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

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.

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

Funding: This project was supported by the Peter Munk Cardiac Centre. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

two features when reading records of patients with mesothelioma symptoms. Moreover, doctors can exploit our machinery to predict patient diagnosis when only lung side and platelet data are available.

1 Introduction

Mesothelioma is a major type of lung cancer. Incidence varies markedly by country [1, 2]. Between 2004 and 2008, 23,869 people in the Americas, 49,779 people in Europe, and 12,012 people in Asia died of mesothelioma [3].

Pleural mesothelioma makes up approximately 75% of all mesotheliomas, and affects the two membranes of the lung: the visceral pleura and parietal pleura. Other subtypes include pericardial mesothelioma, which develops in the membrane around the heart, the pericardium. In many cases, pericardial mesothelioma goes undiagnosed until autopsy [4]. Mesotheliomas are always malignant, but some patients with mesothelioma symptoms might have pleural plaques instead [5], without mesothelioma.

The most important symptoms for diagnosis include pain, dyspnoea (shortness of breath), cough, pain and dry cough, pleural effusion, chest pain, and shoulder pain. In more advanced stages, other symptoms can show up: weakness, fever, hoarseness, hypoxemia (low level of oxygen in the blood), dysphagia (difficulty swallowing), fever, night sweats, and weight loss [6, 7].

In contrast with symptoms, clinical features provide quantitative information to aid diagnosis. Existing models that forecast patient survival use clinical features such as histologic subtype, time since diagnosis, platelet count, hemoglobin, and disease stage [6]. In mesothelioma, occupational history generally can serve as a particularly informative feature, as it shows previous exposure to asbestos. Long workplace exposure to asbestos makes development of pleural mesothelioma extremely likely [8].

Mesothelioma diagnosis generally requires expensive imaging and laboratory medicine resources [9], such as X-rays, magnetic resonance imaging (MRI) and positron emission tomography (PET) scans, biopsies, and blood tests. Even if precise and efficient, the medical imaging machines are expensive and uncommon in remote regions. Medical tests like biopsies, in addition, are quite invasive and painful for patients.

To speed diagnosis and minimize the use of these tests, researchers have used machine learning methods to solve health informatics classification tasks [10]. Machine learning methods provide useful tools to classify, process, and analyze health records in minutes or seconds.

We analyzed a dataset of mesothelioma health records of 324 mesothelioma patients from the Diyarbakır region of southeast Turkey [11]. This area has endemic, natural asbestos fibers in the soil, mostly tremolite fibers, but also chrysotile fibers. This provides a unique dataset with a high incidence of mesothelioma within a population of highly asbestos-exposed individuals. From this dataset, we ascertained risk factors for mesothelioma. Each patient record in this dataset contains multiple clinical features and a diagnosis label. The diagnosis label has two categories: mesothelioma or non-mesothelioma. The “non-mesothelioma” patients have similar clinical features as those with mesothelioma, such as pleural plaques. Nonetheless, physicians did not diagnose these patients with mesothelioma.

It is often difficult to distinguish clinically between patients with mesothelioma with asbestos-exposed individuals who have pleural plaques and clinical features suggestive of mesothelioma, but who lack the disease. Asbestos exposure itself can lead to pleural plaque development, pleural effusions, and other radiologic changes that mimic mesothelioma.

Machine learning methods are well-established in scientific research for cancer predictive diagnosis [12–15]. The mesothelioma dataset used here comes from a previous effort to diagnose mesothelioma using a probabilistic neural network (PNN) [10]. Probabilistic neural networks have also been used to diagnose potential cancer patients [16–18], and to predict anti-HIV drugs [19]. To replicate the approach used by the original dataset authors [10], we started this study by reimplementing a probabilistic neural network [20], and then compared this algorithm with other machine learning models such as artificial neural networks [21, 22], random forests [23, 24], decision trees [25], and one rule [26].

We chose these methods because they are particularly appropriate for the dataset we analyze, and because they have proven successful and suitable in solving similar health informatics problems in the past [27, 28]. Artificial neural networks, for example, have been used to predict the sequence specificity of DNA-binding and RNA-binding proteins [29], and classify micrographs of breast cancer [30]. Random forests, also, have seen extensive use in bioinformatics and biomedical informatics contexts [31–33], such as for the classification of gene expression microarray data [34]. Additionally, researchers have used random forests to classify other cancer types, including renal cell carcinoma data [35], and lung cancer data [36]. Even if machine learning experts often suggest to start with a simple machine learning algorithm [37], such as logistic regression, we decided to avoid this method because it can be imprecise when applied to data having highly correlated features [38]. The mesothelioma dataset, in fact, contains highly correlated data features, generated from the same clinical tests. In addition to random forest's use for classification, we also employed it for feature selection [31, 35, 39], to understand which patient traits and clinical features best predict mesothelioma.

Our findings and our methods can be useful for physicians and medical doctors, in several contexts. Our discoveries about the importance of lung side and platelet count in the dataset suggest physicians should focus on these two features, when reading the electronic health record of a patient. Additionally, physicians can take advantage of our method to predict if a patient is going to have mesothelioma or not, by inputing his/her clinical profile to our software.

2 Dataset

The dataset consists in real electronic health records of 324 patients collected at the Dicle University Faculty of Medicine Hospital (Diyarbakir, southeastern Turkey), before October 2011 [10]. Of these 324 patients, 96 have mesothelioma, and 228 have do not have mesothelioma. Regarding dataset imbalance, the data contains 29.63% positive data instances (patients with mesothelioma), and 70.37% negative instances (patients without mesothelioma).

We represent the dataset as a table of 324 rows, each row corresponding to one patient with potential mesothelioma symptoms. Each row has 35 columns, representing the observed features for that patient (Table 1). One of the features is the diagnosis label, "class of diagnosis". This feature states whether the patient actually has mesothelioma (1, "mesothelioma" label), or or not (0, "non-mesothelioma" label).

The dataset curators published the first analysis of this dataset in October 2011 [10], and subsequently released the dataset publically on the University of California Irvine Machine Learning Repository in January 2016 [40]. Beyond the data origin, feature names, and their values, Er et al. [10] provided no other details about the dataset. We describe the features in more details here (Tables 1 and 2; S1, S2, S3, S4 and S5 Figs). The "diagnosis method" feature has identical values to "class of diagnosis" and we therefore removed it for classification and feature selection purposes. Of the 33 remaining features, 10 features are boolean, 14 are real

Table 1. Dataset features with ranges and measurement units. We removed “diagnosis method” from the classification and feature selection phases, because it has the same values of “class of diagnosis” target we predict. We changed some feature names to add clarity: “blood lactic dehydrogenise (LDH)” into “lactate dehydrogenase test”, “cell count (WBC)” into “white blood cells (WBC)”, “cytology” into “cytology exam of pleural fluid”, “hemoglobin (HGB)” into “hemoglobin normality test”, “keep side” into “lung side”, “pleural glucose” into “pleural fluid glucose”, and “white blood” into “pleural fluid WBC count”.

feature name	value range	measurement unit
ache on chest	0, 1	boolean
asbestos exposure	0, 1	boolean
cytology exam of pleural fluid	0, 1	boolean
dead or not	0, 1	boolean
diagnosis method	0, 1	boolean
dyspnoea	0, 1	boolean
hemoglobin normality test	0, 1	boolean
pleural effusion	0, 1	boolean
pleural level of acidity (pH)	0, 1	boolean
pleural thickness on tomography	0, 1	boolean
weakness	0, 1	boolean
city	[0, 8]	category
gender	0, 1	category
habit of cigarette	0, 1, 2, 3	category
lung side	0, 1, 2	category
performance status	0, 1	category
type of malignant mesothelioma	0, 1, 2	category
age	[19, 85]	years
duration of asbestos exposure	[0, 70]	years
duration of symptoms	[0.5, 52]	years
albumin	[1.5, 6.9]	g/dL (grams per deciliter)
alkaline phosphatise (ALP)	[41, 489]	IU/L (international units per liter)
C-reactive protein (CRP)	[11, 103]	mg/L (milligrams per liter)
lactate dehydrogenase test (LDH)	[55, 1128]	IU/L (international units per liter)
glucose	[60, 421]	mg/dL (milligrams per deciliter)
platelet count (PLT)	[111, 3335]	kilo platelets per mcL (microliter)
pleural albumin	[0, 4.4]	g/dL (grams per deciliter)
pleural fluid WBC count	[742, 21500]	cells per microliter (mcL)
pleural fluid glucose	[2, 151]	mg/dL (milligrams per deciliter)
pleural lactic dehydrogenase	[110, 7541]	IU/L (international units per liter)
pleural protein	[0, 6.7]	g/L (grams per liter)
sedimentation rate	[7, 129]	mm/hr (millimeters per hour)
total protein	[3.1, 8.5]	g/dL (grams per deciliter)
white blood cells (WBC)	[4, 22]	cells per mcL (microliter)

https://doi.org/10.1371/journal.pone.0208737.t001

values, 3 are time values, and 6 are categorical. We describe the features in depth (Tables 1 and 2, Supplementary Information) and confirmed our interpretation with the dataset curators (Orhan Er, personal communication).

It is also worth noticing that the dataset is well structured and complete, and contains no missing or ambiguous values. The dataset contains only real patients’ data, and no simulations. The completeness of the dataset is a rare quality in electronic health record (EHR) collections, and allows us to make a more precise and accurate analysis than other cases where some data values are missing (for example, [41]).

Table 2. Meaning of each feature of the dataset. We reported a detailed description of each feature in the Supplementary Information.

feature name	meaning
ache on chest	presence or absence of pain in the chest area
asbestos exposure	if a patient has been exposed to asbestos during life
cytology exam of pleural fluid	test to detect cancer cells and certain other cells in the area that surrounds the lung
dead or not	if a patient is still alive
diagnosis method	if the patient has had a mesothelioma diagnosed by a common diagnosis method
dyspnoea	shortness of breath
hemoglobin normality test	test that measures how much hemoglobin is in blood
pleural effusion	presence of effusion, common symptom that can inhibit the normal function of the organ
pleural level of acidity (pH)	if the pleural fluid pH is lower than the normal pleural fluid pH, that it's neutral
pleural thickness of thickness	any form of thickening involving either the parietal or visceral pleura
weakness	lack of strength
city	place of provenance of the patients
gender	female or male
habit of cigarette	four categories for the habit of smoking
lung side	the side of the lungs which is experiencing pleural plaques or mesothelioma traces
performance status	patient's ability to perform normal tasks
type of malignant mesothelioma	mesothelioma stage to which the symptoms seem to belong, according to the TNM Classification of Malignant Tumors
age	the age of the patients
duration of asbestos exposure	how long has been the environmental exposure to asbestos
duration of symptoms	the time period, in years, in which the patients show symptoms
albumin	level of blood albumin
alkaline phosphatase (ALP)	test used to help detect liver disease or bone disorders
C-reactive protein (CRP)	acute phase reactant, significantly elevated in patients with pleural mesothelioma (MPM)
glucose	test which measures the amount of glucose in a sample of blood
lactate dehydrogenase test (LDH)	protein that helps produce energy in the body
platelet count (PLT)	test to measure how many platelets patients have in the blood
pleural albumin	level of albumin in the pleural fluid
pleural fluid WBC count	the count of leukocytes in the pleural fluid
pleural fluid glucose	low level can be linked to infection or malignancy
pleural lactic dehydrogenase	its levels indicates if the fluid is exudate or transudate
pleural protein	pleural effusions are classified as transudates or exudates on the basis of the fluid protein level
sedimentation rate	test to measure how quickly erythrocytes settle in a test tube in one hour
total protein	biochemical test for measuring the total amount of protein in serum
white blood cells (WBC)	test measures the number and quality of white blood cells

https://doi.org/10.1371/journal.pone.0208737.t002

3 Methods

In the first part of the project, we used machine learning to perform a supervised binary prediction of the two possible patient diagnoses (mesothelioma or non-mesothelioma).

To this end, we took advantage of several models. We started with PNN, since it was the method applied by the original dataset authors [10], and we wanted to replicate their approach.

To further investigate the effectiveness of artificial neural networks, we then used a perceptron-based neural network.

Afterwards, we decided to move to tree-like graph models (decision trees, random forest, and one rule), because these methods are unmoved by statistical correlations between features, which are very common in electronic health record datasets. Clinical data, in fact, contain features that have strong relationships between each other, since each aspect of the health of a patient is deeply related to her/his other health aspects, at any level. Tree-like graph models usually are minimally affected by feature correlations, and therefore they can be efficient and robust when applied to patient clinical datasets, as in this case.

In the second part of the project, we investigated the most relevant features associated with mesothelioma. For this purpose, we decided to use random forest feature selection because this method achieved the best results in predicting the diagnosis (Results). We also wanted to take advantage of its ensemble learning approach and importance rates (accuracy decrease and Gini impurity decrease), which let us understand the importance of each feature both statistically and informatively. We decided to avoid employing the other methods for this feature selection phase because they do not provide an informative content such as the Gini impurity decrease [39]. Additionally, even if our feature selection results might be biased towards random forest, we preferred this technique because *bootstrap aggregating* [42] makes ensemble learning methods more robust than neural networks, decision trees, and association rule learning algorithms, regarding feature selection [36, 43, 44].

3.1 Probabilistic neural network

The probabilistic neural network is an artificial neural network algorithm based upon a Bayesian statistical network and a Fisher kernel discriminant analysis model [20] (Fig 1).

A typical artificial neural network contains one input layer, several hidden layers, and an output layer. Each neuron of the input layer contains a value that propagates to the first hidden layer neurons. In feed-forward neural networks (such as probabilistic neural networks and perceptrons), each hidden layer neuron reads the input layer values, multiply them by its weights, sums the temporary results up, applies an activation function, and propagates its result to the next layer of neurons. A multi-layer feed-forward perceptron is a typical artificial neural network, made of one input layer, several hidden layers, and an output layer (Fig 2).

Unlike the classical multi-layer perceptron [45], which has a back-propagation method that updates the weights of the neurons at each iteration, the probabilistic neural network computes as output values as probability of class membership. A probabilistic neural network consists of an input layer, a pattern layer, a summation layer, and an output layer. The input layer reads the input values, while the pattern layer computes the radial distance between the values of each pair of input neurons, through a Gaussian function. In the summation layer, the neural network sums all the values output by the previous layer, generating probability values that estimate the likelihood of class membership in the output layer. For a supervised binary classification, the method assigns each value to the most likely boolean category it can belong, true or false (Fig 1).

This particular artificial neural network is a lazy learning model, meaning that it does include an iterative training procedure. When using a probabilistic neural network, we do not train the neurons' weights, but rather assign values to them (Methods).

Following the strategy initially adopted by the dataset curators [10], we implemented and tested a probabilistic neural network. We set the model Gaussian function to have a standard deviation value of 0.1.

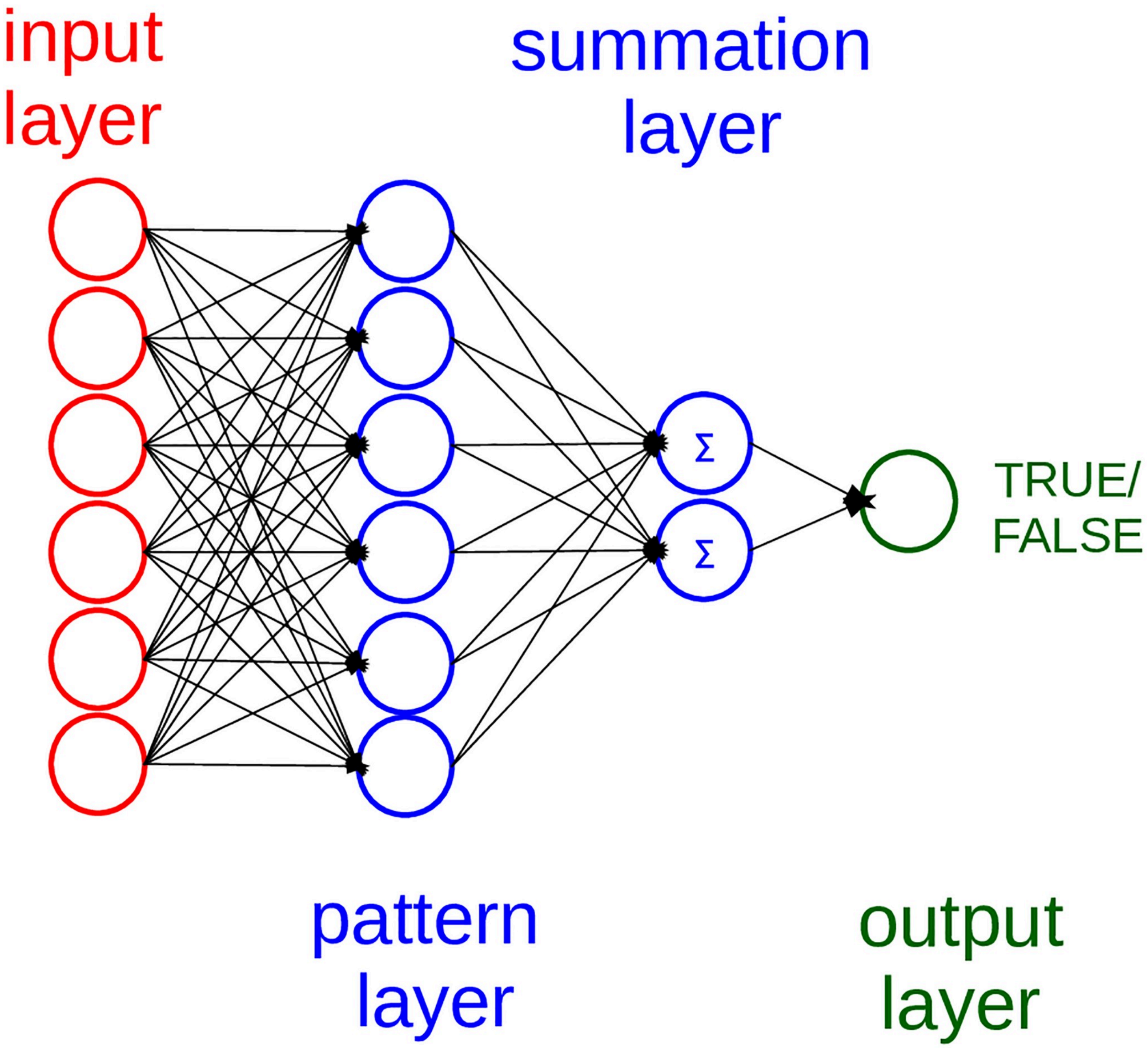

Fig 1. Architecture of the probabilistic neural network. In our model, there are 33 neurons in the input layer, 33 neurons in the pattern layer, and 2 neurons in the summation layer.

https://doi.org/10.1371/journal.pone.0208737.g001

We read the 33 feature values of each patient in the input layer, and we processed their values in the hidden layer. Then, in the output layer, we estimated whether the patient belongs in the mesothelioma or non-mesothelioma diagnosis class. We used 5-fold cross-validation. In each cross-validation fold, we trained on a randomly chosen 80% of the patients, and test on the remaining 20% of the patients. The algorithm finally states if each patient profile is more likely to to belong to the mesothelioma class, or to the non-mesothelioma class.

For our tests, we split the dataset into training set and test set, as made by the dataset curators [10]. We trained our model on the training set, and then applied the trained model to the test set. Best practices in machine learning suggest to split the original dataset into three independent subsets (training set, validation set, and test set) [37], but we decided to use only two-subset split to reproduce the probabilistic neural network used [10]. We then split the dataset

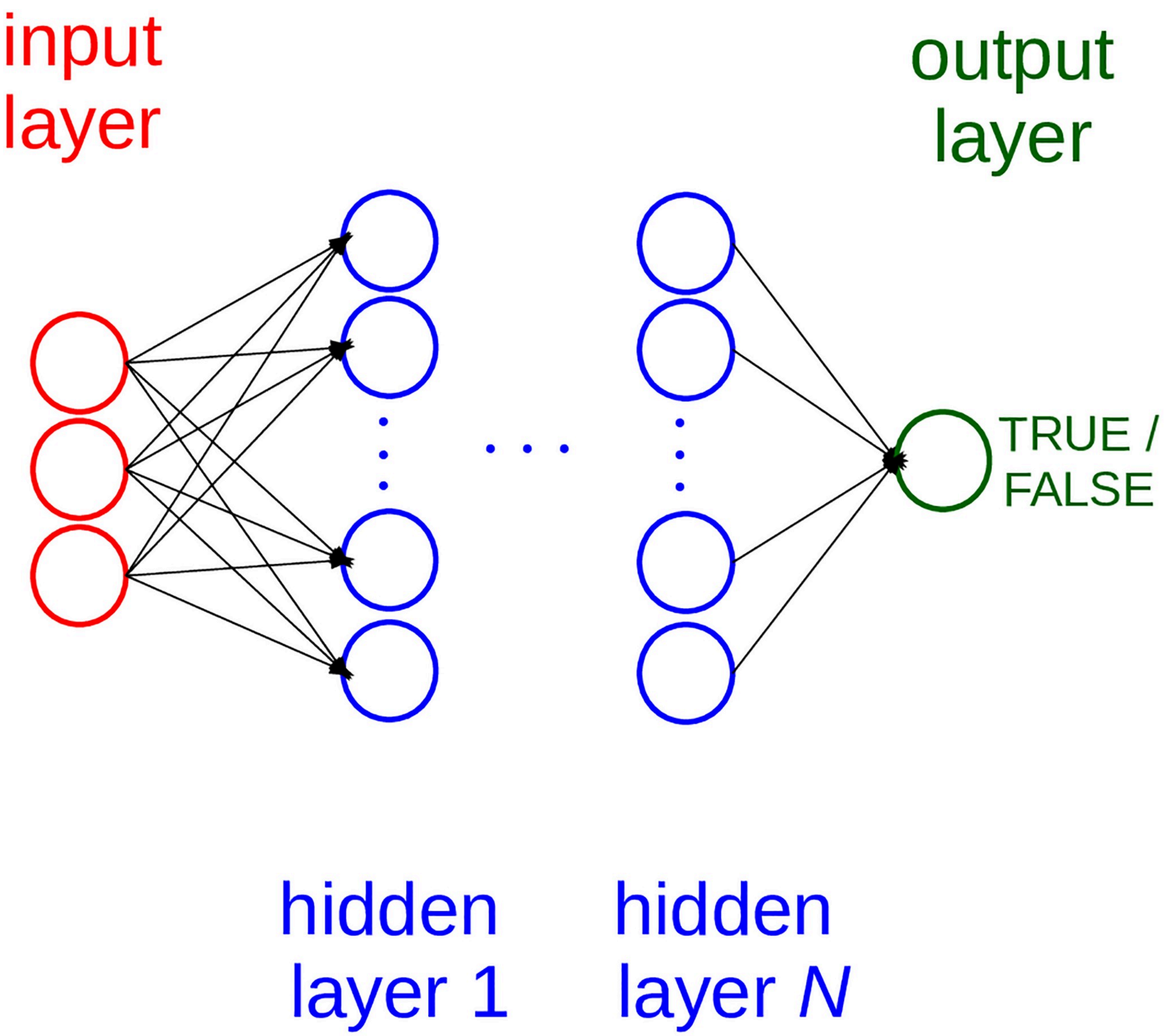

Fig 2. Architecture of a multi-layer perceptron-based neural network. In our model, the input layer neurons are 33. We found different optimized numbers of hidden layers and hidden units, for each program execution. The top architecture among the ten executions had 20 hidden units and 1 hidden layer.

https://doi.org/10.1371/journal.pone.0208737.g002

into training set, validation set, and test set for the perceptron; we then trained each perceptron model on the training set, evaluated it with different hyper-parameters on the validation set, and finally applied the top performing model to the test set.

3.2 Perceptron-based neural network

The main difference between a perceptron and a probabilistic neural network comes from back-propagation. In the perceptron, once the values propagate 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 [46], and they update their weights accordingly.

We read the 33 input values of each patient profile in the input layer, then learned a hidden representation of the profile, and finally translated it into a single real-valued score in the output layer.

We set a confusion matrix threshold τ to 0.5. During testing, we scaled all the values outputted by the neural network through the $z = (x + 1)/2$ formula, where x is the output of the perceptron, and z is the actual value used in the confusion matrix. If the prediction generated a score greater than the likelihood threshold τ, we assigned the patient to the non-mesothelioma class. Otherwise, we assign the patient to the mesothelioma class.

Our multi-layer perceptron used a learning rate of 0.01, and 200 iterations in training. We computed the confusion matrix with the likelihood threshold $\tau = 0.5$. We normalized the input data by column, by scaling every value into the [0; 1] interval, before the application of the perceptron.

We optimized the hyper-parameters (number of hidden layers and number of hidden units) through a grid search, by testing several possible values (hidden layers = [1, 2, 3] and hidden units = [5, 10, 20, 25, 75, 100]). We randomly separated the original dataset into three independent subsets: training set (60% patients, randomly selected), validation set (other 20% patients, randomly selected), and a test set (the remaining 20% patients).

During optimization, for each hyper-parameter configuration, we trained the perceptron on the training set and tested in on the validation set, by computing the Matthews correlation coefficient (MCC) [37, 47]. At the end of the optimization phase, we selected the model which led to the highest MCC score, and applied it to the test set.

Our optimization tests led to different optimized number of hidden units and hidden layers each time, and obtained the top prediction results (MCC = +0.27) with 1 hidden layer and 20 hidden units. In our neural network, we used the sigmoid as activation function.

3.3 Random forest and decision trees

Random forest build upon decision tree learning, in which a set of predictive decision trees maps each input item into an output category, by processing it through its tree leaves [23, 24].

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. 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 child nodes, and so on. The algorithm repeats this procedure until it reaches the end of the tree.

Random forest is an ensemble learning method: it generates multiple classifiers and then aggregates their results. During training, random forest applies a bootstrap aggregating (bagging) method to its trees. It selects random subsets from the input dataset, and applies a decision tree to each of them. To select the final classification outcome, it selects the outcome produced by the majority of the trees, much like a voting system.

The algorithm creates several random decision trees, in which every node corresponds to a feature, randomly selected (Fig 3). The algorithm applies a decision function to each patient profile. For example, for each boolean feature, the node function is "Is the value true?". By applying decision functions repeatedly at each node, the algorithm finally classifies a whole data sample as true or false (mesothelioma or non-mesothelioma in our case). In the end, the random forest outputs the outcome class corresponding to the majority of the outcome classes of the random decision trees (Fig 3), and classifies it as true or false. We trained our random forest classifier on randomly selected 80% data instances and tested on the remaining 20%. In our random forest implementation, we generated 500 trees and tried 11 variables at each node

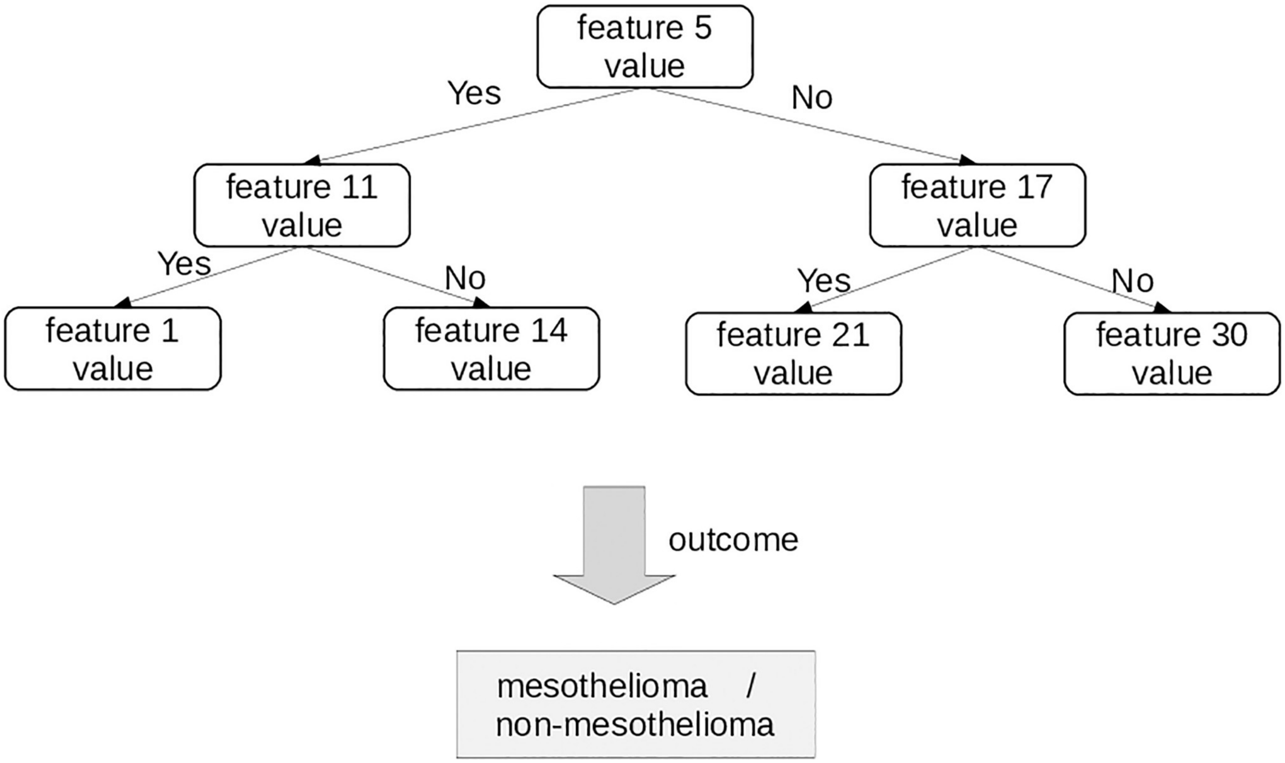

Fig 3. Decision tree. An example of decision tree, which can classify each patient as healthy (non-mesothelioma) or unhealthy (mesothelioma). Random forest generates a set of predictive decision trees.

https://doi.org/10.1371/journal.pone.0208737.g003

split. Differently from the numbers of hidden units and hidden layers of the previously described perceptron-based neural network, we did not run an optimization procedure for the number of generated trees. Increasing the number of trees, in fact, does not improve the performance of random forest, if the the number of trees is sufficiently larger than the number of features [48, 49].

After the diagnosis classification phase, we decided to investigate the most important features of the dataset. We again chose to use random forest for this scope, because this method provides both a statistical outcome (accuracy decrease) and a content-informative outcome of the importance of each feature (Gini node impurity). All the other methods previously used for classification in this project (PNN, perceptron, and one rule) do not produce this twofold outcome.

To rank the importance of each feature, we applied the random forest algorithm to the dataset 33 times. Each time, we removed one feature of the 33 and then computed the accuracy (Eq 1) and the Gini node impurity [39] of the prediction during the decision tree training.

In a confusion matrix, where FP: false positives; FN: false negatives; TP: true positives; TN: true negatives, the accuracy formula is the following:

$$\begin{gathered}\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \\ (\text{accuracy} : \text{worst value} = 0; \text{best value} = 1)\end{gathered} \tag{1}$$

Then, we measured the accuracy and the Gini node impurity decrease between the random forest with all features and the random forest with one feature removed. The top importance features are the ones whose difference in the accuracy and in the Gini node purity is higher, because its absence shows the largest change in the prediction of the diagnosis.

Even if accuracy is less informative than MCC on imbalanced dataset [37], we decided to stick with this score in the feature selection phase to be consistent with the original random forest model introduced by Breiman [23].

We applied the random forest algorithm to rank the features based upon their importance. We measured feature importance with two statistical rates: the proportion of the decrease of the mean square error (MSE, Eq 2) when each feature is missing from the dataset, measured against the 0 or 1 target value (Results), and the tree node Gini impurity (Results).

$$\begin{gathered}\mathrm{MSE}(x,y) = ||x-y||^2 \\ (\text{where } x \text{ is predicted score, and } y \text{ is the corresponding ground truth target})\end{gathered} \tag{2}$$

We computed the percentage of the decrease of the mean square error in the following way. The random forest algorithm computed the accuracy α_{all} of the prediction of the targets by using a decision tree which takes advantage of all the features. Then, the random forest algorithm calculated the accuracy α_{i} of the prediction of the targets by using a decision tree which takes advantage of all the features, except the i_{th} feature.

Afterwards, for each feature i, it computed the percentage mean square error between α_{all} and α_{i}, and assigned it to the i_{th} feature as its percentage of the decrease of the mean square error (Results).

The random forest algorithm computes the impurity of each tree node measured by the Gini index in the following way. For each decision tree split, the method calculated the decrease of Gini index impurity between the node before the split and the node after the split [39].

The larger the impurity decrease after a specific split, the more informative is the feature related to that split [50]. The algorithm summed over all the splits for that feature, over all the trees, and generates its final value (Results).

Feature selection measures the importance of each dataset feature through the accuracy decrease and the Gini impurity decrease. We can consider the accuracy decrease percentage of the mean square error as an importance measure in a statistical sense, and the tree node impurity as an importance measure in an informative content sense.

We applied the random forest feature selection algorithm on all the dataset, and computed the accuracy decrease and the Gini purity decrease for each feature (Results). Here there is no need to split the dataset into training set and test set, because random forest feature selection uses a technique called bagging (or bootstrap aggregation), which generates multiple data subsets by sampling with replacement from the full dataset [42].

We took the accuracy decrease ranking and the Gini impurity ranking, and created a merged ranking by using Borda's method [51]. For each feature f, we sum its position in the first list $p_1(f)$ to its position in the second list $p_2(f)$, and save this value in the ranking score variable $score_f$.

We then sorted all the features from the one having the lowest $score_f$ value to the one having the highest score value (Results).

3.4 One rule

For a baseline comparison, we also implemented and applied the one rule algorithm [26]. Considered one of the simplest machine learning methods existing, one rule is based upon association rules, which involve just one data feature value in each rule condition. In the one rule application, we used randomly selected 80% of the data instances for training, and the remaining 20% for testing (Results).

3.5 Prediction using only two selected features

Since the random forest feature selection highlighted “lung side” and “platelet count (PLT)” as the most relevant features in the dataset (Results), we used a decision tree to predict mesothelioma diagnoses based solely upon these two features (Fig 3). We applied classification and regression tree (CART) [25] to the dataset made only of lung side and platelet count. In the dataset table, we kept only the “lung side” and “platelet count (PLT)” columns, we removed all the other columns (features), and then we applied the CART method. We avoided using random forest in this case because there are only two features: as we described earlier, random forest creates decisions trees based on random combinations of feature subsets, and there would be no possible subset combinations on a dataset containing only two features.

We decided to employ decision tree in this phase because lung side and platelet count were identified as the two most important features by random forest, and random forest is based upon decision trees [23]. If we used another method such as perceptron-based neural network at this stage, it could potentially disagree with random forest on which features are the most important, and therefore generate inconsistent diagnosis prediction results.

Moreover, a methodological advantage of decision tree is that its operating principles and results are easy to understand and interpret [52, 53]. In a scenario where a biomedical doctor has to figure out if a patient had mesothelioma by just looking at the values for lung side and platelet count in the medical record, he/she could diagram all the decision tree steps and understand the reasons beyond the outcome generated. This information would be pivotal for the doctor’s decision making, and would help him/her better interpret the patient’s situation [54, 55]. On the contrary, understanding the operating principles beyond more complex machine learning methods (such as the neural networks used in this study) would be difficult, or even impossible, in a health decision making context [56]. Therefore, to make a critical decision about the therapy for a patient, a biomedical doctor would trust an explainable decision tree more than a black box neural network.

To verify that the predictive power of the lung side and platelet count was valid not only for decision trees, we also applied one rule to this dataset made by only the two selected features (S4 Table).

We used the 80% randomly selected patient profiles for training and the remaining 20% profiles for testing (Results).

3.6 Prediction measurement and dataset split

To state the effectiveness of our prediction methods, we used Matthews correlation coefficient (MCC, Eq 3), accuracy (Eq 1), F_1 score (Eq 4), sensitivity (true positive rate, Eq 5), and specificity (true negative rate, Eq 6) rates.

$$\begin{gathered}\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\cdot(\mathrm{TP}+\mathrm{FN})\cdot(\mathrm{TN}+\mathrm{FP})\cdot(\mathrm{TN}+\mathrm{FN})}}\\ (\mathrm{MCC}:\text{worst value} = -1;\text{best value} = +1)\end{gathered} \tag{3}$$

$$\begin{gathered}\mathrm{F}_1\ \text{score} = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}\\ (\mathrm{F}_1\ \text{score}:\text{worst value} = 0;\text{best value} = 1)\end{gathered} \tag{4}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ (\text{sensitivity} : \text{worst value} = 0; \text{best value} = 1) \tag{5}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ (\text{specificity} : \text{worst value} = 0; \text{best value} = 1) \tag{6}$$

We optimize and evaluate our methods by using the MCC because it weights each class of the confusion matrix, in proportion to the number of positive elements and negative elements in both the gold standard and the prediction [47]. Sensitivity (Eq 5) generates the rate of the correctly predicted true positives on the total positive data instances, while specificity (Eq 6) produces the rate of the correctly classified true negatives on the total tally of false data instances. Accuracy (Eq 1) measures the proportion of the correct predictions (true positives plus true negatives) on the total data instances, while the F_1 score (Eq 4) reports the is the harmonic mean of precision and sensitivity.

As described earlier, we used different strategies for the dataset split, in accordance to the need of optimizing hyper-parameters or not, for each method. Since our probabilistic neural networks, one rule, random forest, and decision trees have no hyper-parameter to optimize, we split the dataset into training set (80% of the data instances, randomly selected) and test test (the remaining 20% data instances) for all the analyses [37].

For the perceptron-based neural network, instead, we ran an optimization procedure to find the best number of hidden units and number of hidden layers. To do so, we split the dataset into training set (60% data instances, randomly selected), validation set (20% of the remaining data instances, randomly selected), and test set (the remaining 20% data instances). We trained each architecture model on the training set, and evaluated it on the validation set. At the end of the optimization phase, we selected the model which obtained the highest prediction score (MCC) on the validation set, and finally applied to the test set.

We later report the results related to the test sets (Results). Each test set used by each method in this project was completely independent from training set and validation set, and has no element in common with them. Each test set employed for each method execution contains 20% of the dataset: 65 randomly selected patients for the complete imbalance dataset tests, and 39 randomly selected patients for the under-sampled balanced dataset tests.

To make an even more precise comparison of the classifiers we employed, we recognize that it would have been ideal to initially put aside a *held-out* data subset as an additional final test set [57], then apply all the optimized trained methods on this *held-out* subset, and finally compare the results obtained by each method (similarly to what happens in the DREAM Challenges [58, 59]). Unfortunately, because of the small size of the dataset analyzed in this study (324 patients), we could not take advantage of this strategy, otherwise we would have not had enough data instances to properly train the models. Our results attained by randomly selecting and shuffling the data instances for each test set, however, confirmed the generalisability of our methods.

3.7 Regression analysis

In addition to the neural network, random forest, one rule, and decision tree prediction and random forest feature selection approaches, we also applied a traditional regression analysis to the dataset. We compared of clinical, radiographic, demographic, and laboratory characteristics between the mesothelioma patients and those who do not have cancer (non-mesothelioma) but who are asbestos exposed by Wilcoxon rank sum tests for continuous variables [60], and Fisher exact tests for categorical variables [61]. We applied univariate logistic regression models to assess the effect of each individual factor or characteristics on diagnosis, reporting estimated odds ratios (OR) and 95% confidence intervals (CI) after applying two-sided statistical tests. A multivariate logistic regression model included all variables with a threshold alpha set at 0.10 or lower, followed by backwards selection of variables with a threshold set at alpha of 0.05 or lower.

3.8 Execution details

On a Dell Latitude 3540 computer with a Intel Core i3-4030U CPU 1.90 GHz processor, with 3.7 GB of random-access memory (RAM), and running a Linux CentOS 7 operating system, the execution of the probabilistic neural network on Python 3.5 with NeuPy [62] execution takes around 1 second, while the execution of the perceptron-based neural network on Torch 7 with the nn and optim packages [63] execution takes around 2 minutes and 30 seconds. The perceptron prediction takes longer because of the optimization phase, which lacks for the other algorithms. We applied random forest through the R randomForest package, and its execution lasted around 1 second, both for classification and feature selection. We applied one rule through the R OneR package [64] and the CART decision tree through the R rpart package [65], and their execution lasted around 1 second, too.

4 Results

4.1 Predictions of patients diagnosis on the complete imbalanced dataset

Er et al. [10] reported top prediction accuracy of 0.98, but, upon investigation, we noticed that one of their input feature ("diagnosis method") duplicated the target diagnosis class. This input feature makes it trivially easy to obtain perfect and almost perfect prediction accuracy, but it is unlikely to exist in a real-world setting. Therefore, we excluded this feature from our analysis.

We generated prediction results through probabilistic neural network, perceptron-based neural network, one rule, decision tree, and random forest classifier applied to the all the features and to all the data instances, and through decision tree applied to the two top selected features and to all the data instances (Table 3).

Our probabilistic neural network achieved the lowest prediction score among the methods we tried, obtaining a result similar to a random prediction (MCC = +0.03, Table 3). This

Table 3. Results of the computational predictions of patient diagnosis on the complete dataset. Matthews correlation coefficient (MCC): Eq 3. Accuracy: Eq 1. F_1 score: Eq 4. Sensitivity (true positive rate): Eq 5. Specificity (true negative rate): Eq 6. The scores are the medians of the results' ten separate program executions. We report the results of the application of the methods on all the dataset features, plus the results of the decision tree only to the two selected features: the row entitled "Decision tree (applied only to lung side & platelet count)". Dataset imbalance: 29.63% positive data instances (all the 96 mesothelioma patients), and 70.37% negative data instances (all the 228 non-mesothelioma patients).

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

https://doi.org/10.1371/journal.pone.0208737.t003 method showed flaws in predicting true positives (sensitivity = 0.32) but did sufficiently well on predicting true negatives (specificity = 0.71).

The other artificial neural network we used, the multi-layer perceptron, and decision tree attained a low general scores: MCC = +0.11 and MCC = +0.19, respectively (Table 3). This deep learning model obtained a low prediction score on the true negatives (specificity = 0.42) but very good prediction score on the true positives (sensitivity = 0.66, (Table 3). The CART decision tree achieved specificity (0.80) but low sensitivity (0.39).

Regarding tree like graph models, one rule (MCC = +0.27) achieved very low results on the sensitivity (0.17) but almost perfect predictions on the specificity (0.97). Random forest outperformed all the other methods, achieving MCC = +0.37, with a low true positive rate (sensitivity = 0.28) and an almost perfect true negative rate (specificity = 0.97).

We also took advantage of the feature selection discoveries, and applied a CART decision tree [25] only to the "lung side" and "platelet count (PLT)" patients values. The prediction results showed an MCC of +0.28, higher than the results obtained with one rule (MCC = +0.27), of multi-layer perceptron (MCC = +0.11), of the probabilistic neural network (MCC = +0.03), and even of decision tree itself (MCC = +0.19) applied to the complete dataset (Table 3). These results confirmed that "lung side" and "platelet count (PLT)" are the most relevant features of the dataset in our analysis, and are alone sufficient to run a reliable computational prediction of the patients' true negative outcomes. To prove these results on the selected two features are unbiased towards the CART decision tree, we applied one rule to the same dataset and obtained similar results, even if slightly lower (MCC = +0.27, S4 Table).

Generally, random forest outperformed all the other methods on the MCC and true negative rate (specificity), but was outperformed by perceptron and probabilistic neural network on the true positive rate (sensitivity). The decision tree applied to the two features obtained the top accuracy, while the multi-layer perceptron was the only algorithm which achieved high prediction results for true positive patients (sensitivity = 0.66), while all the other methods obtain sensitivity scores lower than 0.5, so cannot be considered reliable in detecting true positive patients (Table 3). The multi-layer perceptron attained a true negative rate (specificity = 0.42) lower than all the other methods (Table 3).

Sensitivity and specificity results show that all the methods except the multi-layer perceptron had better capability in predicting true negatives than true positives (Table 3). We believe these results are caused by an imbalanced ratio (29.63% positive data instances, and 70.37% negative instances) of the dataset. Since the models see more negative elements during training, they are better at predicting negative data instances during testing. We therefore tacked the dataset imbalance problem with the under-sampling technique [66], and we show the results in the next section.

However, this inability to predict true positives does not regard the multi-layer neural network, which achieved a high true positive rate (sensitivity = 0.66) without any data imbalance handling strategy.

Even if correctly classifying patients with mesothelioma (sensitivity) and patients without mesothelioma (specificity) are both relevant tasks, we give more importance to the former, because it can identify the patients that need to be cured through a therapy, and possibly have their life saved by an early detection of mesothelioma. To this end, it is relevant to notice that the decision tree applied only to the lung side and platelet count features gained a higher sensitivity (0.28) than the random forest classifier and one rule (Table 3). Regarding true negatives, it is worth mentioning that random forest classifier and one rule obtained an almost perfect specificity score (0.97) that outperformed all the other models (Table 3).

It is relevant to notice that each of the five confusion matrix scores we listed (MCC, accuracy, F_1 score, sensitivity, specificity, Table 3) generate different rankings of our methods, confirming the importance of comparing different rates and not focusing on a single one. As mentioned earlier, we optimized our methods based upon the Matthews correlation coefficient, because it is the only rate that considers all the four categories of the confusion matrix and the balance of the dataset.

For a complete comparison to the reported results of Er et al. [10], we also computed the predictions on the original dataset including the problematic "diagnosis method" feature (S3 Table). As expected, random forest achieved perfect MCC of +1.00, but such classifier would have limited utility in clinical settings.

4.2 Predictions of patients diagnosis on the under-sampled balanced dataset

As already mentioned, our methods applied to the complete dataset obtained generally good results on the true negative rate, and low results on the true positive rate (Table 3). The dataset imbalance is the cause of this inability to predict true positives. The dataset, in fact, contains 228 negative data instances, and just 96 positive data instances. During training, then, each model learns well how to recognize negative elements, but does not learn well how to identify positive elements.

There are many techniques to tackle this dataset imbalance problem: data class weighting [67], over-sampling [68], and under-sampling [66], for example. Here we decided to use under-sampling because this approach does not involve any manipulation or weight assignment to the data instances, making its application more realistically usable in clinical environments than other techniques.

We implemented under-sampling in the following way. The minority class in our dataset contains 96 elements (positive data instances), while majority class contains 228 elements (negative data instances). We created a balanced subset containing all the 96 positive data instances, and 96 negative data instances randomly selected from the majority class. The balanced subset created now contained 192 data instances, with 50% perfect balance. We then applied all the methods to this balanced subset (with the same dataset split and execution configurations of the previous tests) and recorded their results (Table 4).

Compared to the results obtained on all the dataset (Table 3), here all the methods achieved lower specificity and higher sensitivity (Table 4) correctly reflecting the change of ratio positive and negative data instances. After under-sampling, in fact, the percentage of negative data instances moved from 70.37% to 50%, while the percentage of positive data instances increased from 29.64% to 50%. These changes made all the methods able to learn a larger ratio of positive

Table 4. Results of the computational predictions of patient diagnosis, after under-sampling. Matthews correlation coefficient (MCC): Eq 3. Accuracy: Eq 1. F_1 score: Eq 4. Sensitivity (true positive rate): Eq 5. Specificity (true negative rate): Eq 6. The scores are the medians of the results' ten separate program executions, run with different subset content selected randomly for training set, validation set, and test set every time. We report the results of the application of the methods on all the dataset features, plus the results of the decision tree only to the two selected features: the row entitled "Decision tree (applied only to lung side & platelet count)". Dataset balance: 50% positive data instances (all the 96 mesothelioma patients), and 50% negative data instances (96 non-mesothelioma patients, randomly selected). Perceptron: learning rate = 0.1.

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

https://doi.org/10.1371/journal.pone.0208737.t004

elements, and a smaller ratio of negative elements during training, and their consequences clearly influenced the results (Table 4).

Random forest outperformed again all the other methods (MCC = +0.64), by obtaining a high true positive rate (sensitivity = 0.75) and a very high true negative rate (specificity = 0.86). Among all the methods tried, random forest attained the best MCC, accuracy, F_1 score, and true negative rate. Random forest, however, did not attain the top true positive rate, which was achieved again by the multi-layer perceptron neural network (sensitivity = 0.95). Perceptron-based neural network obtained the highest true positive rate both on the complete imbalanced dataset (Table 3) and on the under-sampled balanced dataset (Table 4).

Conversely from the results obtained on all the data instances (Table 3), decision tree applied on all the features of the under-sampled dataset achieved the second top performance among all the methods (MCC = +0.59) and outperformed decision tree itself applied just to the lung side and platelet count (MCC = +0.41). These results show that decision tree applied only to the two selected features works well if there are enough data instances to train and test the model; otherwise, more features lead to better prediction scores. On the complete imbalanced dataset, in fact, there are 324 patients for which the lung side and platelet count features are available. Here, instead, decision tree applied to the two-feature dataset made of just 192 patients did not have enough data instances to outperform decision tree applied on all the features.

On the complete imbalanced dataset, less features, more data instances, and data imbalance led to better predictions for decision tree. On the under-sampled balanced dataset, more features, less data instances, and data balance led better predictions for decision tree.

Perceptron-based neural network obtained an almost perfect score for sensitivity (0.95), confirming again its predictive power in classifying true positive patients. This neural network, however, obtained the worst results on specificity (0.20) among all the methods tried. Compared to the complete imbalanced dataset tests, one rule dropped its general performances score from MCC = +0.27 to MCC = +0.15. Probabilistic neural network obtained again the worst general prediction scores (MCC, accuracy, and F_1 score) among all the models.

4.3 Feature selection

On the feature selection content, the features “lung side” and “platelet count (PLT)” resulted as the most predictive ones among the 33 dataset features (Figs 4 and 5). We measured the importance of each feature with the mean square error decrease (Fig 4) and the Gini node impurity decrease (Fig 5), and these measures highlight “lung side” and “platelet count (PLT)” as the most relevant features for the dataset. In other words, the removal of these two features from the dataset would influence the prediction of the diagnosis more than the removal of the other ones. We selected only “lung side” and “platelet count (PLT)” as top features because they both occupy the first and second positions in both the random forest rankings (Figs 4 and 5).

The merged ranking confirmed the importance of “lung side” and “platelet count (PLT)”, followed by four non-clinical features: “duration of symptoms”, “age”, “city”, “duration of asbestos exposure” (Table 5).

The ranking indicated that the less influential features of the predictions are “dead or not”, “weakness”, “pleural effusion” and “ache on chest” (Table 5). Some of these features even have a negative effect on the prediction. “pleural fluid WBC count”, “total protein”, “alkaline phosphatise (ALP)”, “dyspnoea”, “pleural level of acidity (pH)”, “ache on chest”, “pleural effusion”, “weakness”, “dead or not” have negative values in the tree node impurity value list (Fig 4 and MSE accuracy column of Table 5). These features appear not to add useful information, and might even cause overfitting.

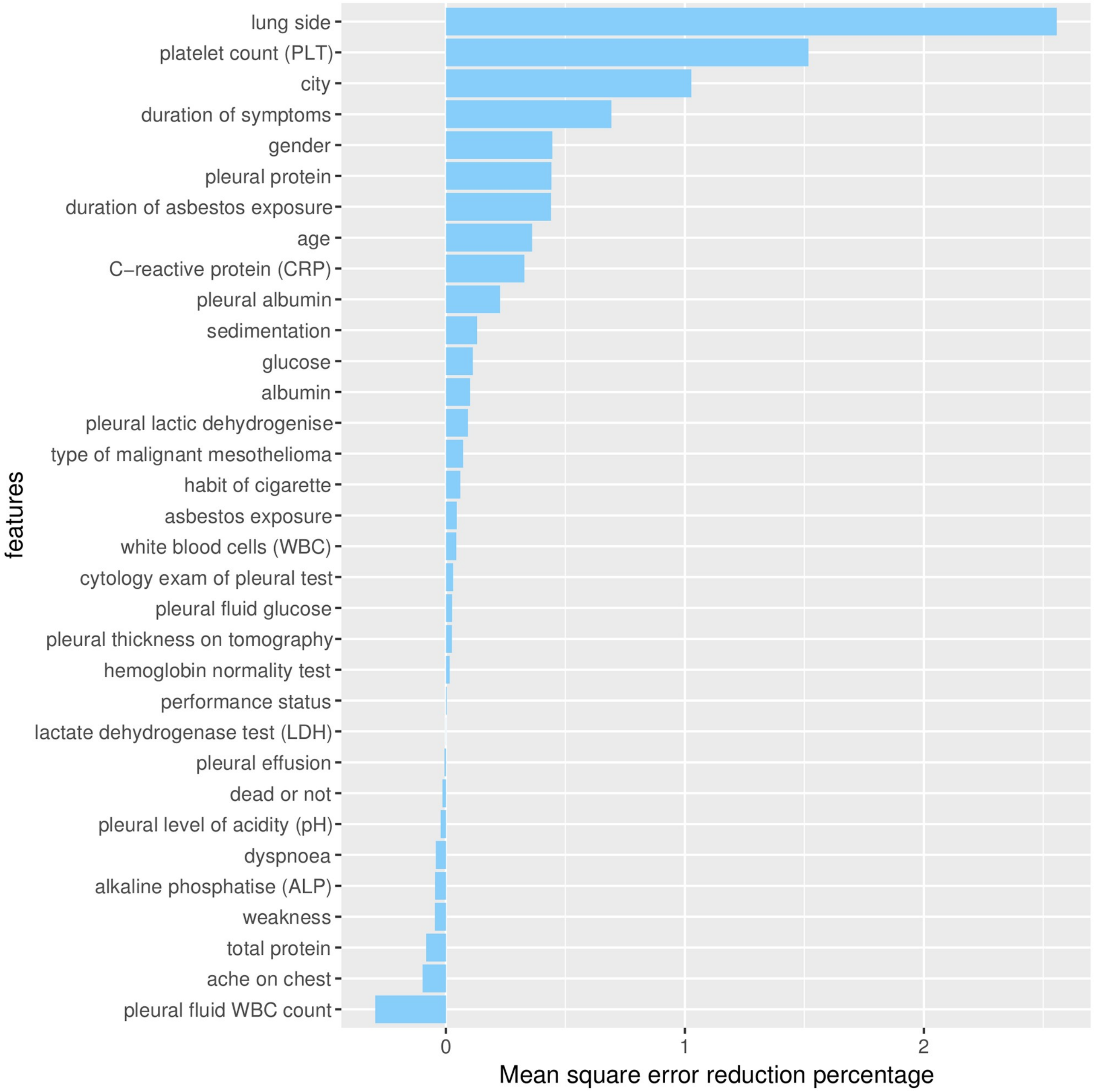

Fig 4. Mean square error (MSE) decrease in accuracy for each feature removal. Random forest feature selection rely on bootstrap aggregation (bagging), and therefore does not have training set, validation set, and test set [69]. The bars represent the drop in the accuracy of the prediction made on the patients' dataset each time a feature is removed. For each feature, the higher is its accuracy drop when removed, the more important the feature is (Methods).

https://doi.org/10.1371/journal.pone.0208737.g004

The random forest percentage decrease in the Gini node impurity error does not fully confirm this negative effect of the aforementioned features, by for example selecting "pleural fluid WBC count" as the thirteenth most important feature (Fig 5). The difference on the feature selection of these two indexes is caused by their different meaning. The mean square error decrease, in fact, is based upon prediction statistics, while the Gini impurity node decrease is based upon the dataset content information. This meaning difference might lead to such

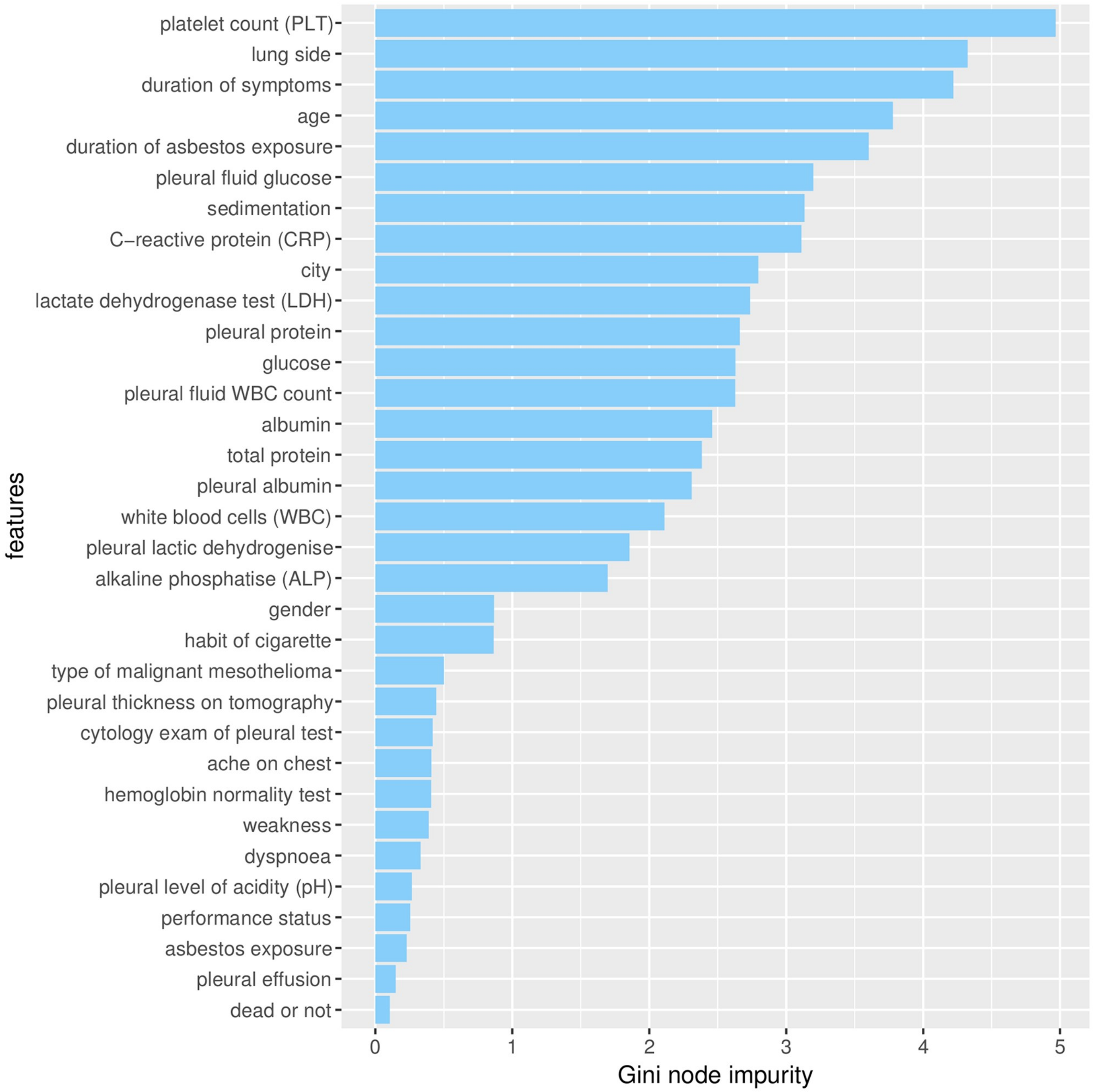

Fig 5. Gini impurity decreases of each random forest tree node. Random forest feature selection rely on bootstrap aggregation (bagging), and therefore does not have training set, validation set, and test set [69]. The bars represent the importance of each feature, measured through the sum of all the Gini impurity index decreases for each specific feature [39] (Methods).

https://doi.org/10.1371/journal.pone.0208737.g005

ambiguous cases. Since we want to find the most relevant features of the dataset, and not the least important, we focus on the top features found by both this measures ("lung side" and "platelet count (PLT)") (Table 5).

4.4 Biostatistics analysis

As regression methods are more commonly used to identify variables associated with an outcome (which in this case was presence of mesothelioma among asbestos-exposed individuals),

Table 5. Merged rank of features. We sorted the features by combining ranking of the node impurity and the ranking of the percentage of MSE decrease in accuracy (Methods).

merged ranking position	feature name	MSE decrease in accuracy %	tree node purity decrease
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 fluid	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 normality test	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

https://doi.org/10.1371/journal.pone.0208737.t005

we also performed this traditional statistical modeling technique to allow comparison with our machine learning approaches.

Before regression takes place, it is usual to explore the nature of the relationships between clinical, demographic, radiographic, and laboratory characteristics and the outcome of interest. For this dataset at a significance level of 0.05, patients with mesothelioma were slightly younger (Wilcoxon rank-sum test, $p = 0.03$), more likely to be male (Fisher exact test, $p = 0.01$), were more likely to have mesothelioma in its initial phase (T1 phase in the TNM Classification of Malignant Tumors [70]) (Fisher exact test, $p = 0.01$), and pleural plaques on both lung side (Fisher exact test, $p = 0.001$)(Supplementary section 5). Univariate logistic regression methods (Supplementary section 5) identified age, gender, and lung side as statistically different between cases and controls, when alpha was set at 0.05. CRP levels, duration of symptoms, and duration of asbestos exposure resulted in non-significant trends at an alpha between 0.05 and 0.10. In subsequent multivariate regression analyses, only lung side remained significant.

5 Discussion

Our results highlighted several interesting aspects, both regarding the diagnosis prediction and the feature selection. Random forest classifier predicted mesothelioma patients' diagnosis with high accuracy, both on the complete imbalanced dataset and on the under-sampled balanced dataset. The random forest classifier, in fact, outperformed the probabilistic neural network model previously used to predict the diagnosis of the patients, and all the methods employed in this study. The multi-layer perceptron and one rule outperformed the probabilistic neural network too, but were outperformed by the random forest classifier (Results). These results suggest further usage of random forest and ensemble learning in health informatics.

Our perceptron-based neural network can precisely identify true positive patients having mesothelioma, while our random forest classifier and one rule models can detect true negative patients without mesothelioma with almost perfect specificity. Random forest, in fact, obtained the top prediction results measured with the Matthews correlation coefficient and specificity but, regarding sensitivity, the perceptron resulted in the top performing method with the only sensitivity rate able to predict the majority of true negative elements (both on the complete imbalanced dataset and on the under-sampled balanced dataset). In this scenario, we would suggest biomedical doctors to take advantage of our multi-layer perceptron to predict true positive patients, and to employ our random forest and one rule methods to identify true negative patients.

The presence of pleural plaques on both the lung sides is highly predictive for malignancy. According to our feature selection analysis, "lung side" feature is the most important sign of mesothelioma. If a patient is found to have pleural plaques on both sides of the lung, that patient has a high probability of having a mesothelioma. In fact, the presence of pleural plaques in both sides of the lung as proof of mesothelioma is well known fact in the biomedical community [71]. Doctors consider cancer appearing on both lung sides as a sign of progress in mesothelioma staging, precisely in the advance from stage IIIA to stage IIIB [72]. Also, the association of the "lung side" feature value with the mesothelioma patients' status confirms the importance of this feature. In this dataset, 22 patients have pleural plaques on both sides (”lung side” value: 2). Among these patients, 16 have mesothelioma in the dataset, meaning the 72.72%. We therefore can see this inference as a correct positive control test for our method.

Low platelet count is strongly related to mesothelioma. According to what our feature selection found, the "platelet count" feature is another influential sign of mesothelioma. Following this indication, we studied the values of this feature and observed that, if patients have a low level of platelet count, they have a high probability of having a mesothelioma.

The "platelet count" feature turned out to be the second most relevant predictive feature of the dataset, ranking second in the list of the features sorted by the mean square error accuracy decrease (Fig 4), first in the list of the features sorted by node impurity (Fig 5) generated by the random forest algorithm, and second in the merged list (Results).

The association of the "platelet count" feature value with the mesothelioma patients' status confirms the importance of this feature. As mentioned before, the normal range of platelet count for a patient is between 150k and 400k platelets per microliter. The patients having platelet count lower than 150k per microliter in the original dataset are 42. Among these, the mesothelioma ones are 23, that is 54.76% of the total.

We then can state that if a patient has a platelet count value smaller than the lower normality limit (150k platelet per microliter), he/she probably experiences mesothelioma.

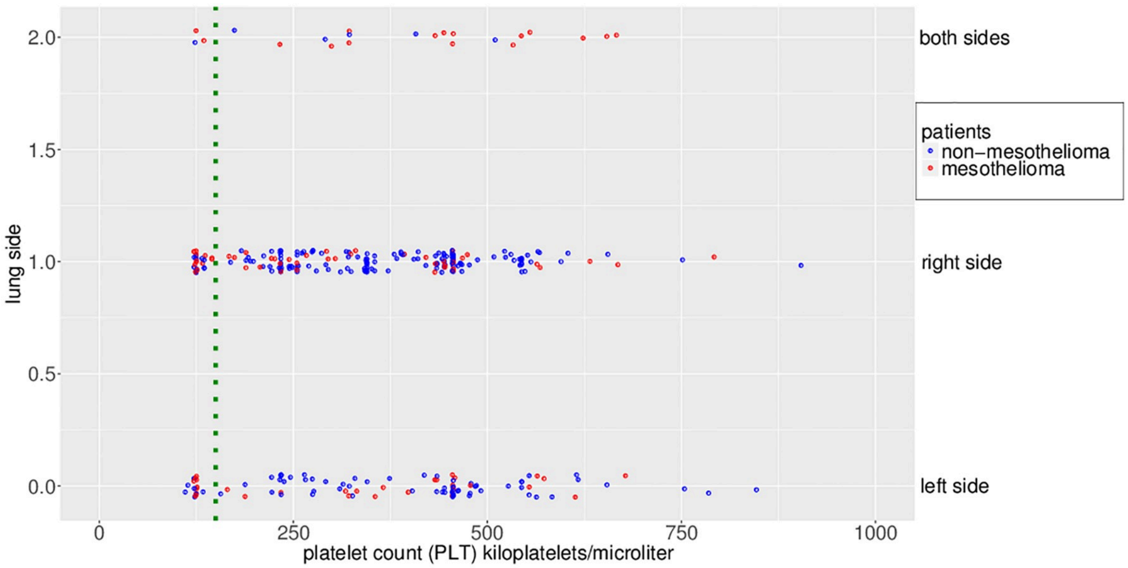

Fig 6. Strip plot of platelet count (PLT) by lung side. We exclude one outlier on the X axis with 3,335 platelet/microliter. Vertical blue dotted line: lower boundary of the platelet count normality test.

https://doi.org/10.1371/journal.pone.0208737.g006

In Fig 6, in fact, the majority of dots on the "both sides" horizontal line in the plot are red, meaning that most of these patients have mesothelioma. And also the majority of dots on the left of the green dotted vertical separator (set at the lowest normality limit, 150k platelet per microliter) are red, confirming that the most of patients having platelet count lower than the normality range limit have mesothelioma.

Several research studies (for example, [73]) confirm that low platelet count strongly relates to mesothelioma and it can also happen as a consequence of chemotherapy [74].

The duration of asbestos exposure is an important risk factor, but not among the most important features, according to our random forest feature selection. As we mentioned previously (Introduction), physicians commonly consider the duration of asbestos exposure and the occupational history of the patient as the most relevant risk factors for mesothelioma diagnosis. No information about the occupational history of the dataset patients is available. Regarding "duration of asbestos exposure", our feature selection model ranked this feature as the sixth most important feature among 33 (Results).

A decision tree applied the two main features selected by random forest ("lung side" and "platelet count") alone predicted diagnosis of mesothelioma patients with higher accuracy than all the other methods (including decision tree itself) applied to the complete imbalanced dataset.

After having identified the two main features, we trained and tested a classification and regression tree on dataset made only by those two features and by all the 324 patients. Results showed high MCC prediction scores, confirming the importance of "lung side" and "platelet count" in the dataset. These results suggest that physicians could focus on these two features, when analyzing the health record of a patient with signs of mesothelioma, if other feature values were unavailable in his/her medical charts.

Our results about the "lung side" and "platelet count" features can be useful for medical doctors and physicians dealing with patients having mesothelioma symptoms. Our results state that, when analyzing health records of patients having mesothelioma symptoms, physicians should pay more attention to these two highly informative features than to the other features available. Therefore, extra analysis and tests on lung sides and platelet count can be pivotal to diagnose mesothelioma. Additionally, in case only data related to "lung side" and "platelet count" were available for some patients, doctors and biomedical researchers can take advantage of our trained machine learning system to predict their diagnosis. To the best of our knowledge, physicians have not used lung side and platelet count of patients' health records for mesothelioma diagnosis. The results achieved after applying under-sampling, however, showed that decision tree applied to all the features obtained better prediction scores than decision tree applied only to platelet count and lung side, on the under-sampled balanced dataset. This outcome shows that decision tree applied to the two selected features needs more data instances to outperform decision tree applied to all the features. Decision tree applied to all the features, instead, beats decision tree applied to the two top features, on a perfectly balanced dataset containing the same number of positive data instances and negative data instances.

Additionally, we showed that random forest feature selection provides more insight than standard biostatistics analysis. Random forest, in fact, identified a substantially larger set of important factors that affected mesothelioma risk, when compared to traditional regression methods. Both identified the overwhelming covariate of lung side, but regression methods did not identify platelet count, city, or pleural protein that random forest highlighted.

Regarding the limitations of this study, we have to report that our approach might not generalize well in the mesothelioma context, because of the specificity of the features (for example, the "city" feature, which is the distance from downtown). Our approach, however, can be applied to any patients dataset of any disease available, generate reliable models for diagnosis prediction, and identify the most relevant clinical feature in any of these cases. About feature selection, we have to reaffirm that we built this phase only on random forest, and therefore its results might be biased towards this algorithm (Methods). This limitation might be addressed in the future by employing multiple feature selection methods, and then by comparing and aggregating their results afterwards through advanced correlation rates (such as Spearman's ρ and Kendall τ rank correlation coefficients [75], for example).

Under-sampling confirmed its utility to improve the classification results on the minor class (true positives, in our case), even if it brought the limitation of discarding some useful data instances. The under-sampling prediction results, in fact, relate only to 192 patients, and not to the complete dataset made of 324 patients.

Feature work will also include the enhancement of the presented machinery by applying alternative techniques to handle the data class-imbalance [37, 67, 68], the application of our algorithm combination to other disease health record datasets (for example, [41]), the application of alternative machine learning algorithms (for example, latent Dirichlet allocation [76] or probabilistic latent semantic analysis [77]) for the diagnosis prediction, and the possible usage of semantic similarity measures to incorporate similarity information between features (for example, through latent semantic indexing [78]). We also plan to explore the feature dependence in the dataset, to see what feature influence which other features and how.

Supporting information

S1 File. Feature meanings. In this file, we report an accurate description of the meanings of the features.
(PDF)

S1 Fig. Barplots of the values of the boolean features. On the left, the patients who do not have mesothelioma; on the right, the mesothelioma patients.
(TIFF)

S2 Fig. Barplots of the values of the category features. On the left, the patients who do not have mesothelioma; on the right, the mesothelioma patients.
(TIFF)

S3 Fig. Histograms of the values of the time features. On the right, the patients having mesothelioma; on the left, the non-mesothelioma patients.
(TIFF)

S4 Fig. Histograms of the real-valued features (part 1/2). On the left, the patients who do not have mesothelioma; on the right, the mesothelioma patients.
(TIFF)

S5 Fig. Histograms of the real-valued features (part 2/2). On the left, the patients who do not have mesothelioma; on the right, the mesothelioma patients.
(TIFF)

S1 Table. Statistical analysis table summary. Descriptive statistics with median and range for continuous factors and frequencies and percentages for categorical factors.
(TEX)

S2 Table. Univariate logistic regression table. We provide descriptive statistics with median and range for continuous factors and frequencies and percentages for categorical factors. NA (not available): not estimable value. reference: reference value used for computing the odd ratios (OR). p-value: statistical p-value for each feature value, computed in relation to the reference value. global p-value: statistical p-value that tests the hypothesis of all the feature values together.
(TEX)

S3 Table. Results of the computational predictions of patient diagnosis, including the "diagnosis method" feature in the dataset.
(TEX)

S4 Table. Results of the computational predictions of patient diagnosis, regarding one rule applied to the two top selected features. The scores are the medians of the the results of ten separate program executions. We applied one rule only to the selected features "lung side" and "platelet count (PLT)". Dataset imbalance: 29.63% positive data instances (mesothelioma), and 70.37% negative data instances (non-mesothelioma).
(TEX)

Acknowledgments

The authors thank Khalid Amir Mahmoud Abdulla (the University of Melbourne) for his help in multiple programming tasks, Wail Ba-Alawi (Princess Margaret Cancer Centre) for his technical help, Christina Cumbaa (Krembil Research Institute) for his help in the review guidance, Crystal Kantores (Peter Munk Cardiac Centre) and Laura Aronoff (University of Toronto) for their help in the English proof-reading of the manuscript, and the University of California Irvine Machine Learning Repository for providing the mesothelioma dataset to the public. The authors especially thank Slava Epelman and the Peter Munk Cardiac Centre for having supported this publication. This paper is dedicated to all the patients having mesothelioma.

Author Contributions

Conceptualization: Davide Chicco, Cristina Rovelli.

Formal analysis: Davide Chicco.

Investigation: Davide Chicco, Cristina Rovelli.

Methodology: Davide Chicco, Cristina Rovelli.

Project administration: Davide Chicco, Cristina Rovelli.

Software: Davide Chicco.

Supervision: Davide Chicco, Cristina Rovelli.

Validation: Davide Chicco.

Visualization: Davide Chicco.

Writing – original draft: Davide Chicco, Cristina Rovelli.

Writing – review & editing: Davide Chicco.

References

1. McDonald JC, McDonald AD. The epidemiology of mesothelioma in historical context. European Respiratory Journal. 1996; 9(9):1932–1942. https://doi.org/10.1183/09031936.96.09091932 PMID: 8880114
2. Dollinger M, Tempero M, Mulvihill S. Everyone's guide to cancer therapy: how cancer is diagnosed, treated, and managed day to day. Andrews McMeel Publishing; 2002.
3. Delgermaa V, Takahashi K, Park EK, Le GV, Hara T, Sorahan T. Global mesothelioma deaths reported to the World Health Organization between 1994 and 2008. Bulletin of the World Health Organization. 2011; 89(10):716–724. https://doi.org/10.2471/BLT.11.086678 PMID: 22084509
4. Cancer Research UK. Types of mesothelioma; 2016. http://www.cancerhelp.org.uk/help/default.asp?page=4398. URL visited on 31st October 2016.
5. Yadav A, Kumar S, Khulbe S, Sharma S, Ahlwat V, Yadav RK, et al. Rare case of benign pleural fibrous mesothelioma: a surgical experience. Indian Journal of Thoracic and Cardiovascular Surgery. 2004; 20(3):142–143. https://doi.org/10.1007/s12055-004-0068-x
6. Whitwell F, Rawcliffe RM. Diffuse malignant pleural mesothelioma and asbestos exposure. Thorax. 1971; 26(1):6–22. https://doi.org/10.1136/thx.26.1.6 PMID: 5101273
7. Coates A, Forbes J, Simes RJ. Prognostic value of performance status and quality-of-life scores during chemotherapy for advanced breast cancer. The Australian New Zealand Breast Cancer Trials Group. Journal of Clinical Oncology. 1993; 11(10):2050–2050. PMID: 8410129
8. Whitwell F, Scott J, Grimshaw M. Relationship between occupations and asbestosfibre content of the lungs in patients with pleural mesothelioma, lung cancer, and other diseases. Thorax. 1977; 32(4):377–386. https://doi.org/10.1136/thx.32.4.377 PMID: 929482
9. Robinson BW, Musk AW, Lake RA. Malignant mesothelioma. The Lancet. 2005; 366(9483):397–408. https://doi.org/10.1016/S0140-6736(05)67025-0
10. Er O, Tanrikulu AC, Abakay A, Temurtas F. An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease. Computers & Electrical Engineering. 2012; 38(1):75–81. https://doi.org/10.1016/j.compeleceng.2011.09.001
11. Yazicioglu S, Ilcayto R, Balci K, Sayli BS, Yorulmaz B. Pleural calcification, pleural mesotheliomas, and bronchial cancers caused by tremolite dust. Thorax. 1980; 35(8):564–569. https://doi.org/10.1136/thx.35.8.564 PMID: 7444823
12. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine. 2001; 23(1):89–109. https://doi.org/10.1016/S0933-3657(01)00077-X PMID: 11470218
13. Li M, Zhou ZH. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2007; 37(6):1088–1098. https://doi.org/10.1109/TSMCA.2007.904745
14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015; 13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696

15. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine. 2016; 375(13):1216. https://doi.org/10.1056/NEJMp1606181 PMID: 27682033

16. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Computing and Applications. 2013; 23(6):1737–1751. https://doi.org/10.1007/s00521-012-1134-8

17. Bao FS, Lie DYC, Zhang Y. A new approach to automated epileptic diagnosis using EEG and probabilistic neural network. In: 20th IEEE International Conference on Tools with Artificial Intelligence. vol. 2. IEEE; 2008. p. 482–486.

18. Shan Y, Zhao R, Xu G, Liebich H, Zhang Y. Application of probabilistic neural network in the clinical diagnosis of cancers based on clinical chemistry data. Analytica Chimica Acta. 2002; 471(1):77–86. https://doi.org/10.1016/S0003-2670(02)00924-8

19. Vilar S, Santana L, Uriarte E. Probabilistic neural network model for the *in silico* evaluation of anti-HIV activity and mechanism of action. Journal of Medicinal Chemistry. 2006; 49(3):1118–1124. https://doi.org/10.1021/jm050932j PMID: 16451076

20. Specht DF. Probabilistic neural networks. Neural Networks. 1990; 3(1):109–118. https://doi.org/10.1016/0893-6080(90)90049-Q

21. Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning. 2009; 2 (1):1–127. https://doi.org/10.1561/2200000006

22. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–444. https://doi.org/10.1038/nature14539 PMID: 26017442

23. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

24. Denisko D, Hoffman MM. Classification and interaction in random forests. Proceedings of the National Academy of Sciences (PNAS). 2018; 115(8):1690–1692. https://doi.org/10.1073/pnas.1800256115

25. Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1(1):14–23.

26. Holte RC. Very simple classification rules perform well on most commonly used datasets. Machine Learning. 1993; 11(1):63–90. https://doi.org/10.1023/A:1022631118932

27. Cangelosi D, Pelassa S, Morini M, Conte M, Bosco MC, Eva A, et al. Artificial neural network classifier predicts neuroblastoma patients' outcome. BMC Bioinformatics. 2016; 17(12):83.

28. Chicco D, Sadowski P, Baldi P. Deep autoencoder neural networks for Gene Ontology annotation predictions. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2014. p. 533–540.

29. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature Biotechnology. 2015; 33(8):831–838. https://doi.org/10.1038/nbt.3300 PMID: 26213851

30. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 2016; 32(12):i52–i59. https://doi.org/10.1093/bioinformatics/btw252 PMID: 27307644

31. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 2011; 11(1):1. https://doi.org/10.1186/1472-6947-11-51

32. Ward MM, Pajevic S, Dreyfuss J, Malley JD. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests. Arthritis Care & Research. 2006; 55(1):74–80. https://doi.org/10.1002/art.21695

33. Thongkam J, Xu G, Zhang Y. AdaBoost algorithm with random forests for predicting breast cancer survivability. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 3062–3069.

34. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7(1):3. https://doi.org/10.1186/1471-2105-7-3 PMID: 16398926

35. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Modern Pathology. 2005; 18(4):547–557. https://doi.org/10.1038/modpathol.3800322 PMID: 15529185

36. Cai Z, Xu D, Zhang Q, Zhang J, Ngai SM, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molecular BioSystems. 2015; 11(3):791–800. https://doi.org/10.1039/c4mb00659c PMID: 25512221

37. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017; 10 (35):1–17.

38. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. Perspectives in Clinical Research. 2017; 8(3):148. https://doi.org/10.4103/picr.PICR_87_17 PMID: 28828311

39. Breiman L, Cutler A. Random forests—Gini importance; 2004. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#giniimp. URL visited on 31st October 2016.

40. University of California Irvine. Machine Learning Repository; 2016. https://archive.ics.uci.edu/ml/datasets/Mesothelioma%C3%A2%E2%82%AC%E2%84%A2s+disease+data+set+. URL visited on 31st October 2016.

41. Fernandes K, Chicco D, Cardoso JS, Fernandes J. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. PeerJ Computer Science. 2018; 4:e154. https://doi.org/10.7717/peerj-cs.154

42. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 2(3):18–22.

43. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2008. p. 313–325.

44. Guan D, Yuan W, Lee YK, Najeebullah K, Rasel MK. A review of ensemble learning based feature selection. IETE Technical Review. 2014; 31(3):190–198. https://doi.org/10.1080/02564602.2014.906859

45. Ruck DW, Rogers SK, Kabrisky M, Oxley ME, Suter BW. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Transactions on Neural Networks. 1990; 1(4):296–298. https://doi.org/10.1109/72.80266 PMID: 18282850

46. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Cognitive Modeling. 1988; 5(3):1.

47. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. International Journal of Machine Learning Technology. 2011; p. 1–27.

48. Probst P, Boulesteix AL. To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research. 2018; 18(181):1–18.

49. CrossValidated. Do we have to tune the number of trees in a random forest?; 2018. https://stats.stackexchange.com/questions/348245/do-we-have-to-tune-the-number-of-trees-in-a-random-forest. URL visited on 4th October 2018.

50. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010; 26(10):1340–1347. https://doi.org/10.1093/bioinformatics/btq134 PMID: 20385727

51. Sculley D. Rank aggregation for similar items. In: Proceedings of the 2007 SIAM International Conference on Data Mining. SIAM; 2007. p. 587–592.

52. Madhu Sanjeevi. Chapter 4: decision trees algorithms; 2017. https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1. URL visited on 8th November 2018.

53. Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. Journal of Computer Science. 2006; 2(2):194–200. https://doi.org/10.3844/jcssp.2006.194.200

54. Koh HC, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management. 2011; 19(2):65.

55. Welton NJ, Sutton AJ, Cooper N, Ades A, Abrams KR. Evidence synthesis for decision making in healthcare. vol. 132. John Wiley & Sons; 2012.

56. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Annals of Translational Medicine. 2018; 6(11). https://doi.org/10.21037/atm.2018.05.32

57. Skocik M, Collins J, Callahan-Flintoft C, Bowman H, Wyble B. I tried a bunch of things: the dangers of unexpected overfitting in classification. bioRxiv. 2016;(078816).

58. Stolovitsky G, Mangravite L. DREAM Challenges; 2007. https://www.dreamchallenges.org/. URL visited on 12th November 2018.

59. Kueffner R, Zach N, Bronfeld M, Norel R, Atassi N, Balagurusamy V, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. bioRxiv. 2018;(294231).

60. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945; 1(6):80–83. https://doi.org/10.2307/3001968

61. Upton GJ. Fisher's exact test. Journal of the Royal Statistical Society. 1992; p. 395–402. https://doi.org/10.2307/2982890

62. Shevchuk Y. NeuPy, neural networks in Python; 2015. https://www.neupy.com. URL visited on 31st October 2016.

63. Collobert R, Bengio S, Mariéthoz J. Torch: a modular machine learning software library. Institut Dalle Molle d'intelligence artificielle perceptive (Idiap), Martigny, Switzerland; 2002.

64. von Jouanne-Diedrich, Holger. OneR: one rule machine learning classification algorithm with enhancements; 2017. https://cran.r-project.org/web/packages/OneR/. URL visited on 31st July 2017.

65. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines; 1997.

66. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications. 2009; 36(3):5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108

67. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering. 2008;(9):1263–1284.

68. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer; 2005. p. 878–887.

69. Breiman L. Bagging predictors. Machine Learning. 1996; 24(2):123–140. https://doi.org/10.1023/A:1018054314350

70. Malignant Pleural Mesothelioma Staging. TNM classification for malignant pleural mesothelioma; 2016. http://emedicine.medscape.com/article/1999306-overview. URL visited on 31st October 2016.

71. Pass HI, Lott D, Lonardo F, Harbut M, Liu Z, Tang N, et al. Asbestos exposure, pleural mesothelioma, and serum osteopontin levels. New England Journal of Medicine. 2005; 353(15):1564–1573. https://doi.org/10.1056/NEJMoa051185 PMID: 16221779

72. American Cancer Society. Malignant mesothelioma stages; 2017. https://www.cancer.org/cancer/malignant-mesothelioma/detection-diagnosis-staging/staging.html. URL visited on 13th May 2018.

73. Kao SC, Vardy J, Chatfield M, Corte P, Pavlakis N, Clarke C, et al. Validation of prognostic factors in malignant pleural mesothelioma: a retrospective analysis of data from patients seeking compensation from the New South Wales Dust Diseases Board. Clinical Lung Cancer. 2013; 14(1):70–77. https://doi.org/10.1016/j.cllc.2012.03.011 PMID: 22658812

74. MesotheliomaWeb. Thrombocytopenia in mesothelioma patients; 2016. http://www.mesotheliomaweb.org/thrombocytopenia.htm. URL visited on 31st October 2016.

75. Chicco D, Ciceri E, Masseroli M. Extended Spearman and Kendall coefficients for gene annotation list correlation. In: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer; 2014. p. 19–32.

76. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3(Jan):993–1022.

77. Pinoli P, Chicco D, Masseroli M. Computational algorithms to predict Gene Ontology annotations. BMC Bioinformatics. 2015; 16(6):S4. https://doi.org/10.1186/1471-2105-16-S6-S4 PMID: 25916950

78. Chicco D, Masseroli M. Software suite for gene and protein annotation prediction and similarity search. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2015; 12(4):837–843. https://doi.org/10.1109/TCBB.2014.2382127 PMID: 26357324