Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Prognosticating Mesothelioma Using Predictive Analytics

Avishek Choudhury
Research Assistant
School of Systems and Enterprises
Stevens Institute of Technology
https://orcid.org/0000-0002-5342-0709
ResearcherID: P-2415-2018

Correspondence:

Avishek Choudhury (Research Assistant)

School of Systems and Enterprises

Stevens Institute of Technology

Hoboken, USA

Email: achoudh7@stevens.edu

Mobile: +1 (515) 608-0777

© 2019 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

Prognosticating Mesothelioma Using Predictive Analytics

Abstract

Background: Malignant pleural mesothelioma (MPM) is an atypical, belligerent tumor that matures into cancer in the pleura, a stratum of tissue bordering the lungs. Pleural mesothelioma is a common type of mesothelioma that accounts for about 75 percent of all mesothelioma diagnosed yearly in the United States. Diagnosis of mesothelioma takes several months and is expensive. Given the difficulty of diagnosing MPM, early identification is crucial for patient survival. Our study implements artificial intelligence and recommends the best fit model for early diagnosis and prognosis of MPM. **Method:** We retrospectively retrieved patient's medical reports generated by Dicle University, Turkey and implemented multi-layered perceptron (MLP), voted perceptron (VP), Clojure classifier (CC), kernel logistic regression (KLR), stochastic gradient decent SGD), adaptive boosting (AdaBoost), Hoeffding tree (VFDT), and primal estimated sub-gradient solver for support vector machine (s-Pegasos). We evaluated the models, compared and tested using $paired\ T - test\ (corrected)$ at 0.05 significance based on their respective classification accuracy, f-measure, precision, recall, root mean squared error, receivers characteristic curve (ROC), and precision-recall curve (PRC). **Results:** In phase-1 SGD, AdaBoost.M1, KLR, MLP, VFDT generates optimal results with the highest possible performance measures. In phase-2, AdaBoost with a classification accuracy of 71.29% outperformed all other algorithms. C-reactive protein, platelet count, duration of symptoms, gender, and pleural protein were found to be the most relevant predictors that can prognosticate mesothelioma. **Conclusion:** This study confirms that data obtained from biopsy and imagining tests are strong predictors of mesothelioma but are associated with high cost, however, can identify mesothelioma with optimal accuracy. Predictive analytics without using biopsy results can diagnose mesothelioma with acceptable accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

Implementation of phase-2 followed by phase-1 can address diagnosis expenses and maximize disease prognosis. Additionally, results indicate improved MPM diagnosis using AI methods dependent upon the specific application.

Keywords: Mesothelioma; Predictive modeling; Decision support system; Early diagnosis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Prognosticating Mesothelioma Using Predictive Analytics

1. Background

Malignant pleural mesothelioma (MPM) is a hostile tumor of mesothelial cells concomitant with preceding asbestos contact. With an amplified implementation of chemotherapy (Vogelzang, Rusthoven, Symanowski, & al., 2003) (Zalcman, et al., 2016) and a varied gamut of clinical examinations, precise prognostication is a crucial subject for individuals with MPM, doctors, and scholars. However, MPM is an outstandingly different ailment. Staging system (Pass, Giroux, Kennedy, & al., 2016), challenging primary tumor identification process (Gill, Naidich, Mitchell, & al., 2016;) (Frauenfelder, Tutic, Weder, & al., 2011;) and distinct biology (Bueno, Stawiski, Goldstein, & al., 2016;), impedes accurate prediction. MM is a rare disease; it affects about two individuals per million per annum in a general population (McDonald, C., & McDonald., 1996). Comparatively industrialized nations are affected more by MM (Spirtas, et al., 1986;) (Peto, Hodgson, Matthews, & Jones, 1995;) (Leigh, et al., 1991;) due to higher exposure to asbestos (Metintas, et al., 2008). Severity of mesothelioma can be categorized into stage 1, stage 2, stage 3, and stage 4 (cancer). Stage1 and stage 2 symptoms of MPM such as dry coughing, dyspnea, respiratory complications, chest or abdominal pain, fever, pleural effusion, fatigues, and muscle weakness are very weak predictors of mesothelioma (Mesothelioma News, 2018). Since mesothelioma is rare, patients are less likely to be suspected with the disease. Moreover, its initial symptoms during stage 1 and 2 resemble other diseases such as pneumonia or irritable bowel syndrome (Selby, 2018), MM can also be mistaken for an infection or a more common type of non-terminal lung cancer that develops in mucus-secreting glands called adenocarcinoma (Selby, 2018). If mesothelioma is not diagnosed and meets no medical aid at its premature stage, it rapidly burgeons into a stage 3 or stage 4 cancer. Unfortunately, the survival rate after being diagnosed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1
Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

with late stage mesothelioma is typically about a year. In order to treat mesothelioma effectively, an early diagnosis is recommended.

Diagnosing mesothelioma is challenging, and the expenses associated with identifying this disease can ascend rapidly. In fact, since the principal way to diagnose mesothelioma incorporates ruling out other plausible diseases, more frequently than not, many examinations may be administered that aren't exclusive to mesothelioma itself but are for erstwhile disorders instead (Molinari, 2018). Furthermore, it is often suggested to get a second opinion (Molinari, 2018), recapping many of the diagnostic tests over and over. For all these causes, diagnostic expenses for mesothelioma starts piling up even before the required treatment commences. Mesothelioma diagnosis typically implicates taking imaging scans of tumors, examining a biopsy of cancer tissue, and blood tests (Selby, 2018).

Oncologists use imaging tests to look for noticeable signs of tumors. A mesothelioma diagnosis depends on a series of diagnostic imaging tests, including X-rays, CT scans, MRIs and PET scan (Selby, 2018) all of which are expensive.

Two chief factors make imaging tests expensive. Foremost, the specialized imaging equipment is expensive both for an upfront purchase and for maintenance. Secondly, this equipment requires well-trained technicians to ensure apt operation of the device. A patient can presume to spend about $800 – $1,600 (Molinari, 2018) for a single CT, MRI, or PET scan respectively. Moreover, multiple scans may be required during diagnosis (Molinari, 2018), which can quickly bourgeon the overall costs.

The most accurate test for confirming mesothelioma is a biopsy (Selby, 2018). It is a procedure that requires removal of fluid or tissue samples from the tumor or cancer site and their analysis under a microscope. There are many diverse approaches to obtaining a biopsy, and which one to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

be used depends on the suspected tumors' location. Some biopsies embrace making an incision and inserting implements to obtain a sample of the tumor cell, while others only require a needle. Given the wide range of biopsy procedures, its expenses can range from $500 to $700 for a needle biopsy (Molinari, 2018), $3,600 to $ 5000 for pleuroscopy (lungs) or laparoscopy (abdomen) (Molinari, 2018), $7,800 to $7,900 for thoracotomy (lungs) or laparotomy (abdomen) (Molinari, 2018). Like other diagnostic procedures, biopsies may also require to be performed multiple times (Molinari, 2018), increasing the overall diagnosis expenses. Doctors also explore a variety of blood tests such as MESOMARK, SOMAmer, and Human MPF to look for biomarkers that suggest mesothelioma (Selby, 2018). However, currently, no blood tests are precise enough to confirm a diagnosis on their own (Selby, 2018).

2. Problem statement

Malignant Pleural mesothelioma has the potential to grow into cancer and sabotage patient health. Like any other fatal disease, malignant mesothelioma demands early diagnosis and effective treatment. However, effective diagnosis methods such as thoracotomy and pleuroscopy are costly and might not be affordable for patients worldwide (Friedin, 2012) (Pope, 2010). Additionally, about two third of the world do not have adequate access to the required technologies, expensive imaging devices, and expert technicians (Silvester, 2016).

There exists some work of literature that has used artificial intelligence or machine learning algorithms such as decision tree, random forest, support vector machine, and even artificial neural network to identify MM (Choudhury , Identification of Cancer: Mesothelioma's Disease Using Logistic Regression and Association Rule, 2018) (Ilhan & Celik, 2016) but with some limitations. These models (random forest, decision tree, and others) either tend to overfit (Tin, 1995) or fails to generate 100% accuracy or might also fail to converge a large dataset (Lotfi & Keshavarz, 2014).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

In our study, we propose a model that overcome the aforementioned flaws and can diagnose MM with and without requiring data from expensive biopsy and imaging tests.

3. Methodology

Our study uses the patient's medical reports generated by Dicle University. The dataset contains 34 attributes, one binary response variable, and 324 instances. It consists of 41% females and 59% males. The patients involved in this study were in nine different cities. We performed k-fold cross-validation to minimize any bias and variance in the dataset. Cross-validation is a resampling technique used to gauge machine learning models on a limited dataset. In this method, the original data sample is randomly partitioned into *k* proportional subsamples. Of the *k* subsamples, one subsample is retained as the validation data for evaluating the model, and the remaining *k-1* subsamples are used as training data. The cross-validation process is then reiterated *k* times. The *k* results obtained from the *k*-folds are then averaged to produce a single estimation. In this study we considered the value of *k* to be 10 becoming 10-fold cross-validation. The selection of *k* is usually 5 or 10 (Kuhn & Johnson, 2018). There is a bias-variance trade-off related to the value of *k* in k-fold cross-validation. Performing k-fold cross-validation using $k = 5$ or $k = 10$ have empirically shown to yield test error rate estimates that free from extreme high bias and variance (James, Witten, Hastie, & Tibshirani, 2017). All the analysis was performed using R-studio, an open source machine learning and statistical tool, and *Waikato Environment for Knowledge Analysis* (WEKA), a free software suite of machine learning licensed under the GNU General Public License, programmed in JAVA, and developed at the University of Waikato, New Zealand.

Table 1 below lists all the attributes contained in our dataset, it also determines the mean, deviation and logistic correlation of all predictors with the target variable ("class of diagnosis"). In classification applications, calculating logistic dependencies between a single input and single

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

target or class variable is essential. It determines the absolute values of the logistic correlation between all inputs and all targets. The logistic correlation is a numerical value between *zero* and *one* that expresses the strength of the logistic relationship between a single input and output variables. A value close to *one* indicates a healthy relationship and value approaching *zero* denotes weak or no relationship.

Table 1: Data Statistics

Predictor	**Mean**	**Deviation**	**Logistic correlation with the target variable ("class of diagnosis")**
Age	54.74	11.00	0.06
Gender	-	-	0.15
City	NA	NA	0.02
Asbestos exposure	0.86	0.34	0.07
Type of MM	0.05	0.26	0.13
Duration of asbestos exposure	30.18	16.41	0.06
Diagnosis method*	-	-	1.00*
Keep side	0.75	0.56	0.10
Cytology	0.28	0.45	0.02
Duration of symptoms	5.44	4.71	0.02
Dyspnea	0.81	0.38	0.02
Ache on chest	0.68	0.46	0.05
Weakness	0.61	0.48	0.06
Habit of cigarette	0.91	1.15	0.05
Performance status	0.52	0.50	0.03
White blood	9457.45	3450.73	0.05
Cell count (WBC)	9.55	3.34	0.05
Hemoglobin (HGB)	0..42	0.49	0.03
Platelet count (PLT)	369.65	227.55	0.06
Sedimentation	70.68	21.74	0.00

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

Blood lactic dehydrogenize (LDH)	308.91	185.14	0.01
Alkaline phosphate (ALP)	66.16	35.07	0.04
Total protein	6.58	0.82	0.01
Albumin	3.30	0.63	0.04
Glucose	112.41	38.46	0.01
Pleural lactic dehydrogenize	518.47	536.27	0.03
Pleural protein	3.93	1.57	0.03
Pleural albumin	2.07	0.91	0.07
Pleural glucose	48.44	27.23	0.01
Dead or not	-	-	-
Pleural effusion	0.87	0.33	0.03
Pleural thickness on tomography	0.59	0.49	0.01
Pleural level of acidity (pH)	0.52	0.50	0.04
C reactive protein (CRP)	64.18	22.66	0.11

***Diagnosis method contains data obtained from biopsy and imaging tests. It contains binary values where 1= biopsy or imaging test indicates MM; 0 = otherwise.**

Mesothelioma data set can be broadly divided into pre-diagnosis data and post-diagnosis data. Pre-diagnosis data refers to the all the records obtained before mesothelioma was clinically confirmed such as patient age, gender, the city they belonged to, smoking habit, exposure to asbestos, duration of exposure to asbestos, early-stage symptoms including the feeling of weakness, heartache, and dyspnea, and duration of symptoms. Pre-diagnosis data also encompasses blood test results such as white blood cell count, hemoglobin level, platelets count and others.

Post-diagnosis are those data that refers to the records retrieved after mesothelioma was confirmed. Type of mesothelioma detected (type of MM), side effects of chemotherapy (keep

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

side), and survival of the patient after treatment (dead or not) are all post-diagnosis data. This study eliminates the "dead or not" predictor from all analysis.

Table 1 above indicates that the predictor “diagnosis method” is strongly correlated with the target variable. The predictor “diagnosis method” refers to data obtained from invasive biopsy, and imaging test results. Invasive biopsy and imaging tests can accurately identify mesothelioma but are expensive procedures and may require repeated examinations as stated earlier. To advocate the applicability of AI predictive analytics on both pre and post diagnostic data we perform a comparative analysis of classification models into two phases. Phase-1 models use all the predictor variables except "dead or not" as input to produce high classification accuracy. The same set of models in Phase-2 only takes relevant predictors from pre-diagnosis data as its input.

Phase-1 and phase-2 are denoted as *high accuracy* and *low-cost* phases respectively because phase-1 execution demands data from expensive, invasive biopsy and imaging test results which are robust predictors of MM (logistic correlation = 1) and thus the model is expected to yield high accuracy. Whereas, phase-2 considers only predictors with lower logistic correlation (pre-diagnosis data) and eliminates the use of invasive biopsy and imaging test results. Execution of phase-2 also incorporates a feature selection method to enhance its accuracy and reduce computational time.

Data sets are often designated with too many variables for effective model structure (Miron & Witold, 2010). Commonly most of these variables are extraneous to the classification, and perceptibly their relevance is unknown in advance (Miron & Witold, 2010). Several difficulties arise while dealing with large feature sets. One is decently technical — dealing with large feature sets impedes computational speed, consumes too many resources and is merely bothersome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

Another is even more important — many machine learning algorithms reveal a diminution of accuracy when the number of variables is considerably higher than optimal (Ron & George, 1997). Therefore, selection of minimal feature set that can yield the best possible classification outcome is needed for practical reasons (Miron & Witold, 2010). This problem also known as the minimal-optimal problem (Nilsson, Peña, Björkegren, & Tegńer, 2007), has been intensively analyzed and there are several algorithms which are established to reduce the feature set to a manageable and optimal size (Miron & Witold, 2010).

Nevertheless, this genuine goal sleuths another problem — the identification of all attributes which are in certain circumstances germane for classification, the so-called "all-relevant problem" (Miron & Witold, 2010). Finding all relevant attributes, instead of the non-redundant ones, may be beneficial. This is essential when one is involved in understanding the fundamental mechanisms related to the subject of interest, instead of purely building a black box prognostic model. For example, when dealing with classification of Mesothelioma dataset, identification of all predictors which are related to the outcome ("Healthy" or "Diseased") is necessary for complete understanding of the process, whereas a minimal-optimal set of predictors (variables) might be more useful as classification markers. An honest discussion demarcating the importance of finding relevant attributes is given by Nilsson et al. in 2007 (Nilsson, Peña, Björkegren, & Tegńer, 2007).

The phase-2 of our study implements Boruta algorithm for selecting all relevant predictor (Choudhury & Greene, Evaluating Patient Readmission Risk: A Predictive Analytics Approach, 2018). Boruta algorithm is a wrapper built around the random forest classification algorithm (Miron & Witold, 2010) implemented in the R random forest package (Liaw & Wiener, 2002). Boruta algorithm uses Z-score as the importance measure since it considers the fluctuations of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

mean accuracy loss among trees in the forest (Miron & Witold, 2010). Since we cannot use Z-score unswervingly to gauge importance, an external reference is needed to decide whether the importance of any given attribute is significant. To determine the importance of each attribute Boruta algorithm creates an analogous ‘shadow’ attribute, whose values are obtained by shuffling values of the original attribute across objects (Miron & Witold, 2010). Then a classification is performed using all the attributes of the extended system to calculate the importance of all variables. The importance of a shadow attribute can be nonzero purely due to random fluctuations (Miron & Witold, 2010). Thus, the set of the importance of shadow attributes is used as a reference for determining essential attributes (Miron & Witold, 2010).

The following algorithms were implemented, compared and tested using $paired\ T-test\ (corrected)$ at 0.05 significance.

3.1. Algorithms

Stochastic Gradient Descent (SGD)

Gradient descent is a method to determine the local minima. Stochastic gradient descent is gradient descent performed using multiple updates at a time on a small batch (minibatch) of the dataset selected at random (stochastically). Instead of calculating the gradient of the cost (error) based on the whole dataset, SGD break the dataset into mini batches and compute the gradient on each batch separately followed by a neural net update based on the partial gradient. In other words, it is an optimization algorithm that iteratively determines the values of learnable parameters of a function (*f*) to minimize the cost function (error rate). Cost function for our study is root mean squared error, which can be determined using the following equation (eq.1).

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2} \tag{1}$$

Mathematically, SGD is a simplification of gradient descent. Instead of calculating the gradient of $E_n(f_w)$ (empirical risk using gradient descent), each iteration estimates this gradient by a single randomly picked example (eq.2):

$$z_t: w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t). \tag{2}$$

Where *z* is a random pair of input *x* and scalar output *y*; *w* is weight; *γ* is learning rate; *Q(z,w)* is the loss. Since the stochastic algorithm does not require to retain which examples were visited during the previous iterations, it can process examples on the fly in a deployed system.

Adaptive Boosting M1

It is also known as AdaBoost.M1, is a machine learning meta-algorithm that can be implemented in conjunction with other types of learning algorithms to convalesce performance. The output of the other learning algorithms ('weak learners') is merged into a weighted sum that epitomizes the final output of the boosted classifier. AdaBoost is adaptive since it can fine-tune the weak learners in favor of misclassified instances by previous classifiers. AdaBoost-M1 refers to a specific method of training a boosted classifier (eq.3).

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{3}$$

Where *T* is the number of iterations; each f_t is a weak learner that takes an object x as input and returns a value indicating the class of the object. Each weak learner produces an output

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

hypothesis, $h(x_i)$, for each sample in the training set. At each iteration *t*, a weak learner is selected and assigned a coefficient α_t such that the sum of training error E_t (eq.4) of the resulting *t*-stage boost classifier is minimized.

$$E_t = \sum E|F_{t-1}(x_i) + \alpha_t h(x_i) \tag{4}$$

Where $F_{t-1}(x)$ is the boosted classifier that has been built up to the previous stage of training. $E(F)$ is some error function, and $f_t(x) = \alpha_t h(x)$ is the weak learner that is being considered for addition to the final classifier.

Kernel Logistic Regression (KLR)

It is a well-established statistical model for classification. Unlike Logistic Regression, KLR enables the classification of linearly non-separable problems by assigning the input variables to a higher dimensional space, via the kernel trick. The kernel is a conversion function that must satisfy mercer’s necessary and sufficient conditions, which state that a kernel function must be expressed as an inner product and must be positive semidefinite.

Multi-layer Perceptron

The Artificial Neural Network (ANN), also known as a neural network, is a computational prototype based on the biological neural network. Its fundamental theory originated in the connectionism of cognitive science in which several simple computational units are linked to show intelligent comportments. Such a concept is germane to the neurons of the biological neural system and the computational units of computational prototypes. A typical ANN comprises of an input layer, hidden layer(s), and an output layer. The first layer known as the input layer consists of a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

neuron set $\{x_i | x_1, x_2, \dots, x_m\}$ denoting the input variables. Each neuron in the hidden layer transforms the values from the preceding layer with a weighted linear summation $w_1 x_1 + w_w x_2 + \cdots + w_m x_m$, Followed by a non-linear activation function such as hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into the output values.

Figure 1 shows a typical neuron model, which is comprised of two parts. The first part is the accretion of signals, where the input signals (input data) are gathered for a sum. As shown in the following equation (eq.5), each weight (w_i) equals a data dimension (x_i), while (w_0) as a bias is correspondent to the intercept or constant term of the function. While the constant is set to ‘‘1" as the input of 0th dimension, the bias is managed as the weight of 0th dimension. This is also called affine transformation (Lee, Chen, Yu, & Lai, 2018).

$$Z = bias + \sum_{i=1}^{m} X_i W_i = \sum_{i=0}^{m} X_i W_i \quad (5)$$

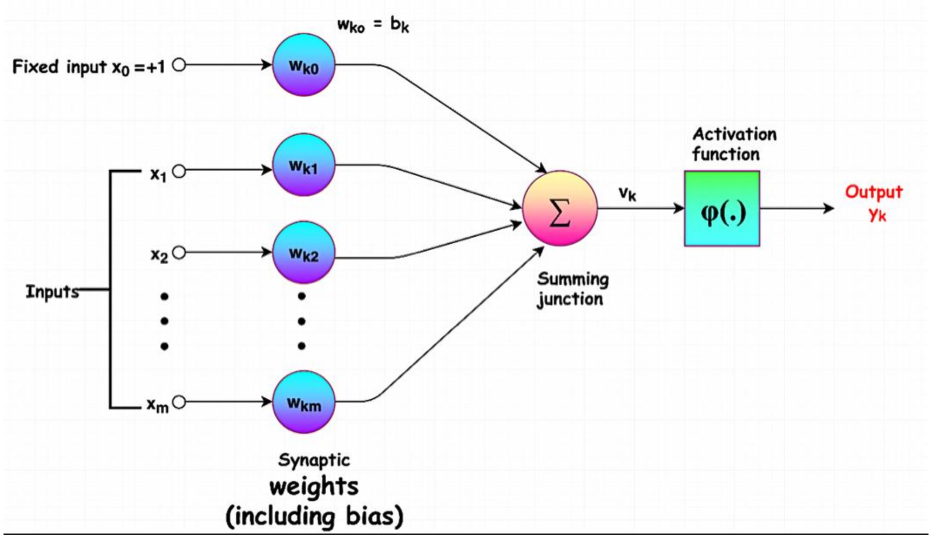

Figure 1.Typical neural network

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

The second part is the initiation of the function, where the obtained activation value is used for the nonlinear compressed transformation to extricate a nonlinear eigenvalue. The frequently-used activation functions include ReLU, Sigmoid, and Tanh (Lee, Chen, Yu, & Lai, 2018). A neural network is a network based on the interconnection between artificial neurons. The feedforward neural network (FNN) or multilayer perceptron (MLP) is a neural network that permits the feedforward connection of neurons. The input of data is known as the input layer, while the output of results is termed as the output layer; the layers between the input layer and the output layer are called the hidden layers (Lee, Chen, Yu, & Lai, 2018). MLP is a supervised algorithm that learns a function $f(.): R^m \rightarrow R^o$ by training on a given dataset, where m is the dimension for input and O is the output dimension. Provided a set of features $X = x_1, x_2, x_3, \dots, x_m$ and target y, it can learn a non-linear function for either classification or regression.

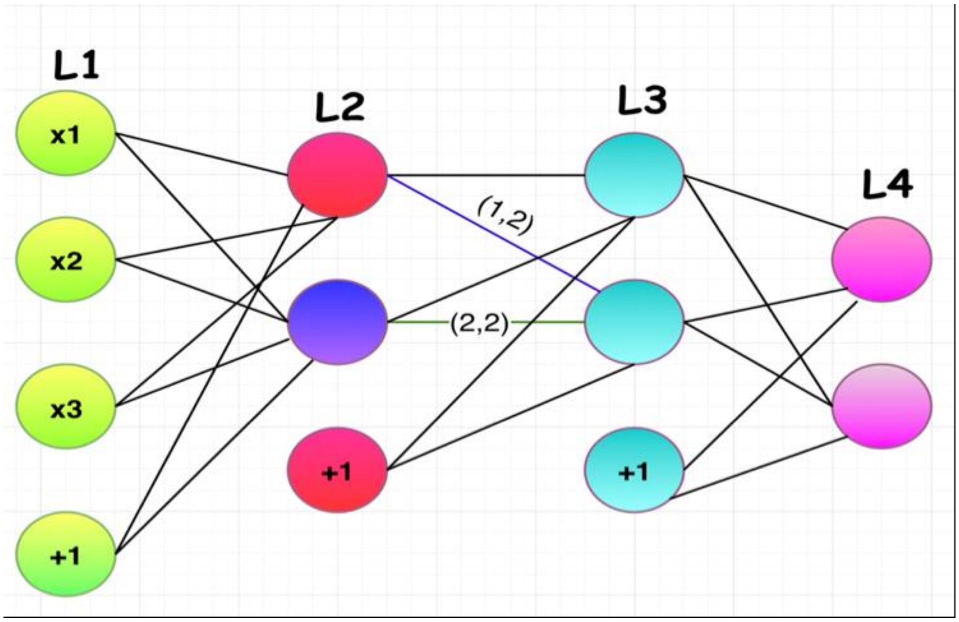

Figure 2.Feed-forward neural network

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Figure 2 shows a 4 layered neural network, where the first layer (L1) is the input layer; L2 and L3 are the hidden layers; L4 is the output layer; $a_{i,j}^{(l)}$ refers to the connection weight of *"i"* (ordinal number) neuron on layer I and *"j"* (ordinal number) neural on layer I+1; a_j^l denotes the connection between the bias on layer I and *"j"* neuron on layer I+1; and a_j^l implies the activation value (output value) of the *"I"* neuron on layer I, and the activation value of the blue neuron in the picture is $a_2^{(2)}$ (Lee, Chen, Yu, & Lai, 2018).

Voted Perceptron (VP)

It is designed for linear classification, that combines the Rosenblatt's perception algorithm with Helmbold and Warmuth's leave-out method. All weight vectors confronted during the learning process vote on a prediction. The measure of the accuracy of a weight vector, based on the number of trials in which it correctly classifies instances, is used as the number of votes given to the weight vector. The output a voted perceptron is given by (eq.6) when given labeled data is (x_i, y_i) where y is *+1* or *-1* (mesothelioma or healthy):

$$y_i = sign\left\{\sum_{p=0}^{P} c_p \; sign(w_p, x)\right\} \qquad (6)$$

Where x are inputs, $p = 0,1,2, \dots, P$; w_p are weights, y_i is the predicted class, and c_p is the survival time (reliability of w_p).

Hoeffding Tree

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

It is also known as Very Fast Decision Tree (VFDT) is a tree algorithm for data stream classification. The Hoeffding tree is an incremental decision tree learner for a large dataset, that assumes that the data distribution is constant over time. It grows a decision tree based on the theoretical guarantees of the Hoeffding bound. In other words, VFDT employs Hoeffding bound to decide the minimum number of arriving instances to achieve a certain level of confidence in splitting the node. The confidence level determines the proximity of the statistics between the attribute chosen by VFDT and the attribute chosen by decision tree for batch learning.

Clojure Classifier (CC)

It is a wrapper classifier developed in Clojure programming language. It mandates to have at least a learn-classifier function and distribution-for-instance function. The learn-classifier function takes an object and a string (nullable) and returns the learned model as a serializable data structure. The distribution-for-instance function takes an instance to be predicted and a model as an argument and returns the prediction as an array.

2.1.1. Primal Estimated sub-Gradient Solver for SVM

It is also known as s-Pegasos. It performs SGD on a primal objective (eq. 7,8) with carefully chosen step size.

$$\min_{w} \frac{\lambda}{2}||W||^2 + \frac{1}{m}\sum_{(x,y)\in S} l\big(W;(X,y)\big) \tag{7}$$

Where

$$l\big(w;(X,y)\big) = \max\{0,1 - y(w,X)\} \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

3.2.Model evaluation

While evaluating supervised machine learning models, it is important to measure each model's classification *accuracy*, *f-measure*, *recall*, *precision*, *root mean squared error* (RMSE), *receiver operating characteristic* (ROC), and *precision-recall curve* (PRC).

Classification accuracy is the metric for evaluating classification models. It is the fraction of predictions or classification that a model performs correctly. Classification accuracy can be calculated by the given equation (eq.9)

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

Where *TP* = Ture positive; *TN* = True negative; *FP* = False positive; *FN* = False negative.

The ROC curve is the graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. In the machine learning domain, a TPR is also known as sensitivity, recall or "probability of detection." Similarly, an FPR is known as the fall-out or "probability of false alarm" and can be calculated as (eq. 10). The ROC curve is thus the sensitivity as a function of fall-out.

$$FPR = (1 - specificity) \qquad (10)$$

Regarding information retrieval undertakings with binary classification (relevant or not relevant), precision is the segment of retrieved instances that are relevant, whereas recall, also known as sensitivity is the fraction of retrieved instances to all relevant instances. In this context of information retrieval, the PRC becomes very useful. PRC is a graphical representation of recall

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

(x-axis) and precision (y-axis), where recall and precision are determined using the given formula (eq. 11,12) respectively.

$$Recall = \frac{TP}{(TP + FN)} \tag{11}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{12}$$

f-measure, also known as F1-score is the harmonic mean of precision and recall (eq.13), where *f-measure* reaches its best at 1 and worst at 0.

$$f1\ score = \frac{2 * (precision * recall)}{precision + recall} \tag{13}$$

The root-mean-square error (RMSE) is a measure of performance of a model. It does this by computing the difference between predicted and the actual values as given below (eq. 14).

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{N}} \tag{14}$$

Where $(x_i - y_i)$ is the difference between predicted and actual value and N is the sample size.

4. Results

Phase 1

As shown in table 2, SGD, AdaBoost.M1, KLR, MLP, VFDT generates perfect results with 100% accuracy, precision, recall, and f-measure. These algorithms also return the highest possible ROC, PRC, and zero RMSE. s-Pegasos also delivers close to the optimal result.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

In this phase, the high accuracy of 100% indicates that results obtained from biopsy and imaging tests are very strong predictors of MM. This result validates the significance of biopsy and imaging results ("diagnosis method") from a data science viewpoint.

Table 2. Comparing classification accuracy (phase-1)

Algorithm	SGD	AdaBoost.M1	KLR	MLP	VP	VFDT	CC	s-Pegasos
Classification accuracy (%)	100	100	100	100	70.38	100	70.38	99.36
f-measure	1.00	1.00	1.00	1.00	0.83	1.00	0.83	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Precision	1.00	1.00	1.00	1.00	0.70	1.00	0.70	0.99
ROC	1.00	1.00	1.00	1.00	0.50	1.00	0.50	0.99
PRC	1.00	1.00	1.00	1.00	0.70	1.00	0.70	0.99
RMSE	0.00	0.00	0.00	0.00	0.54	0.01	0.54	0.04

Phase 2 demonstrates the relevance of pre-diagnosis data. It also shows the behavior of all predicting models post removal of "diagnosis method" and other post-diagnosis data.

Phase 2

Boruta algorithm confirmed five relevant attributes that are enough to predict the presence of Mesothelioma without any loss in model's performance. In other words, the selected attributes alone can prognosticate MM with the same accuracy as all other pre-diagnosis predictors when taken together as input. The relevant predictor identified were *c-reactive protein, platelet count, duration of symptoms, gender*, and *pleural protein*.

This method neither downgrades the remaining predictors nor does it recommend revising the regular clinical procedures. Figure 3 below shows the attributes recognized by Boruta algorithm. Boruta plot generates a box plot for each attribute. The x-axis represents each of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

candidate explanatory variables. The green box plots refer to the relevant attributes whereas the red ones are identified as unimportant (from a data science viewpoint). The blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute created by the Boruta algorithm. The following table 3 compares the different performance measures of each algorithm used in this study.

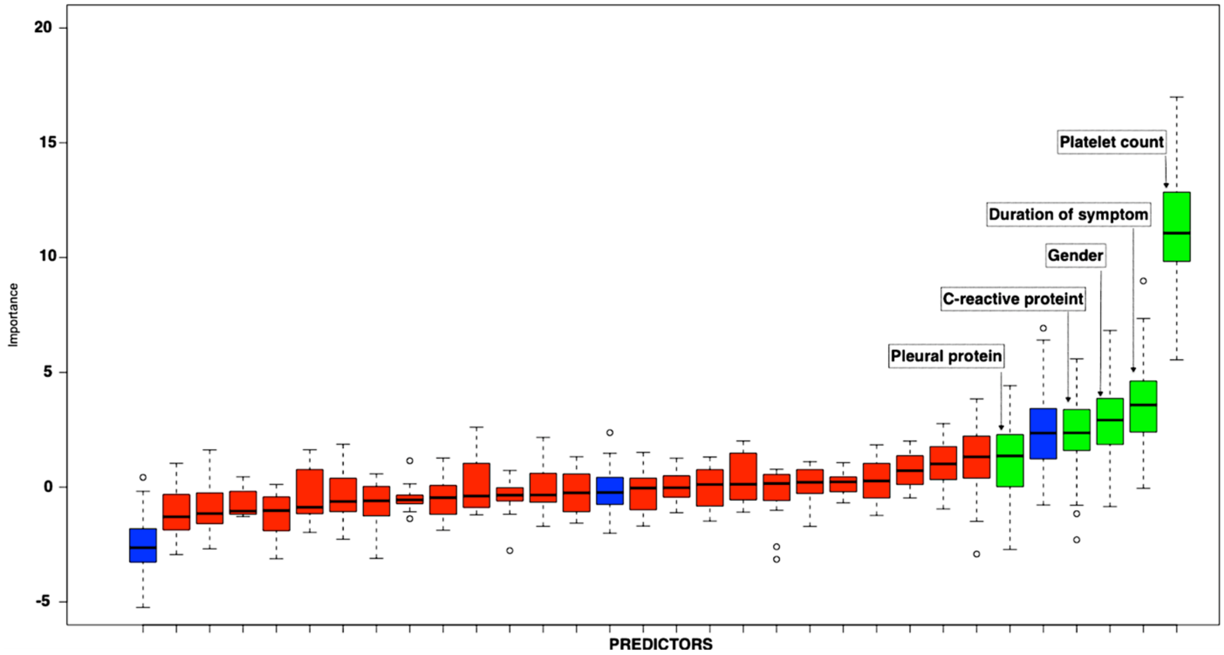

Figure 3.Boruta plot for feature selection

AdaBoost outperformed all other models with the highest classification accuracy of 71.29%. Excluding “diagnosis method” from the prediction model resulted in decreased accuracy. However, this phase has its own advantage. Despite lower accuracy, phase-2 helps reducing diagnostic expenses.

Table 3. Comparing classification accuracy (phase-2)

Algorithm	SGD	AdaBoost.M1	KLR	MLP-C	VP	VFDT	CC	s-Pegasos
Classification accuracy (%)	69.23	71.29	69.51	64.11	70.38	70.38	70.38	67.03
f-measure	0.80	0.82	0.79	0.74	0.83	0.83	0.83	0.77

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

Recall	0.80	0.86	0.93	0.84	1.00	1.00	1.00	0.81
Precision	0.74	0.74	0.76	0.75	0.70	0.70	0.70	0.75
ROC	0.58	0.61	0.65	0.61	0.50	0.50	0.50	0.58
PRC	0.74	0.77	0.82	0.79	0.70	0.70	0.70	0.74
RMSE	0.55	0.45	0.46	0.57	0.54	0.46	0.54	0.57

Discussion

An accurate diagnosis of MM is crucial at both the individual and public health level. It has necessary medicolegal significance due to diagnosis-related compensation (Ascoli, 2015). However, prognosticating MM is challenging due to its composite epithelial pattern and low likelihood of occurrence (Ascoli, 2015). To advocate the prognosis of MM with high accuracy and low diagnostic cost, the current study designed and implemented a prediction model comprising of two phases. (phase 1 and 2).

To our knowledge no previous studies have implemented our AI models and focused on reducing diagnosis expenses by eliminating biopsy and imaging test results from the dataset. Phase-2 of our study proposes AdaBoost.M1 algorithm that can identify high risk patients at lower-cost by taking only blood test results and patient's demographic data. Outcome from phase-2 can provide the doctors with a list of high-risk patients. Doctors and other healthcare providers can then prescribe biopsy tests only to the identified patients for reconfirming MM using phase-1 model with optimal accuracy. This approach will reduce unnecessary biopsy tests and thus reduce overall expenses by up to $7900 (Molinari, 2018).

The recommended model (AdaBoost) in phase-2 requires *c-reactive protein, platelet count, duration of symptoms*, *gender*, and *pleural protein* as its input. The expenses to collect the required input data can range from. $100 to $200 (Practo, 2017) for Protein Total Pleural Fluid (*pleural*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

protein), $40 to $70 (Haiken, 2011) for c-reactive protein test, and $6 to $167 (Pinder, 2012) for complete blood count (*platelet* count) depending up on the location. These factors can also advocate early prognosis of MM; Moreover, studies have shown that higher (*>1 mg/dL*) c-reactive protein influences mesothelioma (Takamori, et al., 2018) (Ghanim, et al., 2012), another study at the University of Maryland determined the clinical significance of preoperative thrombocytosis (high count of platelets), in patients with MPM (Li, et al., 2017).

5. Conclusion

Our study identifies that the *diagnosis method* (biopsy and imaging test results), *c-reactive protein*, *platelet count*, *duration of symptoms*, *gender*, and *pleural protein* plays a significant role in diagnosing MM. However, effective diagnosis methods such as pleuroscopy (lungs) or laparoscopy (abdomen), thoracotomy (lungs) or laparotomy (abdomen), and imaging tests (CT scan and MRI) are expensive. This study proposes two approaches to predict MM, each having its advantages and limitations. The first approach (phase-1) uses all predictors from mesothelioma data and produces 100% classification accuracy. The second approach (phase-2) ensures cost reduction. Our study recommends AdaBoost algorithms for MM prognosis and suggests using pase-2 approach to short list high risk patients followed by phase 1 to confirm MM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

List of abbreviations

- MPM – Malignant Pleural Mesothelioma
- MM – Malignant Mesothelioma
- PM – Pleural Mesothelioma
- ROC – Receiver Operating Characteristics
- PRC – Precision-recall curve
- DT – Decision tree
- VFDT – Very fast decision tree
- MLP – Multi-layer perceptron
- SGD – Stochastic gradient descent
- KLR – Kernel logistic regression
- AdaBoost – Adaptive boosting
- RMSE – Root mean squared error
- ANN – Artificial neural network
- SVM – Support vector machine
- S-Pegasos - Primal Estimated sub-Gradient Solver for SVM
- CC – Clojure classification

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

- VP – Voted perceptron
- TP – Ture positive
- TN – True negative
- FP – False positive
- FN – False negative
- TPR – Ture positive rate
- FPR – False positive rate
- WBC – White blood cell
- HGB – Hemoglobin
- PLT – Platelet count
- LDH – Blood lactic dehydrogenize
- ALP – Alkaline phosphate
- CRP – C reactive protein
- AUC – Area under the curve

Declarations

- Ethics approval and consent to participate - All data were collected with the permission of the organization and the study ensure no leakage of any patient's medical and personal information.
- Consent for publication - Not applicable
- Availability of data and material - All data analyzed during this study are included in this published article and its supplementary information files.
- Competing interests - The authors declare that they have no competing interests
- Funding - Any internal or external source did not fund this study

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

- Authors’ contribution - AC analyzed, interpreted the mesothelioma data. AC performed the time series forecasting and evaluated the model.
- Acknowledgments - Not applicable
-

Reference

Ascoli, V. (2015). Pathologic diagnosis of malignant mesothelioma: Chronological prospect and advent of recommendations and guidelines. *Ann Ist Super Sanità, 51*(1), 52-59.

Bueno, R., Stawiski, E., Goldstein, L., & al., e. (2016;). Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nat Genet, 48*, 407–416.

Choudhury , A. (2018). Identification of Cancer: Mesothelioma's Disease Using Logistic Regression and Association Rule. *American Journal of Engineering and Applied Sciences, 11*(4), 1310-1319.

Choudhury, A., & Greene, C. (2018). Evaluating Patient Readmission Risk: A Predictive Analytics Approach. *American Journal of Engineering and Applied Sciences, 11*(4), 1320-1331.

Frauenfelder, T., Tutic, M., Weder, W., & al., e. (2011;). Volumetry: an alternative to assess therapy response for malignant pleural mesothelioma? *Eur Respir J ., 38*, 162-168.

Friedin, R. B. (2012, May 11). *Am I going to die because I cannot afford the test?* (KevinMD) Retrieved December 26, 2018, from https://www.kevinmd.com/blog/2012/05/die-afford-test.html

Ghanim, B., Hoda, M. A., Winter, M.-P., Klikovits, T., Alimohammadi, A., Hegedus, B., . . . Berger, W. (2012). Pretreatment serum C-reactive protein levels predict benefit from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at Technology and Health Care 2020; doi:10.3233/THC-202237

multimodality treatment including radical surgery in malignant pleural mesothelioma: a retrospective multicenter analysis. . *Annals of Surgery, 256*(2), 357–362.

Gill, R., Naidich, D., Mitchell, A., & al., e. (2016;). North American multicenter volumetric CT study for clinical staging of malignant pleural 30. mesothelioma: feasibility and logistics of setting up a quantitative imaging study. . *J Thorac Oncol , 11*, 1335–1344.

Haiken, M. (2011, July 17). *3 New Medical Tests that Can Save Your Life - But You Have to Ask.* (Forbes) Retrieved December 26, 2018, from https://www.forbes.com/sites/melaniehaiken/2011/06/17/3-lifesaving-new-medical-tests-you-have-to-ask-for/#2df47f75398a

Ilhan, H. O., & Celik, E. (2016). The mesothelioma disease diagnosis with artificial intelligence methods. *2016 IEEE 10th International Conference on Application of Information and Communication Technologies.* Baku.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R.* New Yrok: Springer.

Kuhn, M., & Johnson, K. (2018). *Applied Predictive Modeling.* New York: Springer.

Lee, S.-J., Chen, T., Yu, L., & Lai, C.-H. (2018). Image Classification Based on the Boost Convolutional Neural Network. *IEEE Access, 6*, 12755-12768.

Leigh, J., Corvalan, C., Grimwood, A., Berry, G., Ferguson, D., & al., e. (1991;). The incidence of malignant mesothelioma in Australia 1982–1988. *Am J Ind Med , 20*, 643–655.

Li, Y. C., Khashab, T., Terhune, J., Eckert, R. L., Hanna, N., Burke, A., & Alexander, H. R. (2017). Preoperative Thrombocytosis Predicts Shortened Survival in Patients with Malignant Peritoneal Mesothelioma Undergoing Operative Cytoreduction and Hyperthermic Intraperitoneal Chemotherapy. *Annals of Surgical Oncology, 24*(8), 2259–2265.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Liaw, A., & Wiener, M. (2002). Classification and Regression by random Forest. *R News, 2*(3), 18-22.

Lotfi, E., & Keshavarz, A. (2014). Gene expression microarray classification using PCA-BEL. *Computers in Biology and Medicine,, 54*, 180-187.

McDonald, C., J., & McDonald., A. D. (1996). The epidemiology of mesothelioma in historical context. . *European Respiratory Journal , 9*, 1932-1942.

Mesothelioma News. (2018, July 28). *What Mesothelioma Does to the Body*. Retrieved August 25, 2018, from http://www.mesotheliomanews.com/medical/mesothelioma-diagnosis/pleural-mesothelioma/

Metintas, M., Metintas, S., Guntulu AK, S. E., Alatas, F., Kurt, E., Uugun, I., & Yildirim, H. (2008). Epidemiology of pleural mesothelioma in a population with non-occupational asbestos exposure. *Respirology, 13*, 117-121.

Miron, B. K., & Witold, R. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 2-13.

Molinari, L. (2018, November Thursday). *Mesothelioma Treatment Costs*. (Cancer Alliance) Retrieved December Saturday, 2018, from https://www.mesothelioma.com/treatment/mesothelioma-treatment-costs/

Nilsson, R., Peña, J. M., Björkegren, J., & Tegńer, J. (2007). Consistent Feature Selection for Pattern Recognition in Polynomial Time. *The Journal of Machine Learning Research, 8*, 612.

Pass, H., Giroux, D., Kennedy, C., & al., e. (2016). The IASLC mesothelioma staging project: improving staging of a rare disease through 29. international participation. *J Thorac Oncol, 11*, 2082-2088.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Paydar, K., R, S., Kalhori, N., Akbarian, M., & Sheikhtaheri, A. (2017). A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus. *International Journal of Medical Informatics , 97*, 239-246.

Peto, J., Hodgson, J., Matthews, K., & Jones, J. (1995;). Continuing increase in mesothelioma mortality in Britain. *Lancet, 345*, 535–539.

Pinder, J. (2012, December 27). *How much does a blood test cost? It could be $6, or $167* . (Clear Health Cost Beta) Retrieved December 26, 2018, from https://clearhealthcosts.com/blog/2012/12/how-much-does-a-blood-test-cost-it-could-be-16-or-117/

Pope, T. P. (2010, February 4). *When Patients Can't Afford Their Care*. (New York Times) Retrieved December 26, 2018, from https://well.blogs.nytimes.com/2010/02/04/when-patients-cant-afford-their-care/

Practo. (2017, December 11). *Protein Total Pleural Fluid*. Retrieved December 25, 2018, from https://www.practo.com/tests/protein-total-pleural-fluid/p

Ron, K., & George, H. J. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*, 273-324.

Selby, K. (2018, December 20). *Mesothelioma Diagnosis* . (Asbestos.com and The Mesothelioma Center) Retrieved December 22, 2018, from https://www.asbestos.com/mesothelioma/diagnosis/

Silvester, J. (2016, September 27). *Most of the World Doesn't Have Access to X-Rays*. (Health) Retrieved December 26, 2018, from https://www.theatlantic.com/health/archive/2016/09/radiology-gap/501803/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201908.0061.v1

Peer-reviewed version available at *Technology and Health Care* **2020**; doi:10.3233/THC-202237

Spirtas, R., Beebe, G., Connelly, R., Wright, W., Peters, J., & al., e. (1986;). Recent trends in mesothelioma incidence in the United States. *Am J Ind Med*, *9*, 397-407.

Stefano, P., Rosa, F., & Paola, M. e. (2015). Differential diagnosis of pleural mesothelioma using Logic Learning Machine. *BMC Bioinformatics* , *16*, 1471-2105.

Takamori, S., Toyokawa, G., Shimokawa, M., Kinoshita, F., Kozuma, Y., Matsubara, T., . . . et.al. (2018). The C-Reactive Protein/Albumin Ratio is a Novel Significant Prognostic Factor in Patients with Malignant Pleural Mesothelioma: A Retrospective Multi-institutional Study . *Ann Surg Oncol*, *25*(17), 1-8.

Tin, K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition.* Montreal.

Vogelzang, N., Rusthoven, J., Symanowski, J., & al., e. (2003). Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *J Clin Oncol 2003; 21:2636–44.*, *21*, 2636-2644.

Zalcman, G., Mazieres, J., Margery, J., Greillier, L., Audigier-Valette, C., Moro-Sibilot, D., . . . et.al. (2016). Bevacizumab for newly diagnosed pleural mesothelioma in the Mesothelioma Avastin Cisplatin Pemetrexed Study (MAPS): a randomised, controlled, open-label, phase 3 trial. *Lancet, 387*, 1405-1414.

List of figure legends