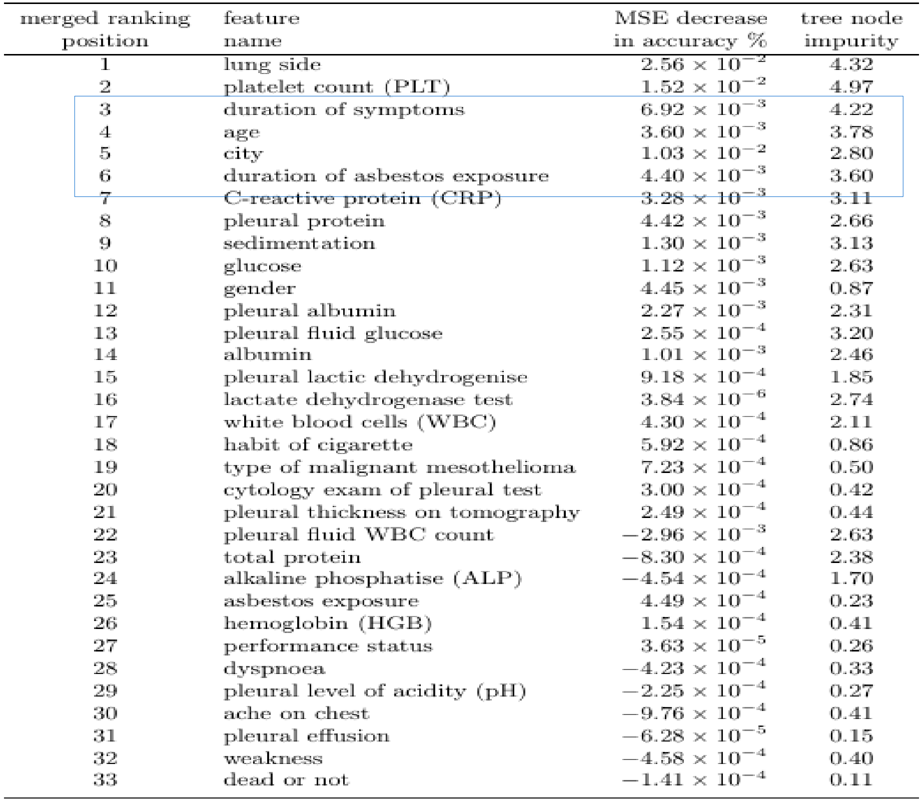
diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

- Random forest: merged ranking of selected features



merged ranking position	feature name	MSE decrease in accuracy %	tree node impurity
1	lung side	2.56×10^{-2}	4.32
2	platelet count (PLT)	1.52×10^{-2}	4.97
3	duration of symptoms	6.92×10^{-3}	4.22
4	age	3.60×10^{-3}	3.78
5	city	1.03×10^{-2}	2.80
6	duration of asbestos exposure	4.40×10^{-3}	3.60
7	C-reactive protein (CRP)	3.28×10^{-3}	3.11
8	pleural protein	4.42×10^{-3}	2.66
9	sedimentation	1.30×10^{-3}	3.13
10	glucose	1.12×10^{-3}	2.63
11	gender	4.45×10^{-3}	0.87
12	pleural albumin	2.27×10^{-3}	2.31
13	pleural fluid glucose	2.55×10^{-4}	3.20
14	albumin	1.01×10^{-3}	2.46
15	pleural lactic dehydrogenise	9.18×10^{-4}	1.85
16	lactate dehydrogenase test	3.84×10^{-6}	2.74
17	white blood cells (WBC)	4.30×10^{-4}	2.11
18	habit of cigarette	5.92×10^{-4}	0.86
19	type of malignant mesothelioma	7.23×10^{-4}	0.50
20	cytology exam of pleural test	3.00×10^{-4}	0.42
21	pleural thickness on tomography	2.49×10^{-4}	0.44
22	pleural fluid WBC count	-2.96×10^{-3}	2.63
23	total protein	-8.30×10^{-4}	2.38
24	alkaline phosphatise (ALP)	-4.54×10^{-4}	1.70
25	asbestos exposure	4.49×10^{-4}	0.23
26	hemoglobin (HGB)	1.54×10^{-4}	0.41
27	performance status	3.63×10^{-5}	0.26
28	dyspnoea	-4.23×10^{-4}	0.33
29	pleural level of acidity (pH)	-2.25×10^{-4}	0.27
30	ache on chest	-9.76×10^{-4}	0.41
31	pleural effusion	-6.28×10^{-5}	0.15
32	weakness	-4.58×10^{-4}	0.40
33	dead or not	-1.41×10^{-4}	0.11

3 duration of symptoms
4 age
5 city
6 duration of asbestos exposure

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

Random forest discoveries: lung side and platelet count

- The presence of mesothelioma on both lung sides is highly predictive for malignancy

In fact, 72% of patients having “lung side = 2” are reported to be unhealthy

- Low platelet count (PLT) is strongly related to mesothelioma.

In fact, 54.76% of the patients having platelet count lower than 150 k/microliter are reported to be unhealthy

- Duration of symptoms, age, city, and duration of asbestos exposure can cause mesothelioma malignancy

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

Prediction by using only the two most relevant features

- We tried to make a prediction of the diagnoses by using only the two features recognized as most relevant (lung side and platelet count)

- We excluded all the other features and applied a classification and regression tree (CART) to all the patients

- The prediction results showed Matthews correlation coefficient +0.28 on average on the complete dataset and +0.41 on the undersampled dataset

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

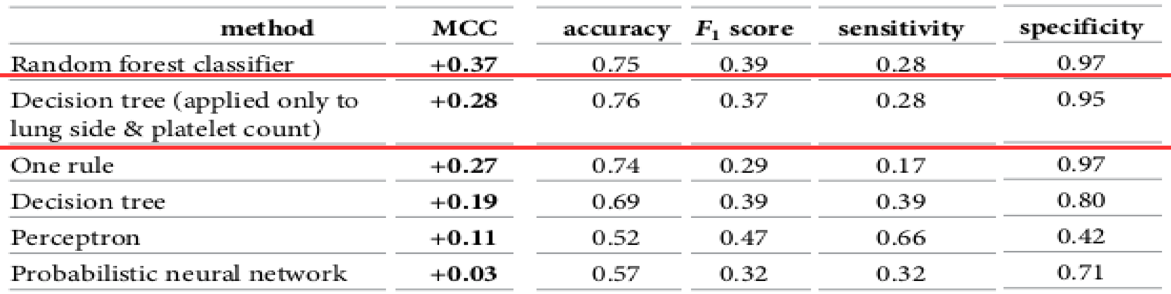
Prediction by using only the two most relevant features

- Therefore, even if lung side and platelet count are the most relevant features, they are sufficient for an accurate prediction

true positive rate

true negative rate

complete dataset



method	MCC	accuracy	F_1 score	sensitivity	specificity
Random forest classifier	**+0.37**	0.75	0.39	0.28	0.97
Decision tree (applied only to lung side & platelet count)	**+0.28**	0.76	0.37	0.28	0.95
One rule	**+0.27**	0.74	0.29	0.17	0.97
Decision tree	**+0.19**	0.69	0.39	0.39	0.80
Perceptron	**+0.11**	0.52	0.47	0.66	0.42
Probabilistic neural network	**+0.03**	0.57	0.32	0.32	0.71

possible values [-1; +1] [0; 1] [0; 1] [0; 1] [0; 1]

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Feature selection results

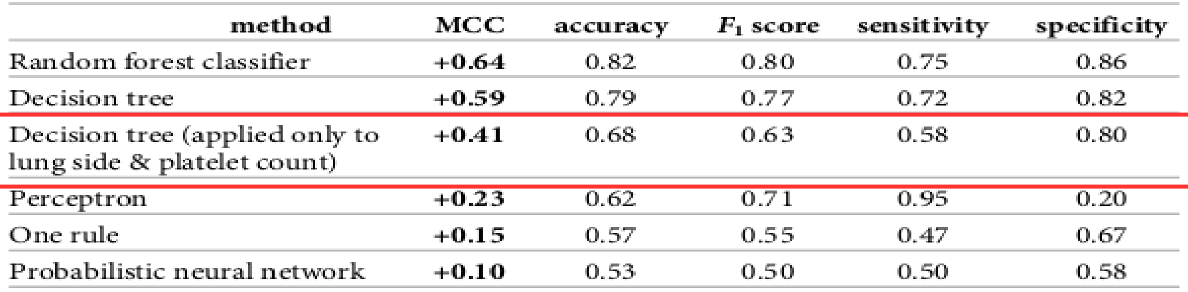
Prediction by using only the two most relevant features

- Therefore, even if lung side and platelet count are the most relevant features, they are sufficient for an accurate prediction

under-sampled dataset

true positive rate

true negative rate



method	MCC	accuracy	F_1 score	sensitivity	specificity
Random forest classifier	**+0.64**	0.82	0.80	0.75	0.86
Decision tree	**+0.59**	0.79	0.77	0.72	0.82
Decision tree (applied only to lung side & platelet count)	**+0.41**	0.68	0.63	0.58	0.80
Perceptron	**+0.23**	0.62	0.71	0.95	0.20
One rule	**+0.15**	0.57	0.55	0.47	0.67
Probabilistic neural network	**+0.10**	0.53	0.50	0.50	0.58

possible values [-1; +1] [0; 1] [0; 1] [0; 1] [0; 1]

problem

dataset analysis

machine learning diagnosis prediction

diagnosis prediction results

machine learning feature selection

feature selection results

Conclusions

- Our neural networks and random forest classifier were able to predict mesothelioma patients’ diagnosis with higher accuracy than probabilistic neural networks

- Our feature selection method was able to identify key causes of mesothelioma (presence of mesothelioma traces in both lung sides, platelet count, duration of symptoms, age, city, and duration of asbestos) in just few seconds

- The paper has been published on the PLOS One journal, within the special collection on machine learning in biomedicine https://collections.plos.org/mlforhealth

- All the code is online on my Github: https://github.com/davidechicco

Discussion points

by Shazia Akbar (1/2 part)

- As mentioned this is a small dataset with few very features. (Understandably because it's a very specific and targeted medical problem). Perhaps a follow-up study could explore detecting and preventing overfitting to this particular dataset. e.g. using the same perceptron model but adding dropout and batch normalization.

- I noticed that the population used to curate this dataset was also fairly selective and included some preconceptions and prior knowledge i.e. patients with mesothelioma were slightly younger (see page 20). Could we use that knowledge to make the predictions models even better?

- The paper mentions that one method was good for identifying true positives and the other for true negatives. Could we combine both methods together to create an ensemble method?

Discussion points

by Shazia Akbar (2/2 part)

- Are the features inputted into the system manually and are they likely to be accurate? For example, would "duration of asbestos exposure" be accurate to 2 decimal places or rounded up?

- Which features were highly correlated - this may be in the paper and I overlooked it

- There was severe imbalance in the dataset and you chose one way to overcome this. What other methods could we use and would they give the same outcome?

- Note: I was unclear what exactly the difference is between the perceptron model and probabilistic neural network. Can you clarify during the talk?

Discussion points

by Nassim Tayari

Objective and study design:

- What is most important to end user? True positive rate or true negative rate?

Data:

- Is there a selection bias? The patients’ population age range?
- How did you handle the imbalanced data issue?
- Lack of follow up study data and longitudinal studies
- Size of data. How one can conclude that data size was sufficient?

Methods:

- Feature selection only based on statistical methods
- Feature correlation analysis
- Why did you use neural networks if the classical methods like random forest obtained better results?
- Is there a benchmark? Other studies?

Thank you!

Believe It.

WE WILL CONQUER CANCER
IN OUR LIFETIME

The Princess Margaret