Journal of Data Science 21 (2), 205–224
April 2023

DOI: 10.6339/23-JDS1090
Data Science in Action

Binary Classification of Malignant Mesothelioma: A Comparative Study

Ted Si Yuan Cheng[1],* and Xiyue Liao[1]

[1]*California State University, Long Beach, United States*

Abstract

Malignant mesotheliomas are aggressive cancers that occur in the thin layer of tissue that covers most commonly the linings of the chest or abdomen. Though the cancer itself is rare and deadly, early diagnosis will help with treatment and improve outcomes. Mesothelioma is usually diagnosed in the later stages. Symptoms are similar to other, more common conditions. As such, predicting and diagnosing mesothelioma early is essential to starting early treatment for a cancer that is often diagnosed too late. The goal of this comprehensive empirical comparison is to determine the best-performing model based on recall (sensitivity). We particularly wish to avoid false negatives, as it is costly to diagnose a patient as healthy when they actually have cancer. Model training will be conducted based on k-fold cross validation. Random forest is chosen as the optimal model. According to this model, age and duration of asbestos exposure are ranked as the most important features affecting diagnosis of mesothelioma.

Keywords *binary classification; cancer; class imbalance; machine learning; mesothelioma; variable importance*

1 Introduction

Mesothelioma is a rare and incurable form of cancer that is closely associated with exposure to asbestos (Spugnini et al., 2006). It affects the lining of the lungs and chest most and the abdomen and sac around the heart less. Because the disease is fatal and is largely resistant to radiation and chemotherapy treatment, diagnosis of the disease at early stages is vital in having a better prognosis (Robinson et al., 2005). Machine learning algorithms can assist in identifying the disease early on to improve the quality of life of a patient. Application of machine learning models in cancer research has seen growing popularity in the past two decades (Cruz and Wishart, 2006). For example, genetic algorithms and artificial intelligence were used for early detection of breast cancer by detecting patterns in spectronomy data (Petricoin and Liotta, 2004). Faitima et al's (2020) comparative analysis of various machine learning techniques on breast cancer data found that support vector machine (SVM) was the optimal algorithm in terms of prediction accuracy.

Here, we explore the potential of machine learning algorithms for mesothelioma diagnosis. The mesothelioma data set discussed in this paper is taken from the University of California Irvine (UCI) Machine Learning Repository, with 324 individuals who are healthy or with malignant mesothelioma. 35 variables are included in this data set. The response variable is “class of diagnosis.” It is imbalanced with 70.37% individuals without mesothelioma and 29.63% individuals with mesothelioma. Although the breast cancer data set is not like the mesothelioma data

*Corresponding author. Email: ted.cheng@student.csulb.edu.

© 2023 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.

Received July 25, 2022; Accepted February 6, 2023

set it terms of clinical characteristics, the machine learning methods in Faitima et al (2020) can be further investigated with the mesothelioma data set, for which the research goal is also to build a binary classifier. In addition, there are many small sample clinical data sets and it is important to investigate the performance of machine learning in this setting, which we do here with mesothelioma data.

Our goal in this paper is to compare different machine learning algorithms for predicting mesothelioma diagnosis. We first explore the data structure and features in Section 2. In Section 3, we give more information on model building and decision making. In Section 4, we review the algorithms studied in the paper in Section 4. We present the results of our analysis in Section 5, and close with a discussion of main findings, limitations, and future work in Section 6.

2 Data Summary and Exploratory Data Analysis

Variables of the data set are either continuous or categorical with no missing values. Table 1 is a summary of continuous variables. Table 2 is a summary of categorical variables. More details about this data set can be found in Chicco and Rovelli C (2019).

Table 1: Summary of Continuous Features.

Name	Data Type	Description	Mean ± SD
Age	Continuous	Age of the individuals (year)	54.74 ± 11.00
Duration of asbestos exposure	Continuous	Duration of the environmental exposure to asbestos (year)	30.19 ± 16.42
Duration of symptoms	Continuous	Time period in which the patients show symptoms (year)	5.44 ± 4.72
White blood	Continuous	Number of white blood cells in the pleural fluid per microliter (mcL)	9457.45 ± 3450.73
Cell count (WBC)	Continuous	Number of white blood cells per milliliter (mL)	9.56 ± 3.35
Platelet count (PLT)	Continuous	Number of platelets in blood per milliliter (mL)	369.65 ± 227.55
Sedimentation	Continuous	Measure of how quickly erythrocytes settle in a test tube in one hour measured in millimeter (mm/hour)	70.69 ± 21.75
Blood lactic dehydrogenise (LDH)	Continuous	Amount of LDH measured in international units per liter (IU/L)	308.91 ± 185.14
Alkaline phosphatise (ALP)	Continuous	Amount of ALP in the blood (IU/L)	66.16 ± 35.08
Total protein	Continuous	Total amount of protein in serum in grams per deciliter (g/dL)	6.58 ± 0.83

Continued on next page

Table 1 – *Continued from previous page*

Name	Data Type	Description	Mean ± SD
Albumin	Continuous	Amount of albumin in blood (g/dL)	3.30 ± 0.63
Glucose	Continuous	Amount of glucose in blood in milligram per deciliter (mg/dL)	112.41 ± 38.46
Pleural lactic dehydrogenise	Continuous	Amount of lactic dehydrogenise in pleural fluid(IU/L)	518.47 ± 536.28
Pleural protein	Continuous	Amount of protein in pleural fluid (g/dL)	3.94 ± 1.58
Pleural albumin	Continuous	Amount of albumin in pleural fluid (g/dL)	2.08 ± 0.92
Pleural glucose	Continuous	Amount of glucose in pleural fluid (mg/dL)	48.44 ± 27.23
C-reactive protein (CRP)	Continuous	Acute phase reactant, significantly elevated in patients with pleural mesothelioma (mg/dL)	64.19 ± 22.66

Table 2: Summary of Categorical Features.

Name	Data Type	Description	Count Per Level

Gender	Categorical	Female or male	Female = 134
			Male = 190

City	Categorical	A multinominal variable indicating how far individuals are from downtown. The value is from closest (0) to furthest(8)	0 = 100 | 1 = 42 2 = 51 | 3 = 25 4 = 24 | 5 = 2 6 = 66 | 7 = 13 8 = 1

Asbestos Exposure	Categorical	A binary variable indicating if or not a patient has been exposed to asbestos	Negative = 44
			Exposed = 280

Type of MM	Categorical	A multinominal variable indicating mesothelioma stages/phases to which the symptoms seem to belong	Negative = 310
			Middle = 11
			Late = 3

Diagnosis Method	Categorical	A binary variable indicating if or not the patient has been diagnosed as with mesothelioma by a common diagnosis method	Uncommon = 96
			Common = 228

Continued on next page

Table 2 – *Continued from previous page*

Name	Data Type	Description	Count Per Level

Keep Side	Categorical	A multinomial variable showing the side of the lungs which is experiencing pleural plaques or mesothelioma traces	Left = 100
			Right = 202
			Both = 22

Cytology	Categorical	A binary variable showing if or not cancer cells and certain other cells in the area that surrounds the lung are detected	Negative = 233
			Positive = 91

Dyspnoea	Categorical	A binary variable showing if or not the patient is short of breath	No = 59
			Yes = 265

Ache on Chest	Categorical	Presence or absence of pain in the chest area	No pain = 103 Pain = 221
Weakness	Categorical	A binary variable showing if or not the patient lacks of strength	Not weak= 126 Weak = 198

Habit of cigarette	Categorical	A multinomial variable showing the habit of smoking	None = 183
			Rare = 37
			Regular = 54
			Frequent = 50

Performance status	Categorical	A binary variable showing the patient's ability to perform normal tasks	Normal = 155
			Difficult = 169

Hemoglobin (HGB)	Categorical	A binary variable showing if or not the patient's hemoglobin is normal	Normal = 137
			Abnormal = 187

Dead or not	Categorical	A binary variable showing if or not the patient is dead	Dead = 18
			Alive = 306

Pleural effusion	Categorical	A binary variable showing the presence of pleurl effusion	None = 42
			Present = 282

Pleural thickness on tomography	Categorical	A binary variable showing if or not the patient has any form of thickening involving either parietal or visceral pleura	None = 131
			Thickened = 193

Pleural level of acidity (pH)	Categorical	A binary variable showing if or not the patient's pleural fluid pH level is lower than the normal level	Normal = 155
			Lowered = 169

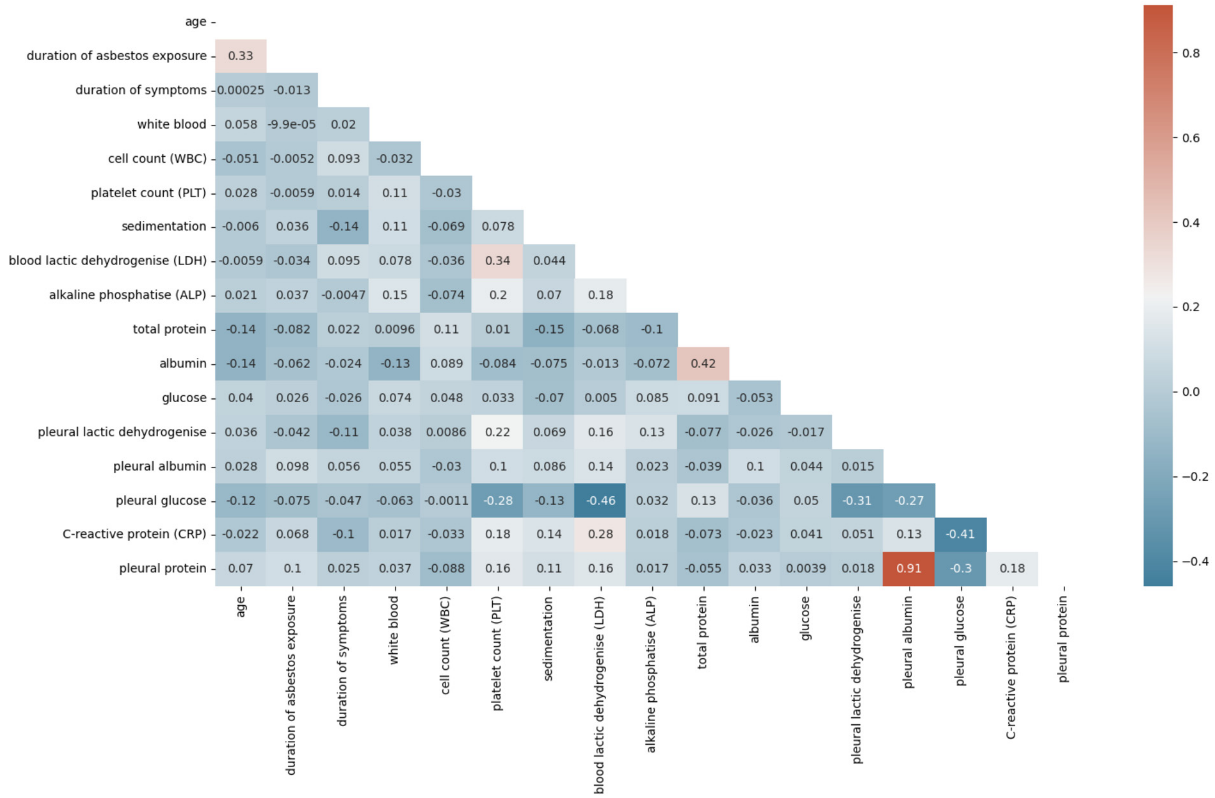

Figure 1: Pearson Correlation heatmap of continuous variables.

A Pearson correlation heatmap is used to check pairwise correlations among continuous variables.

From the results, “pleural albumin” and “pleural protein” have a high correlation of 0.91. The two features have very similar biological meanings, so it is redundant to keep both. In this study, pleural protein is dropped.

3 Methods

3.1 Data Pre-Processing

In this study, we first remove redundant and irrelevant features. In Section 2, we explore the pairwise correlation among continuous predictors, and remove “pleural protein.” Based on the definition of “type of MM” and “dead or not”, we remove the two features because it doesn’t make sense to use them as predictors. Finally, because of the perfect dependence of “diagnosis method” and diagnosis result, we remove “diagnosis method.” All other 30 features are used in modeling.

After removing “pleural protein”, “type of MM”, “dead or not”, and “diagnosis method”, we split the data into training set (80%) and test set (20%). We center and scale each continuous feature. Next, we check the distribution of each continuous feature and log-transformed the following features because of their highly skewed pattern: “duration of symptoms”, “platelet count”, “alkaline phosphatise”, “glucose”, “pleural lactic dehydrogenise”, “cell count”, and “blood lactic dehydrogenise."

The binary response variable is imbalanced and we apply Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla et al., 2002) to the training set to generate data points from the minority class, and make the "diagnosis" variable roughly balanced. SMOTE works by taking a sample of k points from the minority class, creating a synthetic data point, taking the vector between one of those k points and the current synthetic data point. This vector is then multiplied by a random number between 0 and 1, and added to the current synthetic data point to generate a new data point in the minority class.

3.2 Cross Validation and Model Evaluation

We compare 11 models in this study. A brief background of each model is included in Section 4. We use k-fold cross validation to tune parameters and build each model with the training set. Next, we evaluate each model with the test set. Four evaluation metrics defined in Section 4.9 are used.

4 Background

Modeling in this paper is mainly inspired by Fatima et al.'s (2020) comprehensive comparison of various classification algorithms with breast cancer data. We also consider the probabilistic neural network discussed in Er et al. (2012). For certain methods, there are variants such as different kernels for support vector machines (SVM) and distance metrics for k-nearest neighbors. These variants are chosen based on how well documented these methods are for reproducibility, such as the kernels for SVM in the scikit-learn API reference (Pedregosa et al., 2011). Background information of each model we discuss in this paper is included in subsection 4.1 to subsection 4.8.

4.1 Artificial Neural Network (ANN)

Artificial neural networks are computing systems that attempt to emulate neural networks in biological systems, specifically to the human brain (Aggarwal, 2014). ANNs utilizes a myriad of optimization tools to learn from past experiences and use that information to predict and classify new data (Janghel et al., 2010). ANNs are based on neurons, which are defined as atomic parts that compute the aggregation of their input to an output based upon an activation function. An activation function defines the output of that node given a set of inputs. Most commonly activation functions include sigmoid function, ReLU function, and radial basis function.

4.1.1 Multilayer Perceptrons (MLP)

The most commonly used feedforward neural networks is known as multilayer perceptrons. Multilayer perceptrons provide a nonlinear mapping from an input vector to an output vector. It can be used for nonlinear modeling in regression and classification (Hand et al., 2001). The commonly used activation function in MLP (Chollet, 2017) is the sigmoid function, which is defined as

$$S(x) = \frac{1}{(1+e^{-x})}.$$

It is a logistic function where x is the input into the neuron. The function becomes increasingly closer to 1 as $x \to +\infty$ and 0 as $x \to -\infty$ for the probability of classes.

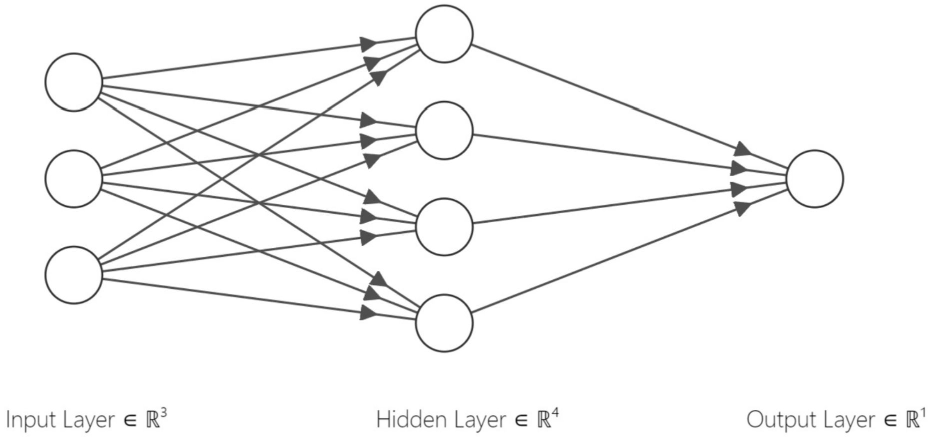
Input Layer ∈ $\mathbb{R}^3$
Hidden Layer ∈ $\mathbb{R}^4$
Output Layer ∈ $\mathbb{R}^1$

Figure 2: Schematic diagram of a multilayer perceptron.

Another commonly used activation function is the Rectified Linear Unit (ReLU) activation function (Logan and Shepp, 1975). It is defined as

$$f(x) = \max(0, x),$$

where x is the input into the neuron. The advantage of the ReLU function is that it is faster to train and has sparse representation when provided with negative inputs.

4.1.2 Probabilistic Neural Network

Probabilistic neural networks (PNN) is another class of feedforward neural networks that are mainly used for classification problems. The method is derived from the Bayesian network and the Kernel Fisher discriminant analysis algorithm. It is a kernelized version of linear discriminant analysis. Compared with MLP, PNN is faster to train and it is less sensitive to outliers. The PNN module from the NeuPy (Shevchuk et al., 2015) package is used in this paper.

The activation function in PNN is the radial basis function (RBF) (Stein and Weiss, 2016). It is defined as

$$\varphi(\mathbf{x}) = \hat{\varphi}(\|\mathbf{x}\|),$$

where φ is a real-valued function and $\|x\|$ is the Euclidean distance between the input x and the origin.

4.2 Support Vector Machine (SVM)

Support vector machine is one of the newer supervised machine learning techniques proposed by Cortes and Vapnik (1995). SVM belongs to a family of generalized linear classifiers. If the training data is linearly separable, then a SVM will construct a separating hyperplane. The classes are divided by the hyperplane, which is of the form

$$w^\top x - b = 0,$$

where w is the normal vector of the hyperplane.

Data points from two classes are separated by the margins

$$w^\top x - b = 1 \text{ and } w^\top x - b = -1.$$

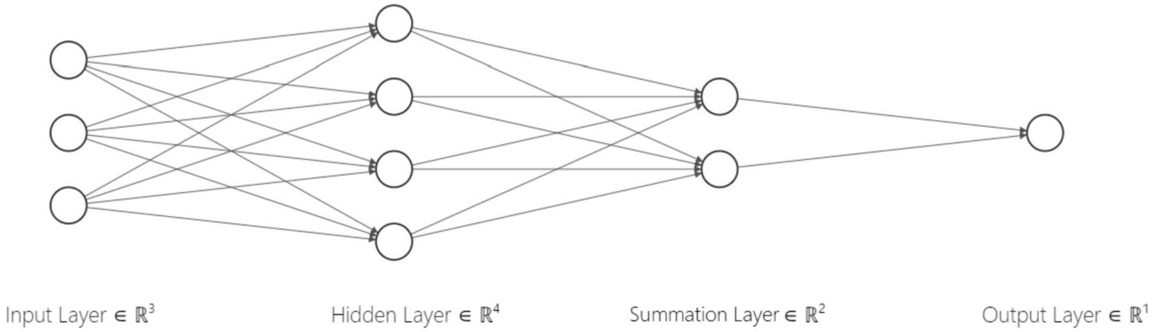

Figure 3: Schematic diagram of a probabilistic neural network.

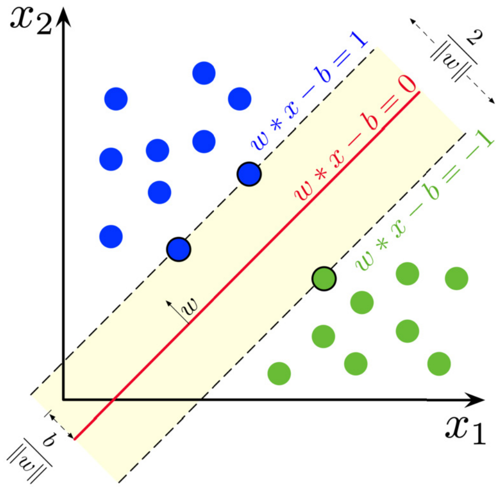

Figure 4: Hyperplane and margins of a SVM for a data set with two classes (Cortes and Vapnik, 1995).

For the ith data point, $i = 1, 2, \ldots, n$, it satisfies:

$$w^\top x_i - b \geqslant 1 \;\text{ if }\; y_i = 1,$$

or

$$w^\top x_i - b \leqslant -1 \;\text{ if }\; y_i = -1.$$

The goal of a linear SVM is to maximize the margin $2/\|w\|$ or minimize $\|w\|$. The optimization problem is to find w and b to minimize

$$\|w\| \;\text{ subject to }\; y_i(w^\top x_i - b) \geqslant 1.$$

When data points are not linearly separable, nonlinear kernels can be used to transform the data points to be in a higher-dimensional feature space. A hyperplane can be defined in this new space to separate the data points (Kotsiantis et al., 2007). Subsections 4.2.1, 4.2.2 and 4.2.3 list 3 non-linear kernels (Pedregosa et al., 2011).

4.2.1 Polynomial Kernel

The polynomial kernel is defined as

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d, \quad i \neq j, \quad i, j = 1, 2, \ldots, n,$$

where d ($d \geqslant 2$) denotes the number of degrees of the polynomial.

4.2.2 Radial Basis Kernel

The radial basis kernel is the most popular kernel used in a SVM. Similar to the k-NN algorithm, it calculates the Euclidean distance between two data points to determine their similarity. It can also be easily tuned with the γ parameter. It is defined as

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad i \neq j, \quad i, j = 1, 2, \ldots, n,$$

where $\gamma > 0$ and $\|x_i - x_j\|$ is the Euclidean distance between x_i and x_j.

4.2.3 Sigmoid Kernel

The sigmoid kernel is popular in neural networks. In a SVM, it is often used as a proxy to a two-layer perceptron. It is defined as

$$K(x_i, x_j) = \tanh(\alpha x_i \cdot x_j + c), \quad i \neq j, \quad i, j = 1, 2, \ldots, n,$$

where $\alpha > 0$ and $c < 0$.

4.3 Logistic Regression (LR)

Logistic regression is a generalized linear regression model using a logistic function to estimate the occurring probability of the event of "success" in a binary response. Typically, the event of success is coded as 1; the event of failure is coded as 0.

The logit link function connecting the occurring probability of success: $p(x)$ for x_i, $i = 1, 2, \ldots, n$, with the predictors is defined as

$$\log\left[\frac{p(x_i)}{1 - p(x_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} +, \ldots, +\beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

When there is no imbalance issue in the binary response, if the estimated $p(x)$ is less than 0.5, then a data point is assigned to the failure class; otherwise it is assigned to the success class.

4.3.1 Regularization

Regularization is defined as a modification made to a learning algorithm that reduces the testing error but not the training error (Goodfellow et al., 2016). Adding a regularization term helps prevent overfitting on the training set. Since small data sets are more prone to overfitting, it is worthwhile to experiment with regularization methods. The regularization methods included in the scikit-learn library (Pedregosa et al., 2011) are Lasso regularization (L_1 penalty), ridge regularization (L_2 penalty), and elastic net regularization (both L_1 and L_2 penalties are added) (Müller and Guido, 2016).

Ridge regularization (James et al., 2013) uses an L_2 penalty. It is similar to the least squares criterion used in ordinary linear regression, except the coefficient estimates β^R are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = RSS + \lambda \sum_{j=1}^{p}\beta_j^2,$$

where λ is a tuning parameter and RSS is the residual sum of squares. Its distance is calculated as the Euclidean distance of the β vector from the origin.

Lasso regularization (James et al., 2013) uses an L_1 penalty. It is an alternative to ridge regression. The lasso coefficients β_λ^L are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|.$$

Lasso regularization can force some coefficient estimates to be 0, while ridge regularization will only have some coefficients to be close to 0 but never to be equal to 0. Therefore, Lasso regularization leads to more interpretable models.

Elastic net regularization is a method that combines both L_1 and L_2 penalties (Müller and Guido, 2016). The elastic net coefficients β^E are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p}\beta_j^2 + \lambda \sum_{j=1}^{p}|\beta_j|.$$

4.4 k-Nearest Neighbor (k-NN)

The k-NN algorithm is one of the most fundamental algorithms for both regression and classification. It is a method that can be used when there is no prior knowledge of the underlying pattern in the data (Peterson, 2009). It is a nonparametric classifier and usually has good performance when the decision boundaries are irregular (Hastie et al., 2009). As the name implies, for an unclassified data point, the k-NN algorithm will find k training samples closest to the unclassified data point according to some distance measure, and then classify the data point using majority vote (Pedregosa et al., 2011). The value of k is usually an odd number to avoid ties in the decision.

Subsections 4.4.1, 4.4.2 and 4.4.3 list three distances commonly used in the k-NN algorithm.

4.4.1 Minkowski Distance

Minkowski distance is a metric used in a normed vector space and is the generalized form of the Euclidean and Manhattan distance metrics. It is defined as

$$d\,(x, y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p},$$

where $p \geqslant 1$.

4.4.2 Euclidean Distance

Euclidean distance is the most popular distance for k-NN. It is defined as

$$d\,(x, y) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2}.$$

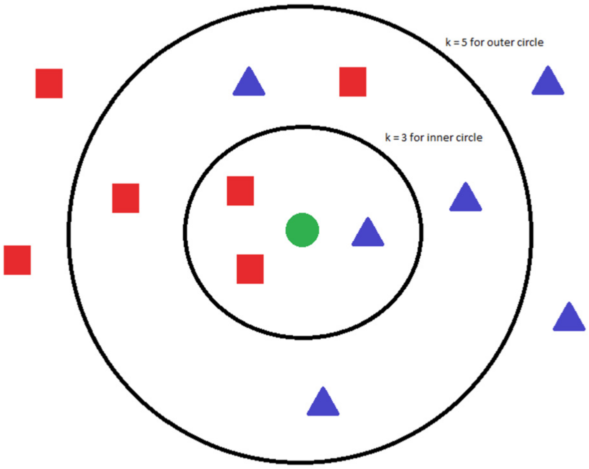

Figure 5: Example of k-NN for two classes where the center dot is the unclassified data point.

4.4.3 Manhattan Distance

Manhattan distance is defined as

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|.$$

It is preferred over Euclidean distance for data sets in a high dimension (Aggarwal et al., 2001).

4.5 Naïve Bayes (NB)

Naïve Bayes classifier is a popular method for data sets whose features are in a high dimension space (Hastie et al., 2009). It is a probabilistic classifier that is based on Bayes' theorem with the assumption of strong independence among the features. The goal of Naïve Bayes classifier is to maximize the posterior probability defined as

$$Pr(Y_i = k \mid x_i) = \frac{\pi_k \times f_{k1}(x_{i1}) \times f_{k2}(x_{i2}) \times \cdots \times f_{kp}(x_{ip})}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_{i1}) \times f_{l2}(x_{i2}) \times \cdots \times f_{lp}(x_{ip})}, \quad i = 1, 2, \ldots, n,$$

for $k = 1, \ldots, K$, where K is the total number of classes. $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is a vector of p features. π_k is the prior probability of an observation falling in class k. f_{kl} is the density function of the lth predictor among observations in the kth class.

The classification is obtained as

$$\hat{Y}_i = \underset{k \in \{1,\ldots,K\}}{\text{argmax}}\, Pr(Y_i = k \mid x_i), \quad i = 1, 2, \ldots, n.$$

Subsections 4.5.1 and 4.5.2 describe 2 commonly used variants of NB. As the mesothelioma data set has both continuous and categorical features, the two NB models will be assessed for categorical and continuous features separately.

4.5.1 Gaussian NB

The Gaussian NB algorithm assumes the likelihood of the features to follow a multivariate normal distribution.

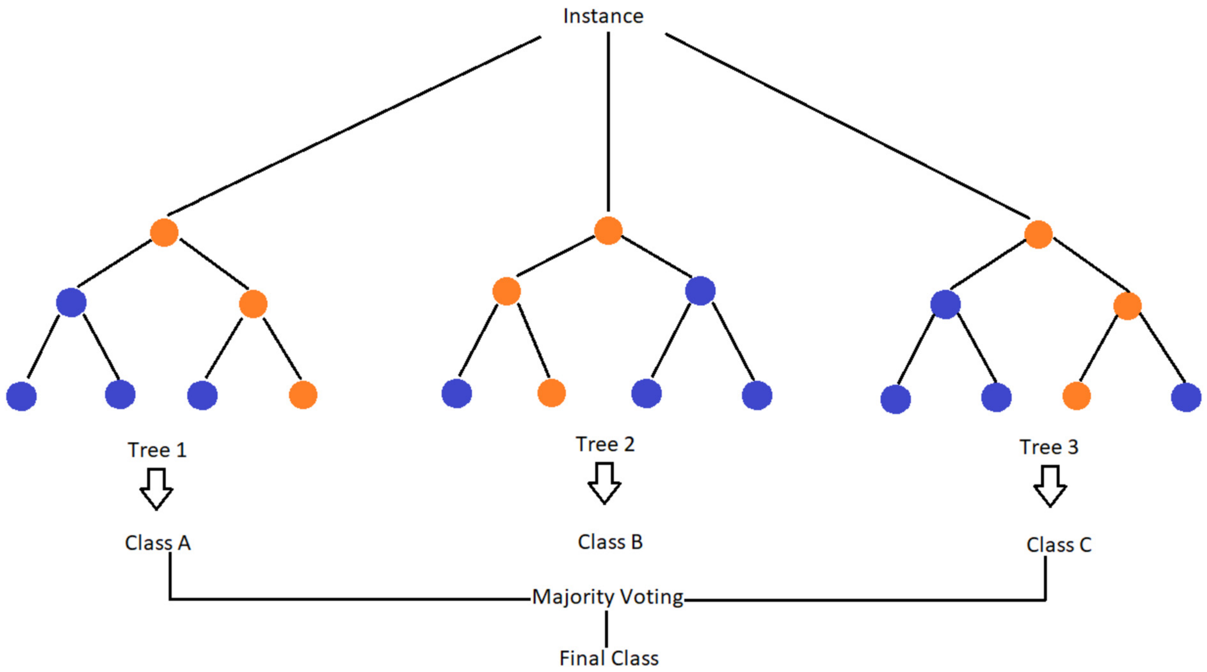

Figure 6: Diagram of a random forest model.

4.5.2 Multinomial NB

Multinomial NB, unlike the Gaussian NB, supports categorical features that follow a multinomial distribution.

4.6 Random Forest (RF)

Random forest is an extension of decision tree. A decision tree makes predictions by greedily splitting the predictors' space into sub-spaces, which are known as nodes of the tree. Predictions are then made at each node by averaging the values in the node. A decision tree can easily overfit a data set and has high variance. Random forests are developed by fitting a number of decision trees and then making predictions by averaging the predicted values of each individual decision tree, to reduce variance of predictions. Meanwhile, when fitting a random forest, only a subset of predictors are used as candidate predictors to make a split, and this de-correlates decision trees (Hastie et al., 2009).

4.7 Gradient Boosting (GB)

Gradient boosting is another algorithm that can be used to improve the prediction of a single decision tree. Unlike random forest, boosting doesn't build a number of trees using bootstrapping samples. Instead, trees are grown sequentially. Each tree is grown using the residual information of the previous tree. Each tree can be rather small with just a few terminal nodes (Hastie et al., 2009).

Subsections 4.7.1 and 4.7.2 list two variants of GB.

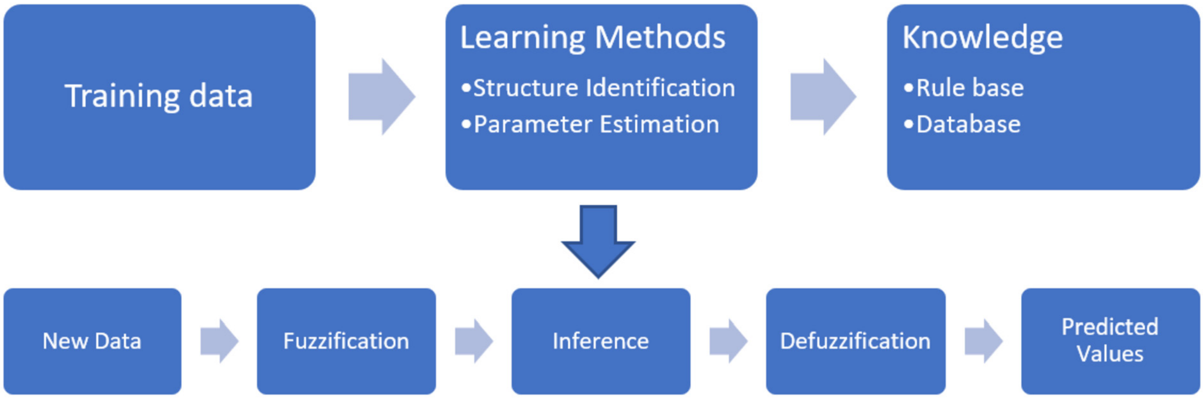

Figure 7: FRBS process based on Riza et al.'s (2015) construction.

4.7.1 LightGBM

LightGBM is a gradient boosting framework that uses two novel techniques to improve model efficiency, namely, Gradient-Based One Side Sampling (GOSS) and Exclusive Feature Bundling Technique (EFB) (Ke et al., 2017). GOSS keeps a small fraction of instances that have large gradient values, which contribute most to information gain, and randomly sample instances with small gradient values to retain the accuracy of information gain. EFB bundles mutually exclusive features into one single feature to improve training speed.

4.7.2 Extreme Gradient Boosting (XGBoost)

XGBoost is an open source software library developed by Chen and Guestrin (2016). XGBoost builds decision trees in a parallel pattern, instead of sequentially like regular GB. This allows fast training if the user has access to a computation platform that allows parallel computing. Besides, XGBoost is a regularized form of gradient boosting, and it reduces overfitting and improves the model's generalizability.

4.8 Fuzzy rule-based systems (FRBSs)

Fuzzy rule-based systems are methods based on fuzzy concepts to tackle problems such as uncertainty, imprecision, and non-linearity in identification, regression, and classification tasks. It is a method of soft computing (Riza et al., 2015). FRBSs are based on fuzzy set theory, which represents human knowledge in a set of fuzzy IF-THEN rules (Zadeh, 1965). The rules are defined as having degrees of membership instead of a binary membership value, such that the values will be between 0 and 1 rather than 0 or 1. A degree of 1 means that an object is a member of the set, a value of 0 means it is not a member, and a value in-between shows a partial degree of membership. Fuzzy logic allows vagueness and overlapping class definitions. Compared with other classifiers, which use crisp sets (sets in which a characteristic function assigns a binary 0 or 1 to each member), FRBSs are more interpretable because they emulates the knowledge of human experts in a set of fuzzy IF-THEN rules.

Subsections 4.8.1 and 4.8.2 list 3 FRBS models. The first 2 models are from the R package frbs (Riza et al., 2015) and the third model is from the R package RoughSets (Janusz and Riza, 2019).

4.8.1 FRBCS.W and SLAVE

The R package frbs implements the most widely used FRBS models. In this package, 5 models are designed for classification tasks. We experiment with FRBCS.W and SLAVE. FRBCS using Ishibuchi's method with weight factor (FRBCS.W) is a fuzzy rule-based classification system imposing certainty grades, i.e., rule weights to IF-THEN rules (Ishibuchi and Nakashima, 2000). The authors of the frbs package applied FRBCS.W to the iris data set (Fisher, 1936) for a binary classification task and its prediction accuracy was about 95%. Structural learning algorithm on vague environment (SLAVE) is a genetic learning algorithm that uses the iterative approach to learn fuzzy rules. Rafique et al. (2011) compared SLAVE with models such as NB, SVM, etc in detecting Short Message Service (SMS) spam, and SLAVE showed the highest detection rate and lowest false alarm rate. The drawback of FRBCS.W and SLAVE is that they can only use continuous features. We only assess them for continuous features in this paper.

4.8.2 Fuzzy-rough nearest neighbors (FRNN)

Onan (2015) proposed a fuzzy-rough nearest neighbors (FRNN) algorithm where the nearest neighbors are used to construct the fuzzy lower and upper approximations of decision classes, and test instances are classified based on their membership to these approximations. In that paper, authors compared FRNN with other classifiers such as SVM, k-NN, and fuzzy nearest neighbors classifier with the Wisconsin breast cancer data set (Street et al., 1993) and showed that FRNN outperformed other methods.

4.9 Model Validation

k-fold cross validation is used to evaluate each algorithm. Evaluation metrics are specified in subsections 4.9.1, 4.9.2, 4.9.3, and 4.9.4. We choose $k = 7$. For each metric, we need the following terms: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For this data set, a positive (negative) case represents an individual with (without) mesothelioma. Each metric ranges from 0 to 1, inclusively. A larger value implies better performance.

4.9.1 Accuracy

Prediction accuracy is the most commonly used evaluation metric. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Accuracy may not be the most useful metric when the outcome is imbalanced. For example, when most data points are in the positive group, then accuracy can be a large value but it doesn't imply that the model is effective in identifying cases in the negative group.

4.9.2 Recall

Recall measures how many out of all positive cases, are predicted correctly. It is a very important metric in evaluating a model's performance when researchers emphasize on correctly detecting positive cases. It is defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

4.9.3 Precision

Precision is the ratio of correctly classified positive cases to all predicted positive values. It is defined as

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

4.9.4 F1-score

F1-score is a harmonic mean of precision and recall. It is used if researchers care about precision and recall equally. It is defined as

$$\text{F1-Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}.$$

4.10 Model Tuning

In this study, we tune 7 models. Specifically, we tune MLP on the number of units for the hidden layer. Candidate values are 1, 8, 15, and 30. For support vector machine, tuning is done on the kernel types, which include linear kernel and non-linear kernels defined in Section 4.2. For logistic regression, tuning is done on the regularization methods, which include logistic regression without regularization and the 3 regularization methods defined in Section 4.3. It is also done on class weight (balanced vs imbalanced). For the sake of interpretation, we fit a logistic regression model with first-order terms only, and the features that have a p-value smaller than 0.05 are included in Table 5.2. For k-NN, tuning is done on the number of neighbors k. Candidate values are a grid starting from 1 and ending at 17, which is the rounded number of the square root of sample size. Step in this grid is 2. For random forest and gradient boosting methods, tuning is done on the number of features used in modeling. We compare the model using all 30 features versus models with a subset of features, which is selected based on features' importance values.

5 Results

Subsection 5.1 includes a table evaluating the models with metrics defined in 4.9. Subsection 5.2 includes a table interpreting the coefficient estimate of significant predictors in the logistic regression model. Subsection 5.3 includes a variable importance ranking plot. Subsection 5.4 includes a table comparing computation time for each model.

5.1 Model Results

Table 3 lists the best performing variant of each model discussed in Section 4. The largest value for each metric is shown with a bold font. k-NN has the largest recall and F1-score value, which means that it identifies true positives or people with the mesothelioma cancer successfully. However, for this variant, $k = 1$ and this implies overfitting, and we prefer not to choose this model because this model is hard to be generalized to another data set.

Models like PNN, SVM, and RF have a relatively high recall value. However, PNN is not well documented, though it has a high recall value. It is included in this paper to be compared with other models as it is one of the original methods discussed in Er et al. (2012). RF has the second largest F1-score, largest accuracy and largest precision value. We consider RF as the optimal model.

Table 3: Model Evaluation Results.

Algorithm	Accuracy	Recall	Precision	F1-Score
MLP (1 layer, 30 units)	80.77%	85.04%	67.29%	74.10%
PNN	74.18%	96.64%	61.20%	72.35%
SVM (Polynomial)	81.59%	85.71%	79.22%	82.03%
LR (Lasso)	65.66%	69.78%	64.36%	66.86%
k-NN ($k = 1$, Manhattan distance)	84.07%	**97.80%**	76.87%	**86.00%**
NB (Gaussian)	58.79%	85.16%	55.87%	67.37%
RF	**86.54%**	85.16%	**88.44%**	85.37%
XGboost	84.07%	81.32%	85.73%	81.85%
FRBCS.W (continuous)	39.47%	63.64%	26.92%	37.84%
FRNN	34.21%	81.82%	28.12%	41.861%
SLAVE (continuous)	36.84%	63.64%	25.93%	36.85%

Table 4: Summary of the Logistic Regression Model.

Variable	Odds Ratio (95% C.I.)	p-value
Age	0.48 ([0.35, 0.67])	0.000
Gender	0.64 ([0.48, 0.85])	0.002
Duration of asbestos exposure	2.34 ([1.46, 3.74])	0.000
Keep side	1.49 ([1.14, 1.94])	0.004
Duration of symptoms	1.37 ([1.03, 1.81])	0.029
Cell count (WBC)	0.74 ([0.56, 0.97])	0.030
Hemoglobin (HGB)	1.37 ([1.05, 1.79])	0.022
Platelet count (PLT)	0.71 ([0.51, 0.998])	0.048

5.2 Summary of the Logistic Regression Model

Table 4 summarizes the odds ratio of each significant predictor from logistic regression. There is no comparison between Table 4 and Table 3. We include Table 4 here because of the interpretability of the logistic regression model.

Table 4 shows that mesothelioma is more prevalent among individuals with a longer asbestos exposure, with pleural plaques being detected in the right lung or both lungs, with a longer duration of symptoms, and with more hemoglobin. It is less prevalent among individuals with more platelets, with more white blood cells, and who are female.

5.3 Feature Ranking of the Random Forest Model

Figure 8 ranks features according to their feature importance value, which is computed as the mean of accumulation of the impurity decrease within each tree. Here, we only include features whose importance values are above the average of all feature importance values.

5.4 Computation Time Comparison

Table 5 compares computation time of each model. In this table, we include the average computation time in k-fold cross validation. Guassian NB is the fastest model. Except for SLAVE

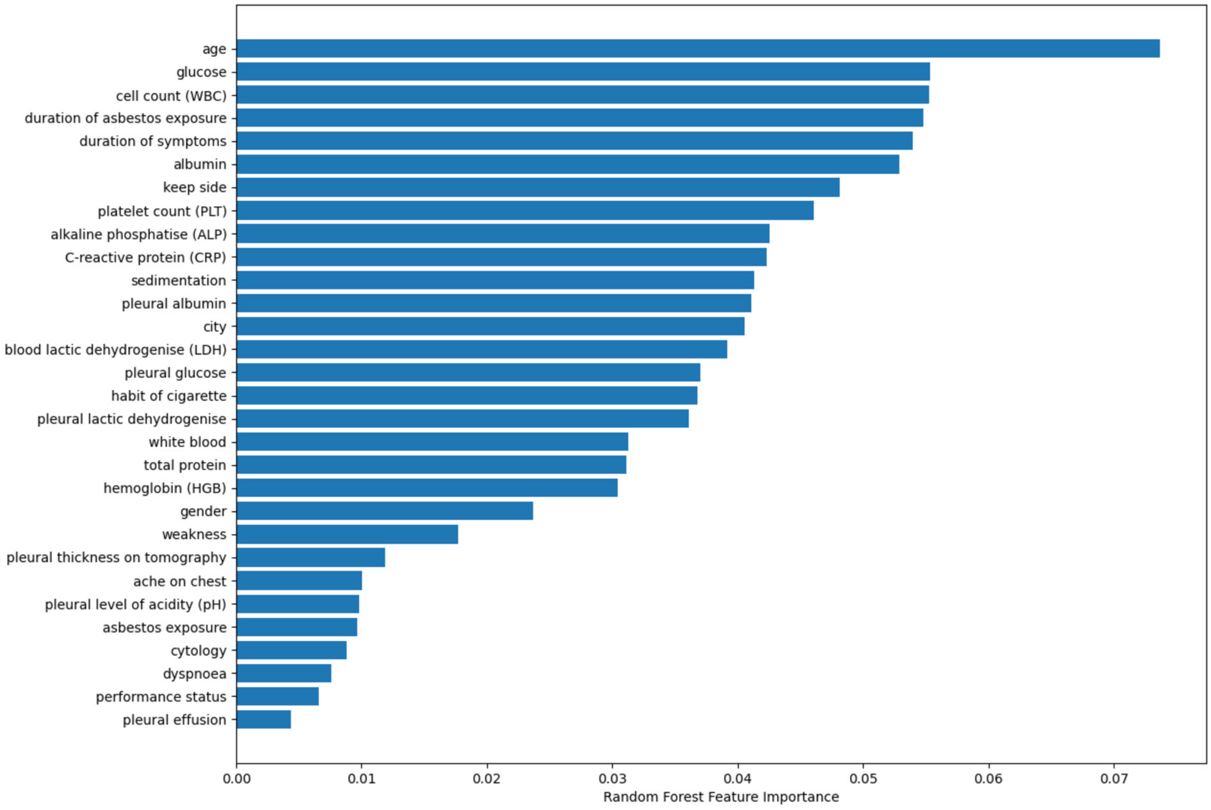

Figure 8: Random Forest Feature Importance Ranking.

Table 5: Computation Time Comparison.

Algorithm	Time (seconds)
MLP (1 layer, 30 units)	0.871
PNN	0.003
SVM (Polynomial)	0.011
LR (Lasso)	0.005
k-NN ($k = 1$, Manhattan distance)	0.005
NB (Gaussian)	**0.002**
RF	0.0100
XGboost	0.046
FRBCS.W (continuous)	1.680
FRNN	0.206
SLAVE (continuous)	60.978

and FRBCS.W, the computation time for every other model is within 1 second.

6 Discussion

In this paper, we have explored 11 different classification models for a mesothelioma data set with 324 observations and 34 features. We present evidence that random forest has high accuracy, precision, recall and F1-score values. Although its recall value is not the highest, it is above 85% and has better generalization than k-NN ($k = 1$). Logistic regression model shows that exposure to asbestos, gender, and keep side are all strongly predictive of the prevalence of mesothelioma. Fuzzy logic-based models don't have good performance with this data set, though they are novel.

Something that was not discussed in most papers using this data set is the deterministic association between "diagnosis method" and the response "class of diagnosis." When the data set was initially assessed with this predictor, most of the algorithms yielded perfect metrics across the board. Once "diagnosis method" was removed from the model, the algorithms yielded significantly worse metrics. It is important to exclude this variable before modeling and this is what we do in this paper.

It is known that old age and extended exposure to asbestos have long been contributing factors to malignant mesothelioma. It is also interesting to note that there is some association with sex, which may be attributable to the over-representation of male subjects in the sample. Biological markers shown in Figure 8 can be scrutinized more closely by individuals with a medical background to provide insight into their dependency with malignant mesothelioma.

In the future, researchers may consider recruiting more individuals to mitigate the overfitting issue with a small data set. However, this can be difficult with a rare disease like mesothelioma. A limitation in this paper is that model tuning is not thoroughly explored. For example, interaction terms or polynomial terms in the logistic regression model are not discussed and they may help improve the performance of logistic regression.

Supplementary Material

The zip supplementary material file contains the Python and R scripts for reading data and preprocessing, exploratory data analysis, and the various models tested.

References

Aggarwal CC (Ed.) (2014). *Data Classification: Algorithms and Applications*. CRC Press, Yorktown Heights, NY, USA.

Aggarwal CC, Hinneburg A, Keim DA (2001). On the surprising behavior of distance metrics in high dimensional space. In: *International Conference on Database Theory*, 420–434. Springer.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357. https://doi.org/10.1613/jair.953

Chen T, Guestrin C (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

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

Chollet F (2017). *Deep Learning with Python*. Manning, New York, NY, USA.

Cortes C, Vapnik V (1995). Support-vector networks. *Machine Learning*, 20(3): 273–297.

Cruz JA, Wishart DS (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2: 59–77.

Er O, Tanrikulu AC, Abakay A, Temurtas F (2012). An approach based on probabilistic neural network for diagnosis of mesothelioma's disease. *Computers & Electrical Engineering*, 38(1): 75–81. https://doi.org/10.1016/j.compeleceng.2011.09.001

Fatima N, Liu L, Hong S, Ahmed H (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8: 150360–150376. https://doi.org/10.1109/ACCESS.2020.3016715

Fisher RA (1936). The use of multiple measurements in taxonomic problem. *Annals of Human Genetics* 7: 179–188.

Goodfellow I, Bengio Y, Courville A (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.

Hand DJ, Smyth P, Mannila H (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA, USA.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer, New York, NY, USA.

Ishibuchi H, Nakashima T (2000). Effect of rule weights in fuzzy rule-based classification systems. In: *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No. 00CH37063)*. volume 1. 59–64. vol.1.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer, New York, NY, USA.

Janghel RR, Shukla A, Tiwari R, Kala R (2010). Breast cancer diagnosis using artificial neural network models. In: *The 3rd International Conference on Information Sciences and Interaction Sciences*, 89–94. IEEE.

Janusz A, Riza LS (2019). RoughSets: Data Analysis Using Rough Set and Fuzzy Rough Set Theories. R package version 1.3-7.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30: 3146–3154.

Kotsiantis SB, Zaharakis I, Pintelas P, et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 160(1): 3–24.

Logan BF, Shepp LA (1975). Optimal reconstruction of a function from its projections. *Duke Mathematical Journal*, 42(4): 645–659.

Müller AC, Guido S (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc., USA.

Onan A (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications*, 42(20): 6844–6852. https://doi.org/10.1016/j.eswa.2015.05.006

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12: 2825–2830.

Peterson LE (2009). K-nearest neighbor. *Scholarpedia*, 4(2): 1883. https://doi.org/10.4249/scholarpedia.1883

Petricoin EF, Liotta LA (2004). Seldi-tof-based serum proteomic pattern diagnostics for early detection of cancer. *Current Opinion in Biotechnology*, 15(1): 24–30. https://doi.org/10.1016/j.copbio.2004.01.005

Rafique MZ, Alrayes N, Khan MK (2011). Application of evolutionary algorithms in detecting sms spam at access layer. In: *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, 1787–1794.

Riza LS, Bergmeir C, Herrera F, Benítez JM (2015). frbs: Fuzzy rule-based systems for classification and regression in R. *Journal of Statistical Software*, 65(6): 1–30. https://doi.org/10.18637/jss.v065.i06

Robinson BW, Musk AW, Lake RA (2005). Malignant mesothelioma. *The Lancet*, 366(9483): 397–408. https://doi.org/10.1016/S0140-6736(05)67025-0

Shevchuk Y, et al. (2015). Neupy. http://neupy.com/.

Spugnini EP, Bosari S, Citro G, Lorenzon I, Cognetti F, Baldi A (2006). Human malignant mesothelioma: Molecular mechanisms of pathogenesis and progression. *The International Journal of Biochemistry & Cell Biology*, 38(12): 2000–2004. https://doi.org/10.1016/j.biocel.2006.07.002

Stein EM, Weiss G (2016). Introduction to fourier analysis on euclidean spaces (pms-32), volume 32. In: *Introduction to Fourier Analysis on Euclidean Spaces (PMS-32)*, volume 32. Princeton university press.

Street WN, Wolberg WH, Mangasarian OL (1993). Nuclear feature extraction for breast tumor diagnosis. In: *Biomedical Image Processing and Biomedical Visualization*, volume 1905, 861–870. SPIE.

Zadeh LA (1965). Fuzzy sets. *Information and Control*, 8: 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X